Driver Attention Assessment from Gaze and Situational Variables

Zur Erlangung des akademischen Grades des
Doktors der Ingenieurswissenschaften
der KIT-Fakultät für Informatik
des Karlsruher Institut für Technologie (KIT)

genehmigte
Dissertation
von

Felix Martin Schmitt
aus Darmstadt

Tag der mündlichen Prüfung: 09. Januar 2018
Referent: Prof. Dr.-Ing. Rainer Stiefelhagen
Korreferent: Prof. Ph.D. Constantin A. Rothkopf

Kurzfassung

Fahrer, die der Fahrsituation nicht genügend Aufmerksamkeit widmen, stellen eine Gefahr für die Verkehrssicherheit dar. Dies liegt daran, dass in diesem Fall das Fahrvermögen der Betroffenen deutlich verringert ist, was in Folge zu einem erhöhten Unfallrisiko führt. Deshalb versprechen Systeme, die die Fahreraufmerksamkeit automatisch beurteilen und entsprechend warnen oder eingreifen können, eine große Verbesserung der Verkehrssicherheit. Hierbei ist aber eine genaue und echtzeitfähige Beurteilung der Fahreraufmerksamkeit bezüglich des damit verbundenen Unfallrisikos erforderlich.

Diese Dissertation führt eine neue Methode zur Beurteilung von Fahreraufmerksamkeit im situativen Kontext ein. Es wird vorgeschlagen angemessenes Blickverhalten durch Blickstrategien in einem entscheidungstheoretischen Formalismus festzulegen. In diesem Ansatz werden Modelle der Fahrsituation sowie der Wahrnehmung und der Fahrzeugführung des Fahrers verwendet. Bisherige Arbeiten beurteilen Fahreraufmerksamkeit zumeist alleine anhand Fahr- und Blickverhaltens. Ein deutlicher Nachteil ist dabei, dass somit das Zusammenspiel aus Fahrerverhalten, Fahrsituation und Unfallrisiko vernachlässigt wird. Das ist umso gravierender, da bekannt ist, dass erfahrene Fahrer an die Fahrsituation abgestimmte Blickstrategien zeigen, die die Beeinträchtigung ihrer Fahrleistung abmildern können. Ähnliche Blickstrategien enstehen auf natürliche Art und Weise aus dem gewählten entscheidungstheoretischen Ansatz.

In der Arbeit wird der entscheidungstheoretische Ansatz beispielhaft an der Fahraufgabe des Spurhaltens untersucht. Hier wird auf die Modellbildung, die Echtzeitberechnung, die passende Parametrisierung sowie auf die Evaluierung der Methode in der Anwendung in einem neuen Warnsystem eingegangen.

Zuerst wird die Aufgabe des Spurhaltens bei einer Nebenaufgabe, die um die visuelle Aufmerksamkeit konkurriert, modelliert. Dazu wird ein *Partially Observable Markov Decision Process* (POMDP) verwendet, der ein kinematisches Model der Fahraufgabe, ein Model der sensorischen Eigenschaften des Fahrers sowie ein Modell der Nebenaufgabe enthält. Danach wird die Berechnung von Strategien in dem POMDP untersucht. Diese Strategien dienen dazu das angemessene Blickverhalten festzulegen. Schließlich wird die Wirklichkeitstreue dieser Strategien überprüft und der erforderliche Rechenaufwand analysiert.

Zweitens wird die Wahl einer passenden Belohnungsfunktion betrachtet. Diese ist deswegen von Bedeutung, da sie schlussendlich das angemessene Blickverhalten festlegt. Es wird ein neues Verfahren der inversen optimalen Steuerung entwickelt, das es vermag Parameter der Belohnungsfunktion aus dem Verhalten erfahrener Fahrer zu schätzen. In einem Experiment im Realverkehr erhobenes Fahrerverhalten wird benutzt um die entwickelte Methode hinsichtlich der Genauigkeit in der Verhaltensvorhersage zu prüfen.

Die vorliegende Arbeit untersucht drittens die Schätzung von Modellen der sensorischen Eigenschaften von Fahrern. Dazu wird der erste allgemeine Ansatz für dieses Inferenzproblem in sequenziellen Entscheidungsproblemen vorgestellt. Darauffolgend wird eine Umsetzung des Ansatzes für den vorherig eingeführten POMDP entwickelt. Das resultierende Verfahren wird mittels Fahrverhaltensdaten aus einem weiteren Fahrversuch geprüft.

Schließlich wird viertens die Entwicklung eines Warnsystems und dessen Einbindung in ein Versuchsfahrzeug verfolgt. Das System zielt darauf ab den Fahrer bei der Aufrechterhaltung von genügen Aufmerksamkeit zu unterstützen. In einem abschließenden Nutzertest wird das entwickelte System mit einem Warnsystem nach dem aktuellen Stand der Technik verglichen, wobei sowohl die Akzeptanz durch die Nutzer als auch die Auswirkungen auf die Fahrleistung untersucht werden.

Im Ganzen verdeutlicht diese Arbeit die Umsetzbarkeit und die Vorteile des verfolgten Ansatzes des angemessenen Blickverhaltens für die automatische Bewertung von Fahreraufmerksamkeit. Es wurde gezeigt, dass der benötigte Rechenaufwand eine Echtzeitanwendung zulässt und dass geeignete Modellparameter automatisch geschätzt werden können. Schließlich wurde die Verbesserung eines Ablenkungswarnsystems belegt. Folglich stellt die Methodologie, die in dieser Arbeit eingeführt

wurde, einen vielversprechenden neuen Ansatz zur Bewertung von Fahreraufmerksamkeit dar, der die Probleme des aktuellen Standes der Technik vermeidet.

Abstract

Drivers who pay insufficient attention to the road scenery may become a road hazard. This is because their driving performance is significantly impaired which results in increased crash risk. Therefore, automatic systems that can assess the driver's attention and intervene or warn accordingly promise a great benefit for road safety. In this context, a precise and real-time assessment of driver attention with respect to the associate crash risk is required.

The present thesis establishes a novel framework for assessment of driver's visual attention in the situational context. It is proposed to define appropriated glance behavior by means of a rational policy in a decision theoretic formalism. This approach features both models of the driving situation as well as the driver's perception and vehicle control. In previous work the driver's attention is mostly assessed based on driving and glance behavior alone. A significant drawback is that the important interactions between driver behavior, situational context and crash risk are neglected. This is especially problematic as experienced drivers haven been shown to apply situationally adaptive glance behavior which can to some extend mitigate impairment of driving performance. In the decision theoretic framework proposed in this thesis, similar rational glance strategies naturally emerge.

In this work the decision theoretic model of appropriate glance behavior is investigated at the exemplar driving task of lane keeping. The thesis addresses model development, real-time computation, suitable parameterization as well as evaluation in application for a novel warning system.

First, the task of lane keeping in presence of an additional task concurring for visual attention is modeled. This is done by means of a Partially Observable Markov Decision Process (POMDP) which features a kinematic model of the driving task, a model of the driver's sensor characteristics and a model of the additional task as well as corresponding reward functions. In this POMDP approaches for computation of rational policies are considered. These policies are used to define appropriate glance behavior. Finally, the policies are evaluated with respect to realism and the involved computational demands.

Second, this work considers the suitable parameterization of the reward functions of the POMDP. This is of importance because the reward parameterization in the end determines the computed appropriate glance behavior. A novel Inverse Optimal Control (IOC) approach is developed for estimation of reward parameters from the behavior of experienced drivers. The methodology is evaluated with respect to prediction of driver behavior recorded in a driving experiment in real traffic.

Third, the estimation of models of the driver's sensor characteristics is addressed. For this purpose, the first general inference framework for sensor models underlying behavior in sequential decision making problems is established. The implementation of the framework for the previously developed POMDP model is derived. An evaluation of the proposed approach using behavioral data from a second driving experiment is conducted.

Fourth and finally, a novel warning system to assist the driver in maintaining sufficient attention is developed and implemented in a test vehicle. The new system is compared to a state-of-the-art warning approach by means of an evaluation in a user study. Here, both user acceptance and effects on driving performance are considered.

Overall, this thesis demonstrated the viability of the pursued concept of appropriate glance behavior for situation-specific assessment of driver attention. It was shown that computational demands are feasible for real-time application and that suitable model parameters can automatically be inferred. Furthermore, the benefits for improvement of a distraction warning system were proven. Consequently, the methodology established in this thesis is a promising new approach for assessment of driver attention that avoids the issues of the current state-of-the-art.

Vorwort

Das vorliegende Werk entstand im wesentlichen während meiner Anstellung in der Gruppe für Nutzermonitoring und Nutzermodellierung (CR/AEU2 bzw. CR/AEY3) im Zentralbereich für Forschung und Vorausentwicklung der Robert Bosch GmbH. Im folgenden möchte ich mich bei allen bedanken, die direkt und indirekt zum Gelingen dieser Arbeit beigetragen haben.

An erster Stelle möchte ich mich bei Professor Rainer Stiefelhagen bedanken, der als Leiter der Gruppe Computer Vision for Human Computer Interaction (CV:HCI) am Karlsruher Institut für Technologie (KIT), meine Betreuung übernahm. Auch wenn die Arbeit in manchen Stellen in Richtungen ging, die nicht ganz in der Kernkompetenz der Forschungsgruppe lagen, so haben die regelmäßigen Diskussionen des Vorgehens und der Resultate wesentlichen Anteil an dem vorliegenden Werk. Genauso danke ich Professor Constantin Rothkopf, Direktor des Centers for Cognitive Science an der Technischen Universität Darmstadt, für sein Interesse an der Arbeit und seiner prompten Bereitschaft zum Koreferat. Darüberhinaus war die Zusammenarbeit und die Diskussionen mit Ihnen, Prof. Rothkopf, eine sehr interessante Möglichkeit das Forschungsgebiet der Fahreraufmerksamkeit aus Sicht der Kognitionswissenschaften zu betrachten.

Neben der wissenschaftlichen Betreuung von Seiten der beteiligten Professuren, wäre dieses Forschungsprojekt nicht ohne die Unterstützung der Robert Bosch GmbH möglich gewesen. Ich danke Dr. Andreas Korthauer als Projektleiter und Dr. Dietrich Manstetten als Chefexperte für die Mensch-Maschinen Interaktion und Abteilungsleiter für die Gewährleistung der Rahmenbedingungen für meine Tätigkeit. Ich konnte mich immer auf eure Unterstützung und der Bereitstellung von Projektressourcen verlassen. Genauso gabt ihr mir in den regelmäßigen fachlichen Rücksprachen viele hilfreiche Hinweise, die zum Gelingen der Arbeit beigetragen haben. Einen besonderen Dank möchte ich Dr. Hans-Joachim Bieg aussprechen. Ohne deine fachliche und wissenschaftliche Betreuung könnte ich sicher nicht auf die abgeschlossene Arbeit zurückblicken. Obwohl ich dir, Joachim, als Kuckuckskind untergeschoben wurde, hast du meine Forschungstätigkeit mit viel Zeit und großem Einsatz in den produktivsten drei Jahren unterstützt. Deine Akribie und dein unermüdlicher Rotstift haben die Gesamtqualität meiner Arbeit gewisslich wesentlich gesteigert. Der Humor ist dabei auch nicht zu kurz gekommen.

Darüber hinaus haben mich viele weitere Personen auf verschiedenste Weisen unterstützt. All den Genannten und Ungenannten möchte ich meinen Dank aussprechen. An erster Stelle danke ich den mehr als 70 Probanden und Probandinnen, die an meinen Fahrexperimenten teilgenommen haben. Sie haben sich leider nicht immer wie die Vorhersage meiner Modelle verhalten, aber ich konnte mich immer auf ihre Geduld, Hilfsbereitschaft und Humor verlassen. Ich danke Michael Herman für die gute Zusammenarbeit und die inspirierenden Diskussionen. Die gemeinsam Forschung an der inversen optimalen Steuerung bzw. des inverse reinforcement learning hat sich meiner Meinung nach für uns beide sehr gelohnt. Ein weiterer Dank sei den Studierenden Kathrin Bromberg, Hao Li, Annika Kaupp und Thakshak Shetty, die mich durch ihre Beiträge tatkräftig unterstützt haben. Marco Quander, Dietmar Martini, Andreas Zehender, Michael Schulz und Simon Hackenbroich danke ich für ihre Hilfe und dem geteilten Wissen bezüglich der Versuchsfahrzeuge und der verbauten Hardware. Genauso gilt ein Dank der gesamten Abteilung CR/AEU für das kollegiale Miteinander. Schließlich bin ich froh darüber, dass ich von den Mitarbeiter und Doktoranden in der CV:HCI Gruppe so freundlich aufgenommen wurde (wenn man von den Mordanschlägen im Mafia-Spiel absieht) und mit Rat und Tat bei organisatorischen Fragen unterstützt wurde.

Die Arbeiten an dem Forschungsprojekt hörten nicht mit dem Verlassen der Arbeitsstätte auf. Ich danke allen Mitbewohnern in der Franklinstraße dafür, dass sie mich in meinem Zimmer in Ruhe arbeiten ließen aber auch für mannigfaltige Ausgleichstätigkeiten zur Verfügung standen. Den Freunden in Karlsruhe, die mich beherbergt haben und von denen mir Dominic Detroit Techno für die nächtlichen Forschungsarbeiten näher gebracht hat, sei ein Dank. Genauso möchte ich mich für den Machine Learning Austausch in der Darmstädter Runde in Stuttgart bedanken.

Zuletzt möchte ich meiner langjährigen Gefährtin Barbara herzlichst danken. Ich bin zutiefst für dein Vertrauen in mich sowie deine Geduld und Unterstützung dankbar. Du hast Grammatik und Orthographie meiner Werke ins Gerade gerückt und standest mir bei Zweifeln und Widrigkeiten in den vergangenen Jahren geduldig zur Seite.

Diese Arbeit widme ich meiner Großmutter Cäcilia Hock. Sie war in ihrer späteren Heimat in Monbrunn in Nordbayern die erste Frau mit Führerschein und fuhr begeistert ihren Passat, genannt “Silberpfeil”. Liebe Großmutter, du hast die Bedeutung meines Forschungsgebiets treffen mit “Felix, ich war ja nur auf der Volksschule, aber ich weiß ganz genau: Beim Autofahren muss man alles im Blick haben" kommentiert. Ich bedaure, dass du die Fertigstellung dieser Arbeit leider nicht mehr erleben konntest.

Stuttgart, 31. August 2017 *Felix Martin Schmitt*

Contents

Nomenclature

Symbols

Latin Letters

Symbol	Description
$a, b, \ldots$	Unspecific objects (lower-case)
$\mathbf{a}, \mathbf{b}, \ldots$	Column vectors (lower-case, bold font)
$\mathbf{A}, \mathbf{B}, \ldots$	Matrices (upper-case, bold font)
$\mathsf{A}, \mathsf{B}, \ldots$	Sets (upper-case, sans serif)
$a : b$	Sequence of integers $a, a+1, a+2, \ldots, b \quad (a, b \in \mathbb{Z})$

Special Matrices

Symbol	Description
$\mathbf{I}^n$	Identity matrix $\mathbf{I}^n \in \mathbb{R}^{n,n}$, $\mathbf{I}^n = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & 1 \end{bmatrix}$
$\mathbf{0}^{n,m}$	All-zero matrix $\mathbf{0}^{n,m} \in \mathbb{R}^{n,m}$
$\mathbf{1}^{n,m}$	All-one matrix $\mathbf{1}^{n,m} \in \mathbb{R}^{n,m}$
$\mathbf{e}^n$	Indicator vector $\mathbf{e}^n \in \mathbb{R}^m, \quad \forall_{i \neq n} : e_i^n = 0,\ e_n^n = 1$

Operators

Symbol	Description
$x = (a, \mathbf{b}, \mathbf{C})$	Tuple x consisting of inhomogeneous elements
$\mathbf{A} = [\mathbf{B}\,\mathbf{C}]$	Horizontal concatenation of matrices $\mathbf{A}$, $\mathbf{B}$
$\mathbf{A} = [\mathbf{B};\,\mathbf{C}]$	Vertical concatenation of matrices, $\mathbf{A} = \begin{bmatrix}\mathbf{B}\\ \mathbf{C}\end{bmatrix}$
$\mathbf{a} = \mathrm{vec}(\mathbf{A})$	Vectorization of matrix $\mathbf{A} \in \mathbb{R}^{n,m}$ by stacking of columns, $\mathrm{vec}(\mathbf{A}) = \begin{bmatrix}\mathbf{A}_{1:n,1}\\ \mathbf{A}_{1:n,2}\\ \vdots\\ \mathbf{A}_{1:n,m}\end{bmatrix}$
$\mathbf{A} = \mathrm{diag}(a_1, a_2, \ldots, a_n)$	Construction of a diagonal matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, $\mathbf{A} = \begin{bmatrix} a_1 & 0 & \ldots & 0\\ 0 & a_2 & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & a_n\end{bmatrix}$
$\mathbf{A} = \mathrm{blk}(\mathbf{B}, \mathbf{C})$	Construction of a block-diagonal matrix $\mathbf{A} \in \mathbb{R}^{m_1+m_2,n_1+n_2}$ $\mathbf{A} = \begin{bmatrix}\mathbf{B} & \mathbf{0}^{m_1,n_2}\\ \mathbf{0}^{m_2,n_1} & \mathbf{C}\end{bmatrix}$ from matrices $\mathbf{B} \in \mathbb{R}^{m_1,n_1}, \mathbf{C} \in \mathbb{R}^{m_2,n_2}$
$\mathbf{P}^{\mathrm{blk}}_{n_x,n_u}(\mathrm{vec}(\mathbf{X}))$	Extracts diagonal blocks of a square matrix $\mathbf{X}$ in vectorized form If $\mathbf{X} = \begin{bmatrix}\mathbf{X}_{1:n_x,1:n_x} & \mathbf{X}_{1:n_x,n_x+1:n_x+n_u}\\ \mathbf{X}_{n_x+1:n_x+n_u,1:n_x} & \mathbf{X}_{n_x+1:n_x+n_u,n_x+1:n_x+n_u}\end{bmatrix}$ then $\mathbf{P}^{\mathrm{blk}}_{n_x,n_u}(\mathrm{vec}(\mathbf{X})) = [\mathrm{vec}(\mathbf{X}_{1:n_x,1:n_x}); \mathrm{vec}(\mathbf{X}_{n_x+1:n_x+n_u,n_x+1:n_x+n_u})]$
$\mathrm{tr}\,\mathbf{A}$	Trace of matrix $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{\min(n,m)} A_{i,i}$
$\det \mathbf{A}$	Determinant of matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, $\det(\mathbf{A}) = \prod_{i=1}^{n} A_{i,i}$
$\mathbf{A}^{-1}$	Inverse of invertible matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, $\det(\mathbf{A}) \neq 0$, $\mathbf{I}^n = \mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1}$
$\mathbf{A}^{+}$	Moore-Penrose pseudo-inverse of matrix $\mathbf{A} \in \mathbb{R}^{n,m}$, unique matrix satisfying $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$, $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$, $(\mathbf{A}\mathbf{A}^{+})^{\top} = \mathbf{A}\mathbf{A}^{+}$, $(\mathbf{A}^{+}\mathbf{A})^{\top} = \mathbf{A}^{+}\mathbf{A}$
$\mathsf{A} = \{\mathbf{B}_{1:n}\}$	Construction of set A from n matrices $\mathbf{B}_1$, $\mathbf{B}_2$, $\ldots$, $\mathbf{B}_n$
$\mathbf{A} = \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x})\vert_{\mathbf{x}=\mathbf{x}_0}$	Derivative of vector-valued function $\mathbf{f}(.)$ evaluated at $\mathbf{x}_0$, $A_{i,j} = \partial \mathbf{f}_i(\mathbf{x})/\partial \mathbf{x}_i\vert_{\mathbf{x}=\mathbf{x}_0}$ we will not differentiated between the derivative and its extension to non-differentiable functions
$\mathrm{softmax}(f(\mathbf{X}))$	Softmax operator evaluated on function $f(.)$, $\mathrm{softmax}(f(\mathbf{X})) = \log \int \exp(f(\mathbf{X}))\,\mathrm{d}\,\mathbf{X}$
$a \oplus b$	Logical xor on binary variables a, b $a \oplus b = 1$ if $((a = 1$ and $b = 0)$ or $(a = 0$ and $b = 1))$, else $a \oplus b = 0$
$\mathbb{I}_{\mathsf{S}}(\mathbf{X})$	Indicator functional $\mathbb{I}_{\mathsf{S}}(\mathbf{X}) = 1$ if $\mathbf{X} \in \mathsf{S}$, $\mathbb{I}_{\mathsf{S}}(\mathbf{X}) = 0$ else

Stochastics

Symbol	Description
$p(\mathbf{X})$	Probability distribution of random matrix $\mathbf{X}$ (this thesis employs the pragmatic approach common in the machine learning literature, where random variables are treated as discrete quantities. We note, that in case of a continuous random variable $p(\mathbf{X})$ mathematically precise specifies the probability *density* function.)
$\mathbf{X} \sim \mathcal{G}, p(.)$	Random matrix $\mathbf{X}$ distributed according to a special probability distribution $\mathcal{G}$ (upper-case, calligraphic) or probability distribution $p(.)$
$\mathbb{E}[\mathbf{F}(\mathbf{X})\|p]$	Expected value of matrix-valued function $\mathbf{F}(.)$ with respect to random matrix $\mathbf{X}$ with distribution $p(.)$, $\mathbb{E}[\mathbf{F}(\mathbf{X})\|p] = \int \mathbf{F}(\mathbf{X}) p(\mathbf{X}) \, \mathrm{d}\mathbf{X}$
$\mathbb{E}[\mathbf{F}(\mathbf{X})\|\mathsf{D}]$	Empirical expected value of matrix-valued function $\mathbf{F}(.)$ with respect to data $\mathsf{D} = \{\mathbf{X}_{1:n}\}$, $\mathbb{E}[\mathbf{F}(\mathbf{X})\|\mathsf{D}] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{F}(\mathbf{X}_i)$,
$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Random variable $\mathbf{x}$ distributed according to normal/Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
$\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Conditional normal/Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
$\mathbf{Y} \sim \mathcal{I}(\mathbf{X})$	Random variable $\mathbf{x}$ distributed according to Dirac distribution i.e. $\mathbf{Y} \equiv \mathbf{X}$
$\mathcal{I}(\mathbf{Y}\|\mathbf{X})$	Conditional Dirac distribution i.e. $\mathbf{Y} \equiv \mathbf{X}$
$x \sim \mathcal{U}_{\mathsf{X}}$	Uniform distribution over set X i.e. $p(x) = u, \forall x \in \mathsf{X}$

Algorithms

Symbol	Description
$\leftarrow$	Assignment operator
$\leftarrow^{+}$	Increment operator
$\leftarrow^{-}$	Decrement operator

Important Objects in Thesis

Symbol	Description
x_t	State of a Markov decision process
u_t	State of a Markov decision process
$r(x_t, u_t)$	Reward function of a Markov decision process
$\mathcal{P}(x_{t+1}\|x_t, u_t)$	Dynamics function of a Markov decision process
$\pi_t(u_t\|x_t)$	Policy in a Markov decision process
z_t	Sensory measurement of a partially observable Markov decision process
$p^z(z_t\|x_t)$	Sensory model of a partially observable Markov decision process
$b(x_t)$	Belief in a partially observable Markov decision process
$\pi_t^\star(u_t\|x_t)$	Optimal policy in a Markov decision process
$\tilde{\pi}_t(u_t\|x_t)$	Maximum causal entropy policy in a Markov decision process
$V_t^\star(x_t)$	Optimal value function in a Markov decision process
$Q_t^\star(x_t, u_t)$	Optimal state-control function in a Markov decision process
$\tilde{V}_t(x_t)$	Soft value function in a Markov decision process
$\tilde{Q}_t(x_t, u_t)$	Soft state-control function in a Markov decision process
$\varphi(x_t, u_t)$	Reward features of a Markov decision process
$\boldsymbol{\theta}$	Reward parameters of a Markov decision process $r(x_t, u_t) = \boldsymbol{\theta}^\top \varphi(x_t, u_t)$
$\mathbf{x}_t^{\mathrm{p}}$	Primary task state of the joint task POMDP
y_t	Vehicle's position in lane in the joint task POMDP
ϕ_t	Vehicle's orientation in lane in the joint task POMDP
α_t	Steering wheel angle in the joint task POMDP
$\dot{\alpha}_t$	Steering wheel velocity in the joint task POMDP
v_t	Vehicle's velocity in the joint task POMDP
κ_t	Lane curvature in the joint task POMDP
u_t^{p}	Primary task control of the joint task POMDP
x_t^{z}	Sensor state of the joint task POMDP
u_t^{z}	Sensor control of the joint task POMDP
x_t^{i}	Secondary task state of the joint task POMDP
u_t^{i}	Secondary task control of the joint task POMDP
$\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$	Sensor state sequence in the joint task POMDP
$\boldsymbol{\mu}_t^{\mathrm{p}}$	Expected primary task states of belief in the joint task POMDP
$\boldsymbol{\Sigma}_t^{\mathrm{p}}$	Covariance of primary task states of belief in the joint task POMDP
d_t	Eyes-off duration in the joint task POMDP

Abbreviations

Abbreviation	Description
ACC	Adaptive Cruise Control Driving assistance system that automatically controls the vehicle's headway
ADAS	Advanced Driver Assistance Systems
AGB	Appropriate Glance Behavior A normative model of glance behavior in driving
CAN	Controller Area Network Central communication network in a vehicle
def. by	defined by
EOD	Eyes-Off Duration Passed time steps since the driver's gaze was on the road for the last time
EOR	Eyes-On-Road algorithm Simple algorithm for assessment of driver attention
IOC	Inverse Optimal Control Obtaining the reward function underlying rational behavior
ISWYS	"I See What You See" Framework for obtaining the reward function as well as the sensor model underlying rational behavior in a (partially observable) Markov decision process
LQG	Linear Quadratic Gaussian partially observable Markov decision process A class of partially observable Markov decision processes in continuous variables that admits efficient solution
LQR	Linear Quadratic Regulator Markov decision process A class of Markov decision processes in continuous variables that admits efficient solution
MCE	Maximum Causal Entropy framework In this work referring to a model of rational decision-making
MDP	Markov Decision Process Decision theoretic problem class
OC	Optimal Control Obtaining a solution of a (partially observable) Markov decision process
POMDP	Partially Observable Markov Decision Process Decision theoretic problem class that involves fusion of noisy sensory measurements.
RMSE	Root Mean Squared Error Defined as $\sqrt{\mathbb{E}[e^2 \vert p(e)]}$ of a random scalar error variable e
SLQG	Sensor scheduling Linear Quadratic Gaussian POMDP A class of mixed discrete-continuous partially observable Markov decision processes
st.	subject to
STD	STandard Deviation Defined as $\sigma = \sqrt{\mathbb{E}[x^2] - \mathbb{E}[x]^2}$ of a random scalar x.
wrt.	with respect to

1 Introduction

Consider your commute to your workplace this morning. If you went by car, this might have been like this: After starting the engine and leaving the parking lot you take a brief glance at the radio to switch to your favorite station "Free Radio Stuttgart". You look back to the road to check for students who are on their way to school in the morning. To make sure to not get fined for speeding you finally have a short glance at the speedometer. Meanwhile you have probably switched your gaze a dozen times.

Frequent gaze switches are required because human vision is limited by various physiological constraints. To cope with the tasks of daily life such as driving to work, humans are required to focus their mental and particular perceptual resources to the important aspects of the current situation or the pursued task. This allocation of resources is generally referred to as attention [97]. Attention is of special importance in automobile driving. This is because manual vehicle control is a task with high demands on sensory processing and dynamic decision making. The driver has to monitor the course of the own vehicle and the instruments as well as the behavior of the other traffic participants. Based on his or her sensory measurements the driver then has to apply the correct steering and accelerator as well as brake movements in fractions of seconds.

Insufficient attention to the driving task, i.e. inattention, can rapidly lead to driving errors [253]. According to the U.S. National Highway Traffic Safety Administration (NHTSA) 25% of police-reported crashes are caused by inattention [183]. An analysis of the 100-Car Naturalistic Driving Study even concludes that approximately 80% of all crashes and 65% of all near crashes involve inattentive drivers as a contributing factor [108]. In crashes attributed to inattention typically the driver's attention was drawn away from the driving task by a *secondary task*, e.g. interacting with passengers, the vehicle's infotainment system or even a hand-held device. This phenomenon is commonly termed *driver distraction* [127]. In the last decade, the awareness for the risks imposed by driver distraction has spread among the public. As a result, several campaigns have been conducted to admonish drivers to concentrate attention to the driving task and to avoid engagement in especially distracting secondary tasks such as typing text messages. In some countries certain secondary tasks are prohibited by law and subject to fining. For example in Germany a driver who uses a hand-help phone while the vehicle's engine is running is fined 60 euros combined with a penalty point in the central database [1]. Despite these actions driver distraction remains an issue: A NHTSA publication recently revealed that 10% of all police reported fatal crashes and 15% of all injury crashes in 2015 were affected by driver distraction [63]. In this context it was noted that due to methodical issues distraction as a contributing factor is typically under-reported by a factor of at least 2 in the considered database. A study conducted by a major German insurance company using customer data of the years 2013 to 2016 revealed that engagement into many common secondary tasks in driving increases crash risk [116].

Despite its impact on road safety the mechanisms of driver distraction are not yet fully understood [127]. In this context it is common to categorize driver distraction into three types: *Visual distraction* is insufficient visual attention to the road scenery and the other traffic participants. This can be the case for example, if the driver is interacting with the vehicle's infotainment system which requires looking at an in-vehicle display. If the driver shows *inappropriate glance behavior* such as spending too much time without looking at the road scenery he or she can for example fail to notice that the vehicle is approaching the lane boundaries. Interacting with in-vehicle infotainment does not only require averting gaze but may also require removing one hand from the steering wheel to press buttons. As a consequence, the driver might not be able to conduct rapid steering movements required for evasive maneuvers. Hence, in addition to visual distraction *manual distraction* is imposed. Finally, distraction can also result from bound cognitive resources, e.g. during intense conversations, which is *cognitive distraction*. During such a conversation it could happen that the driver saw the flashing braking lights of a preceding vehicle but then fails to notice the intensity of the braking maneuver and to react accordingly.

Visual distraction is considered the most important type of distraction to deal with. For example, the analysis of the naturalistic driving study [108] sees visual distraction as the main culprit to crash and near-crash events. As safe driving requires predominately acquiring and processing of visual information [214], distinct negative effects of visual distraction on response to lead vehicle braking [48, 123] and lane keeping performance [229, 139] were established. In contrast, the influence of cognitive distraction on crash risk is less clear: Engagement in moderately cognitive demanding secondary tasks such as conversation with a passenger is associated with reduced crash risk compared to full attentive driving [108]. In addition to that, [55] found decreased deviation from the lane center during cognitive distraction while [139] established a delayed hazard response.

Engagement in secondary tasks is common in natural driving. For example, it was found that in 40% of the driving time drivers engage in a secondary task if a passenger is present [155]. In the case the drivers were alone in the vehicle still in 25% of the driving time such tasks were present. Many of these secondary tasks such as interacting with a navigation system have a utility for the driver. Consequently, it is not possible to ban every potential distracting secondary task. This is the case especially for secondary tasks that compete with the attention to the road scenery but which are required for safe driving. For example, drivers must monitor the vehicle's speed by means of the speedometer and must check the mirrors when planning lane changes. In both cases the gaze must be averted from the road scenery to the vehicle's interior.

Drivers want and need to engage in secondary tasks during driving. These tasks bear the risk of dangerous distraction, especially, if they require averting gaze from the road. This is because of the negative impact on driving performance which can favor crashes. Therefore, it is desirable to develop systems that can assist the driver in safely interacting with secondary tasks. Such systems have the potential to greatly improve road safety. Here, an important component is an algorithm for assessment of the driver's visual attention to the forward road scenery. Based on this assessment the driver could for example receive a distraction warning if insufficient attention is detected. The present thesis contributes to the state-of-the-art in automatic driver attention assessment for real-time distraction warning. This will be done by development of new algorithmic approaches and their validation by means of real driving experiments. Due to its clear and strong relation to decreased driving performance, we will focus on distraction induced by inappropriate glance behavior. That is, we will focus foremost on visual driver distraction.

1.1 State of Research and State of the Art in Driver Distraction Mitigation

After introducing the issue of driver distraction, we will review the literature wrt. current state-of-the art in mitigation of distraction. In this context two facets are considered: We will first address the research on how drivers deal with secondary tasks in driving that compete for attention. Here, especially those works are considered that focus on driver strategies of secondary task interaction and gaze arbitration. Second, technology and systems are reviewed that aim at mitigating driver distraction and its effects. This includes especially approaches for the automatic assessment of driver attention. The purpose of this review is to provide a basis for identifying the current research gap which will lead to the problem statement of this thesis.

This section will not list all the publications that are related to the models and algorithms developed or the experiments conducted in the course of this thesis. This is because much of the previous work that the present thesis is based on is originally unrelated to the subject of driver distraction. Instead, throughout this work each individual chapter will have a separate short review of related publications.

1.1.1 Driver Strategies for Distraction Mitigation

In the previous section, we have explained that driving is a task that poses high demands on sensory processing and consequently misdirected attention can have fatal consequences. However, arbitration of attention and its coordination with actions is successfully employed by humans in many daily activities [76]. Here, human attention arbitration, especially glance behavior, is adapted to the present task [193]. Therefore, it is necessary to review the literature on driver strategies regarding glance behavior and engagement into secondary tasks. If these human strategies can mitigate distraction, then

they should be incorporated in automatic attention assessment. Furthermore, experimental studies help to understand which aspects of driving in presence of a secondary task have the strongest impact on driving performance which should also be considered in algorithmic approaches.

One of the pioneering works on driver strategies with respect to attention is Senders et al.s' work [207]. Here, the authors studied the interaction between the duration of voluntarily chosen occlusion of the forward road scenery and the driving speed. It was shown that given fixed duration of occlusion the drivers adapt their driving speed. That is, the longer the occlusion the lower the chosen driving speed. Conversely, drivers adapt the maximum occlusion time given a certain driving speed in the form that smaller occlusion times are tolerated at higher driving speed. In naturalistic driving familiarity with the road as well as driving experience were found to result in spare capacities of visual attention that can be used to monitor other targets than the road scenery [160, 161]. [37] investigated eye-movement patterns in simulated driving. Here it was shown that increased difficulty of the driving situation came with a stronger concentration of gaze on the road scenery. Furthermore, voluntarily chosen occlusion time in a lane keeping task is closely related to the remaining time until lane departure [69, 68]. This was interpreted as dedicated strategy of attention arbitration. [84] showed that drivers' visual attention is distributed across the driving task and interaction with in-vehicle technology according to the associated information bandwidth, relevance, priory as well as sampling effort. Similar results were found in [228]. Here, driver glance behavior was adapted with respect to incentivation and uncertainty in quantities relevant for either headway control or speed control. More recently, the voluntary chosen occlusion distance, i.e. occlusion duration time divided by the driving speed, was studied [119]. In that work an experiment in a high fidelity driving simulation revealed that the occlusion distance reflects the attentional demands of different driving scenarios such as sub-urban, rural and highway driving environments.

Besides the general strategies of attention arbitration in driving also the effects of the specific structure of a secondary task have been studied. [35] investigated strategies for typing American-style telephone numbers during lane keeping. Here the presentation of numbers in chunks of three resulted in task-interleaving at the boundaries of the blocks. UK-style telephone number are built of blocks of five digits. Hence, in this case block boundaries are too distant to be useful for interleaving [89]. Instead, interleaving strategies were found to be highly sensitive to the performance objective set by incentivation. Furthermore, the occurrence of errors and the possibility to correct in a typing task during driving influences glance behavior [129]. Here, it was shown that detecting a typing error in the secondary task favored gaze return to the road but was dependent on how quickly the drivers obtain feedback on the correctness.

Several studies on drivers' interaction with secondary tasks in naturalistic driving have found evidence for drivers employing dedicate strategies to avoid distraction. For example, engagement in visually demanding secondary tasks is often accompanied by reduced driving speed [12, 180, 147, 80, 162] or increased headway to preceding vehicles [87, 41, 162]. Therefore it was hypothesized that drivers actually apply a "deciding to be distracted" approach [132]. To what extent drivers employ strategies in interaction with a secondary task was investigated in several works: In [205] a driving-simulator experiment was conducted where the participants had the possibility to schedule their engagement in potentially distracting secondary tasks. It was shown that drivers decided when to interact with a task and adapted both their driving and glance behavior to the demands of the driving situation. Driver behavior in an externally-paced secondary task and a self-paced secondary task were compared in a driving simulation in [156]. In the case of the self-pace task glance behavior and engagement in the tasks were highly adapted to the situational demands. Similar results were obtained in an experiment on a test-track [137] and in the analysis of a naturalistic driving study [239]. [247] investigated the effects of externally-paced and self-paced secondary tasks on driving performance. The results demonstrated that scheduling of engagement in the potentially distracting secondary tasks allowed the drivers to prevent large decrements in driving performance. Externally-paced tasks resulted in increased variability of the position in lane as well as increased variability in the time head-way to preceding vehicles. In contrast to the other works in the driving experiment of [81] drivers generally did not strategically postpone engagement in distracting secondary tasks although the demands of the upcoming track were known. Hence, it was hypothesized that drivers can sometimes fail to appropriately assess the demands of the driving situation [82].

In summary this short survey shows that drivers employ strategies for attention arbitration, glance behavior, and engagement in secondary tasks. In addition to that, these strategies are adapted to the secondary task and the demands of the driving task. Furthermore, a coupling with driving behavior was established. Most importantly, this type of driver behavior was found to be rational as it allows to mitigate the effects of distraction and can ensure a certain level of driving performance.

1.1.2 Technology for Distraction Mitigation

Although, the strategies for attention arbitration can mitigate distraction, the facts and figures on the contribution of driver distraction to crash risk [108, 63] indicate that drivers either often fail to apply such strategies or that the applied strategies are not always effective. Therefore, automatic systems that mitigate the effects of distraction or help the driver to maintain sufficient attention are desirable. In the present section we therefore report on such systems namely advanced driver assistance systems, automated driving systems and workload managers. Alternative approaches for automatic driver attention assessment other than the one pursued in the present thesis are listed separately (see Sec. 1.1.3). In this context, we wish to note that a variety of commercial systems for mitigating distraction and its effects have been developed by both car makers and suppliers. However, we will only review systems where a scientific publication regarding their functionality or an evaluation is available.

Advanced Driver Assistance Systems

Based on the observation that human error plays a dominant role in the occurrence of crashes [253], several Advanced Driver Assistance Systems (ADAS) have been proposed that aim at supporting the user in the driving task. Instead of addressing the underlying causes of driving errors these systems instead try to prevent these errors to lead to crashes. In this context, ADAS address both lateral and longitudinal vehicle control.

Lane keeping assistance systems and lane departure warning systems aim at improving lateral vehicle control. A camera system is employed to track the road boundaries and to estimate the vehicles position in the lane [188]. Based on the estimated position lane keeping assistance systems apply a steering torque to help the driver to stay inside the lane boundaries [210]. Lane departure warning systems instead trigger a warning if an imminent crossing of the lane boundaries is anticipated [189]. These systems have generally been found to be effective [7]. Especially, the lane keeping performance of distracted drivers can be improved by an early warning [28].

Forward collision warning systems and automatic emergency brakes can mitigate collision with preceding vehicles. In both systems a sensor monitors the environment in direction of travel of the vehicle. For each obstacle in the driving corridor the collision risk is assessed [96]. Forward collision warning systems trigger a warning at an early stage if the risk is becoming critical. This is required to ensure that sufficient time is left for driver reaction. Instead, autonomous emergency brakes trigger a brake intervention typically at the last possible moment [130]. Similar as in case of assistance systems for lateral vehicle control forward collision warning systems can improve driving safety [49]. [125] demonstrated that early forward collision warnings can redirect the driver's gaze back to the road and thus prevent collisions.

Although ADAS can provide significant safety benefits by mitigating the effects of driving errors, a common problem are false interventions. For example in [7] users reported the lane departure warning system to be annoying. This was the case when drivers deliberately steered the vehicle out of the lane, for example in case of construction works or for lane changing. Furthermore, distracted drivers need earlier warnings to react to an impending threat [125, 28]. However, automatic threat assessment is typically uncertain in early stages of a critical situation due to sensor noise and uncertainty wrt. the evolution of the situation [223]. Consequently, there is a significant risk of false interventions which would be especially annoying for a fully attentive driver. Hence, conventional ADAS could significantly improve through additional assessment of the drivers' attention: [179, 122] proposed ADAS systems for lateral vehicle control that featured assessment of the drivers attention and adapted warnings and intervention. Forward collision warning systems using threat assessment that considers both driving situation and driver attention state were presented in [187, 237].

Automated Driving Systems

A more radical way than mitigating the effects of driver errors is to remove the human from vehicle control in the first place. Indeed, fully autonomous cars could significantly reduce the number of crashes by not suffering from the driving errors humans are prone to. Although in the last years significant progress has been made towards fully autonomous cars a lot of technical and legal issues remain yet unsolved [24]. Automated driving systems that are currently available on the market or that are in series development are partial or conditional automated systems (see [44] for a taxonomy and precise definitions). In partial automation the driver has to continuously monitor the system to detect failures and intervene accordingly. Consequently, in partial automated driving driver distraction is similarly problematic as it is in case of manual driving as known in the literature on supervisory control [208]. In conditional automation the user is not required to monitor the system continuously but must be capable of resuming control if system limits are reached. Typically, take over by the driver in a certain time span is required. In this context driver distraction maintains relevance as it has been shown that engagement in distracting secondary tasks can significantly reduce take-over quality [254]. Summarized we conclude that the assessment of driver distraction remains an important issue in automated driving until finally full autonomy is obtained.

Workload Managers

"Driver distraction is the diversification of attention from the driving task to another task" [127]. Following this definition also the interaction with information technology build into the vehicle can be distracting. For example, an analysis of insurance data revealed significantly increased crash risk for interaction with several in-vehicle technologies [116]. Furthermore, it has been shown that usage of built-in navigation systems can result in decreased driving performance due to distraction [48, 147]. However, the research on driver strategies in scheduling secondary tasks indicates that the effects of distraction can be mitigated if engagement in such tasks is scheduled to low demand driving situations [247].

The idea of *workload managers* is to realize a system for automatic scheduling and prioritizing the information presented to the driver to avoid conflicts with the demands of the driving task. For example certain information could be delayed in a demanding driving situation such as a complicated lane merge. Alternatively, some demanding tasks such as destination entry in a navigation system could be blocked. Several concepts for workload managers have been described in the literature [177, 33, 10]. Notably, also workload managers for smartphone usage during driving have been proposed [119]. The benefits of automatic strategies for blocking engagement in secondary tasks and workload managers have been demonstrated: [177] showed that workload management can decrease overall objective and subjective workload. Furthermore, lane keeping and headway keeping can be improved by blocking secondary tasks in high demand driving situations [51]. The system proposed in [10] resulted in a higher proportion of driver gaze on the road and decreased subjective workload. In [119] a positive effect on glance behavior of secondary task blocking and high perceived usefulness of the workload management system were established.

Workload managers can already provide benefits for the driver if they only assess the driving situation and schedule information as well as block certain secondary tasks. However, increased effectiveness can be obtained if also the driver's attention state and his or her glance behavior in interaction with the information system is considered [10, 119]. In addition to that, workload mangers require knowledge of the secondary tasks the driver is engaging in. Interaction with the vehicle's infotainment system can easily be detect, however this is not the case for other secondary task such as typing on a hand-held smart-phone or reading billboards.

1.1.3 Automatic Assessment of Driver Attention

In the previous review of the technology for distraction mitigation it was revealed that many of the proposed systems could significantly benefit from an assessment of driver attention: The acceptance and effectiveness of ADAS could be improved, sufficient monitoring of partial automated driving and

take-over readiness could be ensured and the specificity of workload managers could be increased by assessment of driver attention.

Furthermore it is also possible to provide a feedback to distracted drivers. That is, if the driver was found not being attentive enough to ensure an acceptable level of safety he or she could be warned to for example interrupt a present secondary task and to return his or her gaze back to the road. Such systems have for example been proposed in [51, 104, 126].

A variety of approaches have been proposed for automatic assessment of driver attention. This section serves to provide an overview over the techniques employed and the approaches proposed in this context. An alternative review can also be found in [50]. Generally, the approaches for driver attention assessment proposed in the literature can be differentiated with respect to the data sources that are used. In this context, driving behavior statistics, statistics of glance behavior and eye movement, measures of brain activity as well as combinations thereof have been employed which we will use to categorize the different works.

At this point, we wish to point out that although many of these works propose to assess driver attention or detect driver distraction actually surrogate classification problems are considered. Specifically, often the presented approaches detect engagement into potentially distracting secondary tasks because a reference for problematic driver distraction is hard to obtain. We will discuss the issues with this approach in Sec. 1.2 and will consequently state the quantity predicted in the reviewed publications.

Unfortunately, most approaches for driver attention assessment were not evaluated on a common data set. Instead, many works use self-collected data sets that significantly varied in the design of the conducted driving experiment. For example, some studies used data obtained in simulated driving and some studies employed data obtained in real traffic. Furthermore, the sensors, e.g. for eye-tracking, employed and their measurement accuracy differed in the individual works. Consequently, obtained prediction performances are of limited comparability between publications. For this reasons we omit to report figures of prediction and classification accuracy in the review.

Driver Attention Assessment from Driving Behavior

Engagement into visually and cognitively demanding secondary tasks during driving can result in characteristic decrements of driving performance [123, 55, 139] as well as compensation strategies [41, 80]. This allows to conversely detect the corresponding periods from measures of driving behavior.

In [241] the quantities obtained by a forward collision warning systems were fed into a random forest classifier for detecting presence of a distracting secondary task. [58] first fitted a multi-layer perceptron to the driver's speed, acceleration and throttle press profile. Presence of a secondary task was then classified based on a support vector machine using the residuals of the neural network model. Driver behavior as measured in series vehicles such as steering angle, driving speed, pedal position as well as the distance to preceding vehicles and the lane borders have been evaluated in [199, 235]. Here a variety of different classification techniques including neural networks, support vector machine and fuzzy logic approaches was employed. In contrast to the other works that used classification techniques from the machine learning domain in [78] a control theoretic driver model was used for assessment of driver attention. That is a neuro-muscular model of steering control was fitted to behavioral data and it was shown that the fitted model parameters were highly sensitive to engagement in secondary tasks during driving.

The advantage of the approaches that build only on statistics of driving behavior alone is that the sensors necessary for classification are already available in modern series cars. Consequently, no additional costly sensors are required to implement the driver attention assessment.

Driver Attention Assessment from Head and Gaze Movements

Drivers engaged into secondary tasks show characteristic driving patterns. These patterns result from diverted attention to the driving task [253]. In the context of visual distraction this diversion of attention can also directly be assessed. This is possible by analyzing the drivers' head and gaze movements. For example, an analysis of the 100 cars naturalistic driving study revealed that a total time of gaze off the forward road scenery of 2 s in a 5 s window was associated with significantly increased relative crash risk [108]. In the last decade several computer vision techniques for estimation of the human head

pose [163, 237] and eye tracking as well as gaze direction [75, 79] have been developed. In addition to that several commercial products are available. Consequently, several approaches have been proposed that assess driver attention from measured glance behavior or head movements although in-vehicle eye-tracking is not yet perfect [5].

In [52] a distraction warning system was proposed that assessed attention from both the proportion of gaze off the road in the last 3 s as well as the duration of a current glance off the road. The 1.5th power of the duration of the current off road glance was employed in [179]. In [105] a distraction index was proposed that was incremented when the driver's gaze was off the road and decremented when the driver's gaze was on the road. Furthermore, hystereses as well as different increment rates dependent on the angular amount of gaze aversion from the forward road scenery were used. Variants of the previously listed algorithms were investigated in [141]. The approaches for driver attention assessment based on driver behavior were built on classification models estimated by applying machine learning approaches to collected data. Here the models were optimized with respect to predicting periods of engagement into distracting secondary tasks. In contrast, it is not clear how the algorithm parameters of the previously listed approaches which use statistics of glance behavior were obtained. An evaluation of several algorithms with respect to predicting relative crash risk was conducted in [141]. Furthermore, [126] reported on an evaluation with respect to predicting the periods of secondary task engagement.

As shown in [245] also cognitive distraction can manifest itself in the driver glance behavior: Under cognitive load the drivers' gaze is stronger concentrated to the center of the road. This phenomenon was utilized in a glance behavior based algorithm for detecting engagement into cognitively distracting secondary task which was described in [126]. Support vector machine and extreme learning machine approaches were employed to detect engagement into cognitively distracting secondary tasks from glance behavior [146]. In this context, it was shown that classification accuracy can benefit from utilizing a semi-supervised learning technique. Recently, a distraction warning system was proposed that used the duration of the off road glance divided by the driving speed as a driver distraction index [119] which was inspired by the driving experiment of [120] mentioned earlier.

The work of [153] indicated, that head movements alone may also allow for assessment of driver attention. In contrast, a significantly larger on-road study conducted in [64] showed that there is strong variation in the amount of correlation between drivers' head movements and eye movements. Consequently, the head pose can substitute an estimate of the drivers' gaze direction only for a subset of the driver population.

In contrast to driver attention assessment from driving behavior an additional sensor is required to measure glance behavior which is a disadvantage. However, driver attention assessment from glance behavior has the advantage of being applicable also in partial and conditional automated driving where no driving behavior of the user is available.

Driver Attention Assessment from Brain Activity

Engaging in a secondary task during driving requires mental effort by the driver. In addition to the mental resources needed for the driving task demands arise from the secondary task and from task interleaving. Consequently, some researchers were successful in assessing driver attention from brain activity. In this context, typically electroencephalography (EEG) was employed. This is an electrophysiological approach which measures voltage dynamics in the brain by means of multiple electrodes placed on the scalp.

[145, 144] used independent component analysis and clustering to identify frequencies band which are sensitive to occurrence of a secondary task in simulated driving. The usage of so-called alpha-spindle features [213] for detecting visual and cognitive distraction was studied in [217]. It was shown that these features can discriminate all three attentive driving, engagement into cognitively and visually distracting secondary tasks using data of simulated driving. In [218] it was shown that alpha spindle features allow for detecting the presence of a cognitive secondary tasks in driving on a test-track with a low average error but significant variation with respect to the individual participants. The potential of different frequency bands and different measurement regions on the scalp for assessment of cognitive distraction was studied in [8]. Furthermore, [248] proposed an adaptive threshold prediction

framework to detect the begin and the end of distracting secondary tasks. In this work several features were constructed from statistics of the different frequency bands and a feature selection approach was employed.

Monitoring of brain activity using EEG has the advantage of more directly measuring the human cognitive processes involved in cognitive distraction. In contrast, other approaches, e.g. using driving or glance behavior, employ indirect measures. Furthermore, [218] demonstrates that EEG based detection of cognitively distracting secondary tasks is robust wrt. the conditions of real driving. However, EEG measurement is highly intrusive to the drivers as it requires a special electrode cap. Consequently, EEG-based detection of driver distraction is currently not suited for a product distraction assessment system.

Driver Attention Assessment by Fusion of Data Sources

The previous sections showed that it is possible to detect engagement in secondary tasks during driving by means of different data sources. While these approaches have the advantage of potentially requiring only a single sensor for estimating the driver state, fusing the different sources has the potential to increase both robustness and effectiveness. Consequently, such an approach has been employed in several works.

[138, 140] used a dynamic Bayesian network for detecting engagement in a cognitively distracting secondary task. Here, both glance and blink statistics as well as measures of driving performance such as standard deviation of lane position in simulated driving were considered. Similar features and support vector machine classification as well as logistic regression were employed in [142]. Notably, different "definitions of driver distraction" were considered: The methods were evaluated for both predicting engagement in a secondary task and predicting the intervals where the drivers showed their worst 25% of driving performance. Here, predicting the periods with presence of the secondary task was most accurate. [157] employed adaboost to detect engagement into two different cognitive tasks during simulated driving. For classification features of the drivers' pupil diameter, glance behavior and heart rate were constructed resulting in higher prediction performance compared to a support vector machine based approach. A system for maneuver and driver specific detection of engagement in secondary tasks was presented in [198]. Here, first the driver was identified by means of audio features, then the current maneuver was recognized by means of a naive Bayes approach using a Gaussian mixture model based on driving behavior measures. Finally, a second driver and maneuver specific naive Bayes classifier was used to detect presence of a distracting secondary task. [250] proposed a long-term-short-term neural network for detecting visual-manual interaction with the vehicle's infotainment system in real driving. In that several features of the driving behavior as well as the drivers' head movements were constructed and fused by means of the neural network. In comparison to a support vector machine based approach improved detection performance could be established which was attributed to the long-term-short-term neural network's capability to account for the temporal driving context. A data set comprising of engagement into several distracting secondary tasks during real-traffic driving was introduced in [88]. In that work self-assessment of distraction by post- driving questionnaires as well as rating of distraction by external assessors using video snippets of a driver facing camera were employed. Furthermore, features of driving behavior, driver's eye glance behavior as well as acoustic features obtained by a micro phone array were analyzed wrt. sensitivity to the secondary tasks. This data set was used in [136] to evaluated k-nearest-neighbor and support vector machine classification to detect the periods of engagement in secondary tasks as well as to discriminate between the different tasks. Furthermore, linear regression and support vector regression were employed to predict the ratings of distraction made by external raters. In addition the features extracted in previous works also statistics derived from facial action units were employed. [135] revisited the data set. In contrast to previous work here visual and cognitive distraction were separately assessed using video snippets of both the driver and the forward road scenery. Furthermore, additional features for prediction were obtained from the video road scenery. Several approaches to predict ratings of distraction as well as to classify into high and low perceived distraction were evaluated. The interaction of glance behavior and steering behavior were analyzed in [251]. It was shown that auto correlation of the individual quantities as well as correlation between the quantities can differentiate between the

three classes attentive driving, engagement in a visually distracting secondary task and engagement into an cognitively distracting secondary task in simulated driving. A comparison of the detectability of engagement into cognitively distracting secondary tasks in different simulated driving scenarios was conducted in [143]. Using support vector machine classification with automatic feature selection similar high classification quality in highway driving and approaching a stop-controlled intersection could be obtained. However, the identified most important features for classification varied between the considered scenarios. Furthermore, significant differences in cognitive load measures according to ISO/DIS 17488 and ISO 15007-2014 between the scenarios could be established.

In many of the reviewed works on driver distraction assessment by fusing diverse sensor modalities feature selection techniques have been employed. Here, the most important features were typically from different sources [250, 135]. This demonstrates that fusion approaches have the potential to improve prediction accuracy. While many of the approaches relied on features obtained from eye-tracking which would require an additional sensor, fusion techniques can be more robust wrt. low sensor quality than approaches solely based on glance behavior. For example in [250, 135] only the drivers' head orientation instead of precise eye-tracking was required.

1.2 Problem Statement

In the present section, we discuss the literature on driver distraction mitigation. This allows us to identify the gap in current research of automatic assessment of driver attention which we will use to define the problem statement of this thesis.

Several works on algorithmic approaches to assessing attention and detecting distraction used the driver's engagement into potentially secondary tasks as a proxy. This is the case for methods that use a statistical classifier that was trained to distinguish between periods of engagement and baseline driving based on features related to driver behavior, e.g. [138, 250]. The same approach has also been used to benchmark and to optimize decision rules on the duration and frequency of glances away from the road [126]. Detecting whether the driver engages in a secondary task has the attractive property that large amounts of sample data can comparably easy be obtained. This is possible by conducting a driving study where the driver or an additional instructor manually logs the periods of these tasks. However, this convenience comes at a cost. First, drivers want to engage into secondary tasks - that is why they paid for an in-vehicle infotainment system. Hence, they should not generally be considered distracted any time they conduct a secondary task. Second, whether or whether not the secondary task poses risks for driving safety is strongly dependent on the driving situation and characteristics of the task. For example, typing text messages on a smartphone is safe when the vehicle is in standstill and highly risky at a significant driving speed. Furthermore, reading a map placed on the adjacent seat is more dangerous than monitoring the speedometer, although both require averting gaze from the road. This is because those secondary tasks differ in the amount of information that can be obtained from the road scenery by means of peripheral vision.

Humans with driving experience are to some extend able to judge the risks to driving safety induced by different secondary tasks. Hence, a model for assessing attention can be trained to predict a human's subjective assessment [136, 135]. This is realized by providing a data set of features of the driver behavior and associated human post-hoc ratings. Such an approach improves over merely detecting secondary tasks. This is because behavioral patterns are related to subjective risk of distraction and therefore a fine grained differentiation is obtained. However, in the end the model rather predicts the demands of secondary tasks than the associated risks: A driver showing shaky steering movements and reducing the driving speed strongly indicates that he or she is engaging in a demanding secondary task. Still, this can be of small accident risk if the driving speed is low and therefore safe driving is not very demanding. Furthermore, subjective ratings can be ambiguous: Raters may come up with different judgments on the demand of a specific task based on their own frequency of engagement during driving and their personal risk attitude. For example in [135] subjective ratings had a correlation of 0.70 - 0.77 when comparing the rating of an individual rater to the average of four other raters.

The literature review showed, several works have found pronounced driver compensation strategies when engaging in secondary tasks. Hence, attention assessment algorithms trained to detect these tasks or assess their demand will likely be sensitive with respect to the occurrence of compensation

strategies. For example, a classification algorithm will predict the driver engaging in a secondary task if the driver deliberately reduces his or her driving speed. In contrast, human factors research concluded that reducing speed is actually a rational strategy [205]. This behavior serves to decrease the demand of the driving task and therefor reduces crash risk. Approaches based on both previous methodologies consider only statistical correlations between behavior patterns and the engagement in secondary tasks. Therefore, distraction may be predicted with high confidence if the driver is employing such a compensation strategy for distraction mitigation. Clearly this is not desirable.

Models for assessment of attention can also be obtained in a different way. This is possible by optimizing the model to discriminate between crash and non-crash events based on the drivers' behavior given the same situational circumstances [141]. Clearly, this approach is most directly related to the goal that shall be achieved. That is, to detect the type of behavior related to attention that has a high risk to result in a crash. However, obtaining the required amount of crash data is extremely cumbersome. Therefore, this approach is not feasible for developing a distraction assessment system based on a specific sensor configuration or a specific vehicle. A similar yet more practicable idea was proposed in [142]. Here, the periods of interaction with an secondary task in which the driving performance metrics were below the participants' individual 0.25 quantile were predicted. While this methodology is feasible for product development of a attention assessment system the choice of the specific quantile is somewhat arbitrary. In addition to that, in both approaches the influence of the situational circumstances on the crash risk or the risks associated with decreased driving performance are neglected.

Humans have been found to show highly adaptive behavior for attention arbitration [193]. Especially, several experiments showed that driver's glance behavior is rational with respect to the demands of the driving situation [207, 69], the characteristics of a secondary task imposed on the driver [34, 129] and preferences or incentives [89, 228]. Furthermore, similar adaptive behavior has also been found in engagement in secondary tasks in naturalistic driving [155]. Summarizing, drivers can to some extent assess the attention demands of the current driving situation and the desired secondary task as well as adapt glance and driving behavior correspondingly to mitigate distraction. Consequently, any distraction warning system that does not consider the context of the driving situation cannot provide optimal assistance and is possibly not considered useful by the drivers.

Based on the preceding discussion of the literature we conclude:

Engagement in potentially distracting secondary tasks during driving requires glance strategies that consider the characteristics and the demands of the driving situation and the secondary task. Drivers are aware of this relation and apply corresponding adaptive behavior. This is not considered by the current state-of-the-art in automatic attention assessment and distraction detection. Hence, current distraction warning systems are neither optimal in terms of effectiveness nor usefulness.

We acknowledge that this research gap has also recently been recognized and discussed in [106] in 2016. That work considers the issue from the perspective of traffic psychology and human factors research and proposes the theoretical concept of minimum required attention. In contrast, this thesis aims at improved attention assessment algorithms that can be used in a real-time distraction warning system. Hence in this context, new mathematical models and algorithmic methodology are required, which are not provided by [106].

In addition to the theoretical framework of [106], heuristic approaches for situation adaptive attention assessment have been employed in [62, 119]. In those works the time passed since the driver's gaze was on the road for the last time was divided by the squared driving speed [62] or the absolute value of the driving speed [119]. The approach of [62] was neither motivated nor evaluated. The heuristic employed in [119] was supported by empirical evidence regarding driver glance behavior in the driving experiment of [119], but its benefits were not specifically evaluated. Both approaches have the disadvantage that they do not establish a mathematical relation to neither decreased driving performance nor crash risk. Consequently, it is not clear which aspects of the problem of appropriate glance behavior in driving in presence of a visually demanding secondary task are accounted for in the algorithms.

Based on previous discussion of the literature and the identified research gap we obtain the following research questions that need to be addressed to improve distraction warning systems:

1. *How can a normative model of glance behavior for engagement in secondary task during driving be obtained, that can be used in a real-time warning system?*
   In order to avoid the insufficiency of previous work the causal relations between drivers behavior and the resulting risks must be considered in automatic driver attention assessment. That is, the model must reproduce how glance patterns lead to specific vehicle control. Furthermore, the resulting accident risk in the current driving situation must be modeled in an appropriate measure of control performance. Finally, the influence of the secondary task on glance behavior must be characterized. Given a model of the aforementioned causal relations, the second step is to compute normative glance behavior. Here, the required properties of the glance behavior are small loss of control performance while not too strongly conflicting with the driver's interest in engaging into a secondary task. Throughout the development of the model of normative glance behavior a good compromise between realism of the computed glance behavior and the computational demand that needs to be feasible for a real-time warning system must be found.

2. *How can suitable model parameters be found?*
   As the normative model of glance behavior tries to address complex real world behavior, it will likely contain several adjustable parameters. For the purpose of using the model in a distraction warning system those parameters must be set to appropriate values. Note, that the model of glance behavior is of normative character. Hence, the chosen parameter values must result in behavior that is accepted by real drivers. Furthermore, the parameters of the individual components of the model should closely match the relations present in real driving. This is a challenge in cases where the components relate to the driver perception as corresponding neuro-biological processes cannot be measured. Hence, new methodology to obtain the parameters with normative function and those related to the driver's sensor characteristics are required.

3. *How can a prototypical distraction warning system based on the normative glance model be developed?*
   Applying the fully specified normative model of glance behavior in a real vehicle poses additional challenges. First, the computation time needs to be reduced to a minimum. Second, we need to robustly deal with noisy sensor inputs. Hence, tailored pre-processing routines must be developed. Finally, the normative model needs to be coupled with a suitable warning system.

4. *Does the proposed model improve a distraction warning system?*
   From the conceptional perspective a normative glance model with the mentioned properties is a great improvement over the state-of-the-art attention assessment approaches. However, it is also necessary to critically evaluate if the new methodology indeed improves a final distraction warning system. This evaluation should be done in real driving to also test the robustness of the proposed approach. In this context, the relevant measures to be evaluated are both system effectiveness and acceptance by potential users.

1.3 Thesis Concept and Outline

This thesis addresses the identified research questions at the exemplar driving task of lane keeping. This is an elementary driving task the driver must address during the entire driving time with very few exceptions. Surely, lane keeping is only a single facet of driving in total and there are other more complex tasks such as lane changes. Note, that we need to develop normative mathematical models of glance behavior in those driving tasks. In contrast to the development of descriptive models, this is a problem that has only scarcely been addressed in previous research. Focusing on lane keeping allows us to investigate all the research questions in detail, which are challenging already in this driving task. Furthermore, insights are gained that are relevant beyond the considered exemplar task and that open up new research questions. Finally, the new methods and algorithms developed throughout this thesis provide a basis for future research on normative glance behavior in other driving tasks.

In the following we give a summary of the individual chapters of this work.

Cpt. 2 reviews the decision theoretical frameworks of Markov decision processes and partial observable Markov decision processes that form the mathematical basis of our normative model of glance behavior. Here, we consider the definitions and fundamental properties of these frameworks as well as

the techniques for computing optimal policies therein that are relevant for the present work. Finally, an extensions of the decision theoretic frameworks that account for imperfect behavior, i.e. sub-optimal policies, as e.g. required to model realistic driver steering are briefly introduced.

Cpt. 3 develops and validates the normative model of appropriate glance behavior in secondary task interaction while driving. This is done defining appropriated glance behavior by means of optimal and rational policies in a partially observable Markov decision process. We build this model by first considering the task of vehicle control under the external influences of track topology and driving speed. Thereafter, practical models of the driver's sensor characteristics and the potentially distracting secondary task are added. Importantly, for all three aspects, the driving task, the secondary task as well as the driver's sensing, individual performance measures, i.e. reward functions, are developed. Having obtained the joint task model and the different reward functions, we address its solution with respect to optimal and more realistic rational policies therein. Here, new solution algorithms are developed. Finally, these techniques and the resulting normative glance behavior are validated. This is done with respect to the realism of the glance behavior and the feasibility of the computational demands for a warning system.

Given the joint task model and the solution techniques, we need to parameterize the reward functions of the individual task. Here, parameters must be found that the normative glance behavior defined by solving the partially observable Markov decision process is accepted by drivers. Furthermore, a prediction model of the future imperfect driver steering and gaze switching behavior is required. These must be quantified to decide on current appropriate glance behavior that suits with respect to realistic driver behavior. Cpt. 4 presents new techniques for estimating these parameters from real driving data using inverse optimal control. Here, we derive a new inference approach for the class of partially observable Markov decision processes the joint task model belongs to. Thereby, the properties can be exploited to efficiently solve the problem class. Furthermore, we introduce a driving experiment on lane keeping in real traffic. The obtained behavioral data is used to evaluate proposed methodology with respect to quality of prediction when using the estimated parameters.

The partially observable Markov decision process model involved in the definition of appropriate glance behavior explicitly models the driver's perception. This done by means of a sensor model of the characteristics of the driver's vision. Suitable values of the sensor model are of crucial importance for the realism of the full model. However, estimating these models is a challenging problem. This is because the visual sensory measurements made by the driver cannot technically be measured in driving experiments. In Cpt. 5, we address this important issue by presenting the first mathematical framework for inference of sensor models underlying real-world motor behavior. The approach is first derived for general partially observable Markov decision processes in conceptual form. Thereafter, algorithms for exact inference for the problem class of the normative model of glance behavior are presented. Finally, a new dataset of driver behavior is introduced. Here, the drivers engaged into three different variants of a realistic secondary task along with varying characteristics of the sensing of the forward road scenery. The data set is used to evaluate the prediction performance under the inferred sensor models.

Cpt. 6 considers the development and an evaluation of a distraction warning system based on the previously obtained normative model of glance behavior. In this context, a robust estimate of the time past since the driver averted his or her gaze from the road is required. For this purpose an estimation approach based on a particle filter is developed. Thereafter, we describe the architecture of the warning system and how it can be implemented in a test vehicle. Finally, the data and results of a user test conducted on a test track are presented. This experiment served the purpose of comparing a warning systems based on a state-of-the-art approach to distraction assessment to the system developed in this work.

Finally, Cpt. 7 reviews the contributions and findings of this thesis. Furthermore, open problems and possible directions for future research are discussed.

We give an schematic overview of this work in Fig. 1.1.

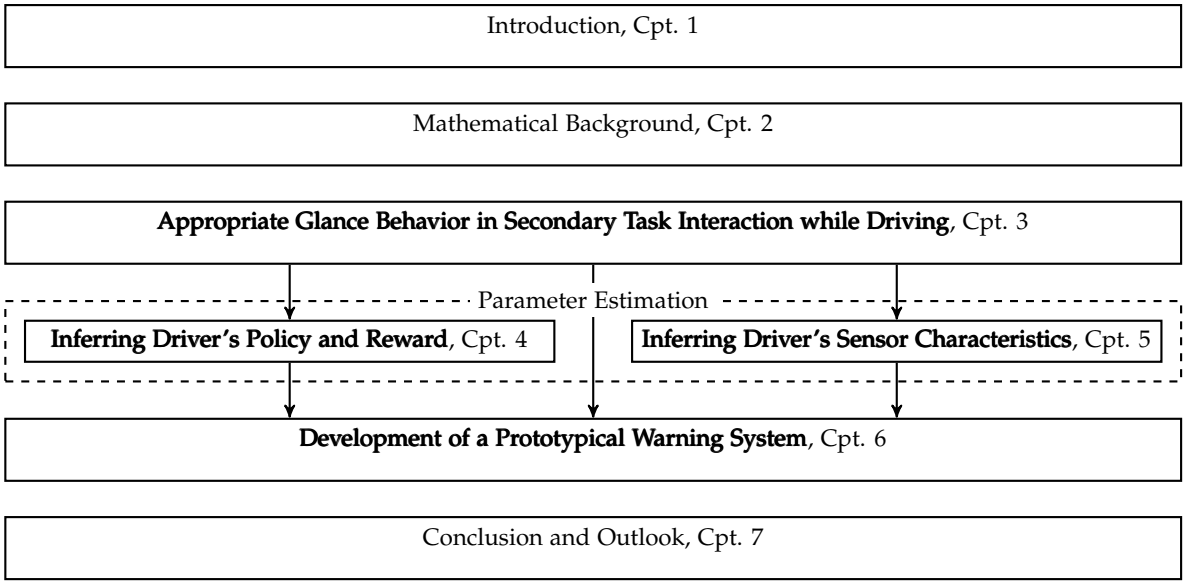

Figure 1.1: Schematic overview and outline of the thesis

As can be seen in the figure, first the model of appropriated glance behavior is obtained in the main part of the thesis. In the following two chapters, we derive new techniques to estimate the model parameters. Finally, the fully specified model is used in the prototypical warning system.

1.4 Contributions

After outlining the thesis, we wish to summarize the main contributions of this work. This thesis contributes to the state-of-the-art in the research on intelligent vehicles and driver assistance systems by developing and validating novel algorithmic methodology for situation specific automatic assessment of driver attention. To the best of the author's knowledge, a similar comprehensive mathematical framework for assessment of driver attention has not been available before. Considering an interdisciplinary subject, the following specific contributions in the fields of machine learning, optimal control, analysis of human machine systems and intelligent vehicles were made:

- **A Computationally Feasible Model of Appropriate Glance Behavior** (Cpt. 3):
  The first comprehensive and mathematical normative model of glance behavior for secondary task interaction in lane keeping is developed. For this purpose, we extend a previous partially observable Markov decision process model as well as previous solution techniques for efficient computing rational glance policies therein. Finally, a thorough evaluation of the glance behavior with respect to realism and computational demand with regards to the application context is contributed.

  This allows to compute situation specific appropriate glance behavior. Furthermore, the obtained algorithms and empirical findings form a basis for developing normative models for other driving tasks.

  Parts of this work were published in:

  F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Predicting lane keeping behavior of visually distracted drivers using inverse suboptimal control. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2016.

  F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Exact maximum entropy inverse optimal control for modeling human attention scheduling and control. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

- **Inverse Optimal Control for a New Class of Partially Observable Markov Decision Processes** (Cpt. 4):
  We derive maximum causal entropy inverse optimal control for the problem class the model of normative glance behavior belongs to. Thereby, we generalize previous work with respect to the model class. In addition to that, we provide the first comparison of the maximum causal entropy approach and the maximum causal likelihood variant of the original framework. Finally, the methodology is evaluated on behavioral data obtained in traffic driving, which is more realistic than simulations used in many other works.

  This contributes to the generalization and better understanding of maximum causal entropy inverse optimal control. In addition to that, we provide practitioners with an efficient and effective algorithmic tool to parametrize the normative model of glance behavior model.

  Parts of this work were published in:

  F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Predicting lane keeping behavior of visually distracted drivers using inverse suboptimal control. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2016.

  F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Exact maximum entropy inverse optimal control for modeling human attention scheduling and control. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

- **A Novel Approach for Inferring Sensor Models and Rewards Underlying Human Motor Behavior** (Cpt. 5):
  This work developed the first general mathematical framework for inference of sensor models and rewards underlying real world motor behavior. In contrast to previous work, here only the assumption of rational behavior in an arbitrary given partially observable Markov decision process is required. A novel algorithmic approach for estimating sensor model in the problem class of the normative glance model is obtained. In addition to that, the developed technique is evaluated using a new data set of four hours of real traffic driving.

  The obtained mathematical concept enables development of new machine learning techniques for inference of sensor model in further specific problem classes. Hence, it can provide a helpful computational methodology for understanding human sensori-motor behavior in cognitive science. The algorithms derived in this work can be used to find an appropriate parametrization of the normative model of glance behavior.

  Parts of this work were published in:

  F. Schmitt, H.-J. Bieg, M. Herman, and C. Rothkopf. I see what you see: Inferring sensor and policy models of human real-world motor behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

- **Development and Evaluation of A Novel Distraction Warning System** (Cpt. 6):
  We develop a novel distraction warning system that takes into account the context of the current driving situation. Furthermore, the first evaluation of adapting distraction warnings to the driving situation in comparison to static warning thresholds is conducted by means of a real driving user test.

  The evaluation demonstrates that the developed normative glance model can indeed be applied to obtain a better a real-time distraction warning system

  This work will be submitted as:

  F. Schmitt, H.-J. Bieg, D. Manstetten, and R. Stiefelhagen. Distraction mitigation by computation of appropriate glance behavior and its evaluation in a user test. *(manuscript in preparation) IEEE Transaction on Intelligent Transport Systems*, 2017.

2 Mathematical Background

A variety of mathematical concepts and notations are employed throughout this thesis. We assume familiarity with the basic concepts of probability theory and linear algebra. This chapter will address models and frameworks of optimal and rational decision making that the algorithms derived in Cpt. 3, Cpt. 4, Cpt. 5 will strongly build on.

Many of the aspects considered in this work will be of temporal nature, such as driver glance patterns. Although certain relevant physical processes such as the kinematics of a vehicle are typically defined in continuous time, all quantities x_t will be assumed to be in *discrete* time by default. Therefore, x_{t-1} will typically refer to the previous value of the quantity and x_{t+1} to the next value.

2.1 Models of Optimal Sequential Decision Making

Automobile driving is a dynamic process. Here, the driver continuously monitors the state of the vehicle and the scenery and applies a certain control input to change or keep the current state. In addition, drivers usually pursue a certain objective in driving e.g. safely reaching their destination in the shortest time. Such tasks where an agent, e.g. the driver, makes sequential decision in pursuit of an objective can be modeled as a *Markov decision process*.

2.1.1 Markov Decision Processes

Definition A *finite horizon* Markov Decision Process (MDP) [22] with *decision horizon* T consists of *states* x_t in a *state space* S and *controls* u_t in a *control space* U available to the decision making agent. The agent's choice of controls is modeled by a potentially stochastic and time-varying mapping from states to controls, a so-called *policy*, $\boldsymbol{\pi}_{0:T}\colon u_t \sim \pi_t(.|x_t)$. Sampled from an *initial distribution* $x_0 \sim p_0$, the states x_t evolve according to a stochastic *process model* $\boldsymbol{\mathcal{P}}_{0:T}$ by means of the dynamics

$$x_{t+1} \sim \mathcal{P}_t(.|x_t, u_t) \tag{2.1}$$

and the applied policy $\pi_t(u_t|x_t)$. The states reached and the controls applied by the agent are evaluated by means of a *reward function* r:

$$r(x_t, u_t)\colon \mathsf{S} \times \mathsf{U} \to \mathbb{R}. \tag{2.2}$$

The architecture of a MDP is outlined in Fig. 2.1.

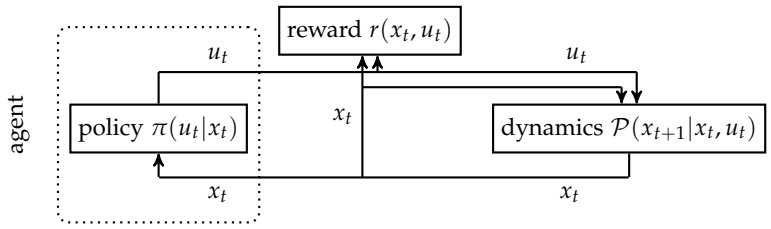

Figure 2.1: Illustration of the model parts of an MDP. Here, the agent chooses a control u_t based on the current state x_t based on a policy π_t. The agent receives a reward $r(x_t, u_t)$ and the next state x_{t+1} is sampled according to the dynamics $\mathcal{P}(x_{t+1}|x_t, u_t)$.

A sequence of states and controls $\mathcal{T} = (\mathbf{x}_{0:T}, \mathbf{u}_{0:T})$ is called a trajectory. Given a reward function $r(x_t, u_t)$, the *return* $R(\mathcal{T})$ of a trajectory $\mathcal{T}$ is defined as

$$R(\mathcal{T}) = \sum_{t=0}^{T} r(x_t, u_t). \tag{2.3}$$

That is, R is the evaluation of decisions $\mathbf{u}_{0:T}$ made by the agent.

Optimal Policies The objective of an MDP is to find a policy $\boldsymbol{\pi}^\star_{0:T}$ that maximizes the expected return R over a horizon T under the dynamics and the initial distribution. Specifically, we seek the optimal solution of the problem

$$\boldsymbol{\pi}^\star_{0:T} = \arg\max_{\boldsymbol{\pi}_{0:T}} \mathbb{E}\Big[\sum_{t=1}^{T} r(x_t, u_t) \Big| \boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big]. \tag{2.4}$$

The most fundamental solution approach to this optimization problem, already introduced in the original work [22], are the so-called *Bellman-Equations*

$$Q^\star_t(x_t, u_t) = \begin{cases} r(x_t, u_t) + \mathbb{E}[\tilde{V}_{t+1}(x_{t+1}) | \mathcal{P}_t(x_{t+1}|x_t, u_t)] & \text{if } t < T \\ r(x_T, u_T) & \text{else} \end{cases} \tag{2.5}$$

$$V^\star_t(x_t) = \max_{u_t}(Q^\star_t(x_t, u_t)) \tag{2.6}$$

$$\pi^\star_t(u_t|x_t) = \mathcal{I}(u_t|u^\star_t(x_t)), \quad u^\star_t(x_t) = \arg\max_{u_t}\big(Q^\star_t(x_t, u_t)\big). \tag{2.7}$$

In this context, the function $Q^\star_t(x_t, u_t)$ is referred to as *optimal state-control-function* while $V^\star_t(x_t)$ is referred to as *optimal value-function*. Interestingly, the optimal policy $\boldsymbol{\pi}^\star$ of an MDP is deterministic and independent of the initial state distribution p_0. In Cpt.s 3-5, however, we will often consider the initial distribution for computing the optimal policy. The reason is that due to the special MDP structure in these cases not all states $x_t \in \mathsf{S}$ can be reached at every time step t. This will be exploited for more efficient optimal policy computation.

While the recursion (2.7)-(2.6) can in principle be used to obtain the optimal policy $\boldsymbol{\pi}$ of any MDP, they are often computationally infeasible. This is because computing $u^\star_t(x_t)$ requires to backup the optimal state-control function values $Q^\star_t(x_t, u_t)$ **for all pairs** x_t, u_t. Although this approach is tractable in small discrete spaces S, U this is not the case in large or even continuous spaces S, U. Here, one needs to resort to approximation methods. An overview over MDP, their properties and solution approaches is provided by [181].

Nomenclature Used in Literature As problems of sequential decision making occur in several domains, several research communities have developed computational solution approaches:

- In control theory and numerical optimization usually MDPs with known models of the reward and dynamics are considered. In this context (approximative) optimal solution is referred to as *Optimal Control* (OC) [25].
- MDPs with unknown dynamics are of interest in machine learning and artificial intelligence. Given an artificial agent that chooses controls and observes the outcomes in form of rewards and successor states, here the goal is to iteratively learn the optimal policy based on the obtained observations. As policy optimization is based solely on the interaction data, this is termed *Reinforcement Learning* (RL) [231]. Notably, there are some RL approaches that build models and solve optimal control problems in iteration [47, 134].

In this thesis, situationally appropriate gaze policies are obtained by first developing and estimating MDP models followed by computing policies therein. Therefore, we will use the term optimal control for computing optimal policies and related matters throughout this work.

2.1.2 Linear Quadratic Regulation

One of the few important MDPs, that can efficiently be solved by means of the Bellman-equations is the class of *Linear Quadratic Regulation* (LQR) problems.

Definition Here, the state and control spaces are given as $\mathsf{S} = \mathbb{R}^{n_x}$, $\mathsf{U} = \mathbb{R}^{n_u}$. The dynamics $\boldsymbol{\mathcal{P}}$ are linear-affine with matrices $\mathbf{A}_t \in \mathbb{R}^{n_x,n_x}$, $\mathbf{B}_t \in \mathbb{R}^{n_x,n_u}$, $\mathbf{a}_t \in \mathbb{R}^{n_x}$ according to

$$\mathbf{x}_{t+1} = \mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t + \boldsymbol{\epsilon}_t^x, \quad \boldsymbol{\epsilon}_t^x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\epsilon^x}), \tag{2.8}$$

subject to random Gaussian noise $\boldsymbol{\epsilon}_t^x \in \mathbb{R}^{n_x}$ and the initial distribution is a Gaussian $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_0^x, \boldsymbol{\Sigma}_0^x)$. Furthermore, the reward is a negative quadratic form

$$r(\mathbf{x}_t, \mathbf{u}_t) = -\mathbf{x}_t^\top \mathbf{C}_x \mathbf{x}_t - \mathbf{u}_t^\top \mathbf{C}_u \mathbf{u}_t, \tag{2.9}$$

with symmetric positive semi-definite matrix $\mathbf{C}_x$, $\min(\mathrm{eig}(\mathbf{C}_x)) \geq 0$ and symmetric positive definite matrix $\mathbf{C}_u$, $\min(\mathrm{eig}(\mathbf{C}_u)) > 0$.

Optimal Policies The most important property of LQR, known at least since E. Kalman [100], is that the optimal state-control and value functions $Q_t^\star(\mathbf{x}_t, \mathbf{u}_t)$, $V_t^\star(\mathbf{x}_t)$ are given by the quadratic forms

$$Q_t^\star(\mathbf{x}_t, \mathbf{u}_t) = [\mathbf{x}_t; \mathbf{u}_t]^\top \mathbf{M}_t^{Q^\star} [\mathbf{x}_t; \mathbf{u}_t] + \mathbf{m}_t^{Q^\star} [\mathbf{x}_t; \mathbf{u}_t] + m_t^{Q^\star,1} + m_t^{Q^\star,1} \tag{2.10}$$

$$V_t^\star(\mathbf{x}_t) = \mathbf{x}_t^\top \mathbf{M}_t^{V^\star} \mathbf{x}_t + \mathbf{m}_t^{V^\star} \mathbf{x}_t + m_t^{V^\star,1} + m_t^{V^\star,2}. \tag{2.11}$$

Here, $\mathbf{M}_t^{Q^\star}$ is a symmetric negative definite matrix $\max(\mathrm{eig}(\mathbf{M}_t^{Q^\star})) < 0$ and $\mathbf{M}_t^{V^\star}$ is a symmetric negative definite matrix $\max(\mathrm{eig}(\mathbf{M}_t^{V^\star})) < 0$. This can be proven by recursively evaluating the Bellman equations. Specifically, it holds for the variables $\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, m_t^{Q^\star,2}$

$$\mathbf{M}_t^{Q^\star} = \begin{cases} [\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star} [\mathbf{A}_t\, \mathbf{B}_t] - \mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{if } t < T \\ -\mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{else} \end{cases} \tag{2.12}$$

$$\mathbf{m}_t^{Q^\star} = \begin{cases} 2[\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star} \mathbf{a}_t + [\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{V^\star} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \tag{2.13}$$

$$m_t^{Q^\star,1} = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{V^\star} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{V^\star} + m_{t+1}^{V^\star,1} & \text{if } t < T \\ 0 & \text{else} \end{cases} \tag{2.14}$$

$$m_t^{Q^\star,2} = \begin{cases} \mathrm{tr}(\mathbf{M}_{t+1}^{V^\star} \boldsymbol{\Sigma}^{\epsilon^x}) + m_{t+1}^{V^\star,2} & \text{if } t < T \\ 0 & \text{else} \end{cases}. \tag{2.15}$$

Next, we define the following variables as

$$\mathbf{M}_t^{Q^\star} = \begin{bmatrix} \mathbf{M}_{t\ 1:n_x,1:n_x}^{Q^\star} & \mathbf{M}_{t\ 1:n_x,n_x+1:n_x+n_u}^{Q^\star} \\ \mathbf{M}_{t\ n_x+1:n_u+n_x,1:n_x}^{Q^\star} & \mathbf{M}_{t\ u,n_x:n_x+n_u}^{Q^\star} \end{bmatrix} =: \begin{bmatrix} \mathbf{M}_{t\ x,x}^{Q^\star} & \mathbf{M}_{t\ x,u}^{Q^\star} \\ \mathbf{M}_{t\ u,x}^{Q^\star} & \mathbf{M}_{t\ u,u}^{Q^\star} \end{bmatrix} \tag{2.16}$$

$$\mathbf{m}_t^{Q^\star} = \begin{bmatrix} \mathbf{m}_{t\ 1:n_x}^{Q^\star} \\ \mathbf{m}_{t\ n_x+1:n_u+n_x}^{Q^\star} \end{bmatrix} := \begin{bmatrix} \mathbf{m}_{t\ x}^{Q^\star} \\ \mathbf{M}_{t\ u}^{Q^\star} \end{bmatrix}. \tag{2.17}$$

According to the Bellman equations (2.6) it holds $V_t^\star(\mathbf{x}_t) = \max_{\mathbf{u}_t}(Q_t^\star(\mathbf{x}_t, \mathbf{u}_t))$. That means for obtaining $V_t^\star(\mathbf{x}_t)$ we need to find the maximum value of $Q_t^\star(\mathbf{x}_t, \mathbf{u}_t) = [\mathbf{x}_t; \mathbf{u}_t]^\top \mathbf{M}_t^{Q^\star}[\mathbf{x}_t; \mathbf{u}_t] + \mathbf{m}_t^{Q^\star}[\mathbf{x}_t; \mathbf{u}_t] + m_t^{Q^\star,1} + m_t^{Q^\star,1}$ for given a given $\mathbf{x}_t$. As $Q_t^\star(\mathbf{x}_t, \mathbf{u}_t)$ is a negative quadratic function its maximum value can be obtained by finding $\nabla_{\mathbf{u}} Q_t^\star(\mathbf{x}_t, \mathbf{u}_t^\star) = 0$. Using the Schur complements of $\mathbf{M}_t^{Q^\star}$ (see e.g. [175]) we can obtain $V_t^\star(\mathbf{x}_t)$ by means of the terms $\mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, m_t^{V^\star,2}$ given as

$$\mathbf{M}_t^{V^\star} = \mathbf{M}_{t\,x,x}^{Q^\star} - \mathbf{M}_{t\,x,u}^{Q^\star}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{M}_{t\,u,x}^{Q^\star} \tag{2.18}$$

$$\mathbf{m}_t^{V^\star} = \mathbf{m}_{t\,x}^{Q^\star} - \mathbf{M}_{t\,x,u}^{Q^\star}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{m}_{t\,u}^{Q^\star} \tag{2.19}$$

$$m_t^{V^\star,1} = m_t^{Q^\star,1} - \frac{1}{4}[\mathbf{m}_{t\,u}^{Q^\star}]^\top[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{m}_{t\,u}^{Q^\star} \tag{2.20}$$

$$m_t^{V^\star,2} = m_t^{Q^\star,2}. \tag{2.21}$$

Note, that we explicitly specified the constants $m_t^{Q^\star,1}, m_t^{Q^\star,2}, m_t^{V^\star,1}, m_t^{V^\star,2}$ which are often omitted in the literature. In LQR these do not affect the policy, however this is not the case in its extensions which we will consider later (see Cpt. 3).

Finally, the optimal policy $\boldsymbol{\pi}^\star$ is given as a linear-affine feedback controller

$$\pi^\star(\mathbf{u}_t|\mathbf{x}_t) = \mathcal{I}(\mathbf{u}_t|\mathbf{u}_t^\star(\mathbf{x}_t)) \tag{2.22}$$

$$\mathbf{u}_t^\star(\mathbf{x}_t) = \arg\max_{\mathbf{u}_t}\left([\mathbf{x}_t; \mathbf{u}_t]^\top \mathbf{M}_t^{Q^\star}[\mathbf{x}_t; \mathbf{u}_t] + \mathbf{m}_t^{Q^\star}[\mathbf{x}_t; \mathbf{u}_t] + m_t^{Q^\star,1} + m_t^{Q^\star,2}\right) \tag{2.23}$$

$$= \mathbf{F}_t^\star \mathbf{x}_t + \mathbf{f}_t^\star, \quad \mathbf{F}_t^\star := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{M}_{t\,u,x}^{Q^\star}, \quad \mathbf{f}_t^\star := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{m}_{t\,u}^{Q^\star}. \tag{2.24}$$

Altogether the optimal policy $\boldsymbol{\pi}^\star$ was obtained using only matrix multiplications and inversion of $[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}$. Therefore, LQRs can quickly be numerically solved using basic linear algebra routines. As the algorithms derived later in this work rely on the solution of a deterministic LQR problem as a subroutine, a pseudo-code algorithm for its solution is given by Algo. 1.

Algorithm 1 Optimal Solution of LQR LQRopt

Require: all

function LQROPT($\mathbf{C}_x, \mathbf{C}_u, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}$)
  $\mathbf{M}_{T+1}^{V^\star} \leftarrow \mathbf{0}^{n_x,n_x}$
  $\mathbf{m}_{T+1}^{V^\star} \leftarrow \mathbf{0}^{1,n_x}$
  $m_{T+1}^{V^\star,1} \leftarrow 0$
  for $t = T, \ldots, 0$ **do**
    $\mathbf{M}_t^{Q^\star} \leftarrow [\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star}[\mathbf{A}_t\, \mathbf{B}_t] - \mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u)$
    $\mathbf{m}_t^{Q^\star} \leftarrow 2[\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star}\mathbf{a}_t + [\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{V^\star}$
    $\mathbf{U} = [\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}$
    $m_t^{Q^\star,1} \leftarrow \mathbf{a}_t^\top \mathbf{M}_{t+1}^{V^\star}\mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{V^\star} + m_{t+1}^{V^\star,1}$
    $\mathbf{M}_t^{V^\star} \leftarrow \mathbf{M}_{t\,x,x}^{Q^\star} - \mathbf{M}_{t\,x,u}^{Q^\star}\mathbf{U}\,\mathbf{M}_{t\,u,x}^{Q^\star}$
    $\mathbf{m}_t^{V^\star} \leftarrow \mathbf{m}_{t\,x}^{Q^\star} - \mathbf{M}_{t\,x,u}^{Q^\star}\mathbf{U}\,\mathbf{m}_{t\,u}^{Q^\star}$
    $m_t^{V^\star,1} \leftarrow -\frac{1}{4}[\mathbf{m}_{t\,u}^{Q^\star}]^\top \mathbf{U}\,\mathbf{m}_{t\,u}^{Q^\star} + m_t^{Q^\star,1}$
    $\mathbf{F}_t^\star \leftarrow -\frac{1}{2}\mathbf{U}\,\mathbf{M}_{t\,u,x}^{Q^\star}$
    $\mathbf{f}_t^\star \leftarrow -\frac{1}{2}\mathbf{U}\,\mathbf{m}_{t\,u}^{Q^\star}$
  end for
  return $\left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_t^\star\right)_{t=0:T}$
end function

Application Although, most relevant dynamical processes do not follow linear-affine dynamics (2.8), LQR is widely used to compute policies in practice [11]. In addition to computational convenience

this is because one is usually interested in keeping a tracking error x_t^e close to zero with least control effort possible as measured by the reward $r(x_t^e, u_t) = -c_x(x_t^e)^2 - c_x(u_t)^2$. In this case, a reasonable approximation of the nonlinear dynamics of the tracking error $x_{t+1}^e = f(x_t^e, u_t)$ is given by its first-order Taylor-approximation at $x_t^e, u_t = 0$,

$$f(x_t^e, u_t) \approx \nabla_{x_t^e, u_t} f(x_t^e, u_t)|_{x_t^e, u_t = 0}[x_t^e; u_t] + f(0, 0).$$

As the optimal policy $\boldsymbol{\pi}^\star$ for the LQR approximation tries to keep both x_t^e, u_t close to zero, the qualtiy of the linear approximation w.r.t. dynamic remains good which results in good performance of LQR policy. A thorough explanation of this robustness property can be found in [11]. Finally, iterative application of LQR has also proven to be an efficient approach to approximately solve nonlinear MPDs in robotics [240, 242].

2.1.3 Partially Observable Markov Decision Processes

In previous section on MDPs, the agent had full knowledge of the states x_t when deciding on controls u_t. However, this is not the case in many real-world decision making problems. For example, in manual automobile driving it is unrealistic to assume that the driver can fully sense every state of the driving situation at any time. Fortunately, MDPs can be extended to explicitly take into account aspects of sensing and perception. This is possible using the class of *Partially Observable Markov Decision Processes* (POMDPs) [216, 215].

Definition In a POMDP the decision making agent relies on noisy and/or incomplete sensory measurements z_t of the "true" states x_t. These measurements lay in a measurement space Z and their conditional distribution with respect to the state x_t is given by a sensor model

$$z_t \sim p^z(.|x_t). \tag{2.25}$$

The objective in a POMDP is to find a policy $\boldsymbol{\pi}_{0:T}$: $u_t \sim \pi_t(.|\boldsymbol{z}_{0:t})$ based on the history of measurements $\boldsymbol{z}_{0:t}$ that maximizes the expected return of the applied actions and the *unknown* visited states. This is formalized in the optimization problem

$$\boldsymbol{\pi}_{0:T}^\star = \arg\max_{\boldsymbol{\pi}_{0:T}} \mathbb{E}\Big[\sum_{t=1}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}_{0:T}, p^z, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big]. \tag{2.26}$$

As the states are not directly accessible, the optimal policy in POMDPs generally depends on the entire history $\boldsymbol{z}_{0:t}$ of measurements instead of only the most recent one z_t. This is because considering the past measurements, effectively improves estimating the current state and therefore leads to better controls. Fig. 2.2 gives a schematic illustration of the previously introduced model aspects.

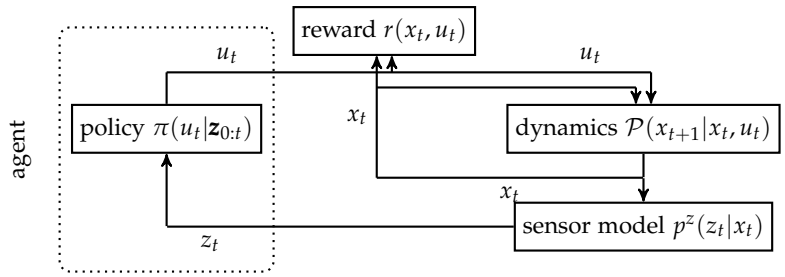

Figure 2.2: Illustration of the model parts of an POMDP. Here, the agent chooses a control u_t based on the history of sensory measurements $\boldsymbol{z}_{0:t}$ based on a policy π_t. The agent receives a reward $r(x_t, u_t)$ and the next state x_{t+1} is sampled according to the dynamics $\mathcal{P}(x_{t+1}|x_t, u_t)$. Based on the new state x_{t+1} a new sensory measurement is generated according to the sensor model $p^z(z_{t+1}|x_{t+1})$.

Belief-MDP A common approach to handle POMDPs is to transform them into equivalent MDPs. This is possible using the *belief* $b_t(x_t)$, i.e. the a-posterior distribution of the state x_t given the history of measurements and applied controls $\boldsymbol{z}_{0:t}, \boldsymbol{u}_{0:t-1}$ under knowledge of the dynamics $\boldsymbol{\mathcal{P}}_{0:t-1}$ and initial state distribution p_0,

$$b(x_t) := p(x_t|\boldsymbol{z}_{0:t}, \boldsymbol{u}_{0:t-1}, \boldsymbol{\mathcal{P}}_{0:t-1}, p_0), \quad p_0^b := b(x_0) = p_0. \tag{2.27}$$

According to *Bayes's* law the belief can recursively be obtained by means of the *belief update*

$$b(x_t) = \frac{\int p(z_t|x_t)\mathcal{P}_t(x_t|u_{t-1}, x_{t-1})b(x_{t-1})\,\mathrm{d}\,x_{t-1}}{\int p(z_t'|x_t)\mathcal{P}_t(x_t|u_{t-1}, x_{t-1})b(x_{t-1})\,\mathrm{d}\,z_t', x_{t-1}} \tag{2.28}$$

given the current sensory measurement z_t, the past belief $b(x_{t-1})$ and the past applied control x_{t-1}. Let $\mathsf{Z}(b(x_t)|b(x_{t-1}), u_{t-1})$ denote the set of measurements z_t that produce the belief $b(x_t)$ by the belief update (2.28) given the previous belief $b(x_{t-1})$ and the previous control u_{t-1},

$$\mathsf{Z}(b(x_t)|b(x_{t-1}), u_{t-1}) = \left\{\hat{z}_t \in \mathsf{Z}\colon b(x_t) = \frac{\int p(\hat{z}_t|x_t)\mathcal{P}_t(x_t|u_{t-1}, x_{t-1})b(x_{t-1})\,\mathrm{d}\,x_{t-1}}{\int p(z_t'|x_t)\mathcal{P}_t(x_t|u_{t-1}, x_{t-1})b(x_{t-1})\,\mathrm{d}\,z_t', x_{t-1}}\right\}. \tag{2.29}$$

We can now define the belief dynamics $\boldsymbol{\mathcal{P}}^b$ by means of

$$\mathcal{P}_t^b(b(x_{t+1})|b(x_t), u_t) := \int_{\mathsf{Z}(b(x_{t+1})|b(x_t),u_t)} \Big(\int p(z_{t+1}|x_{t+1})\mathcal{P}_{t+1}(x_{t+1}|u_t, x_t)b(x_t)\,\mathrm{d}\,x_{t+1}, x_t \Big)\,\mathrm{d}\,z_{t+1}. \tag{2.30}$$

That is the probability of a transition from belief $b(x_t)$ to belief $b(x_{t+1})$ is the integral of the probability of all observations that produce the belief $b(x_{t+1})$ by means of the belief update. Furthermore, taking the expected reward under the a-posterior of the state x_t

$$r^b(b(x_t), u_t) := \mathbb{E}[r(x_t, u_t)|b(x_t), u_t] = \mathbb{E}[r(x_t, u_t)|p(x_t|z_{0:t}, u_{0:t-1}, \boldsymbol{\mathcal{P}}_{0:t-1}, p_0), u_t]. \tag{2.31}$$

yields a reward function on belief and control $b(x_t), u_t$. Finally, the belief-MDP equivalent to the original POMDP is given by reward function r^b (2.31), dynamics $\boldsymbol{\mathcal{P}}^b$ (2.30) and initial distribution p_0^b (2.26) [15]. The relationship between the individual parts of the belief MDP are illustrated in Fig. 2.3.

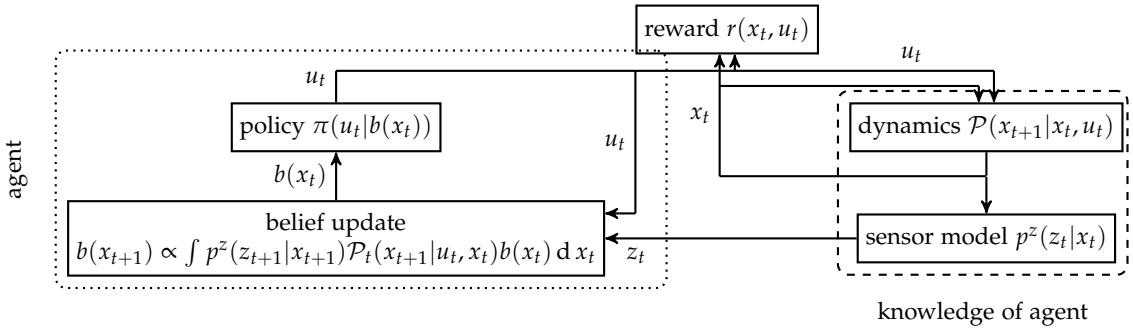

Figure 2.3: Illustration of the model parts of a belief MDP. Here, the agent chooses a control u_t based on the current belief $b(x_t)$ of the state x_t based on a policy π_t. The agent receives a reward $r(x_t, u_t)$ and the next state x_{t+1} is sampled according to the dynamics $\mathcal{P}(x_{t+1}|x_t, u_t)$. Based on the new state x_{t+1} a new sensory measurement is generated according to the sensor model $p^z(z_{t+1}|x_{t+1})$. The sensory measurement z_{t+1} is fused with the belief $b(x_t)$ of the states as well as the knowledge of the agent regarding the dynamics and the sensor model. This leads to the new belief $b(x_{t+1})$.

Note, that the belief MPD implicitly implies that the agent has full knowledge of both the dynamics and the sensor model. Those are required to conduct the belief update (2.28). We will later refer to this aspect and discuss whether this is realistic in the context of drivers' belief of vehicle/driving situation states (Cpt. 5). Furthermore, belief MDPs can very often not exactly be solved by means of the

Bellman equation. This is because the belief is always a continuous variable (a probability distribution) resulting in a MDP in continuous states. Formally, POMDPs can also proven to be of harder complexity class than ordinary MDPs [171]. A review of POMDP models and solution techniques with respect to application for human behavior modeling will be conducted in Cpt. 3.

2.1.4 Linear Quadratic Gaussian Problems

Similar as LQR for MDPs (see subsection 2.1.2), *Linear Quadratic Gaussian Problems* (LQG)s [100] form a notable special class of POMDPs.

Definition In LQGs reward and dynamics are the same as in LQRs

$$\mathbf{x}_{t+1} = \mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t + \boldsymbol{\epsilon}_t^x, \quad \boldsymbol{\epsilon}_t^x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\epsilon^x}), \quad \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_0^x, \boldsymbol{\Sigma}_0^x) \tag{2.32}$$

$$r(\mathbf{x}_t, \mathbf{u}_t) = -\mathbf{x}_t^\top \mathbf{C}_x \mathbf{x}_t - \mathbf{u}_t^\top \mathbf{C}_u \mathbf{u}_t, \tag{2.33}$$

while LQGs additionally feature measurements $\mathbf{z}_t \in \mathbb{R}^{n_z}$ according to a linear Gaussian sensor model

$$\mathbf{z}_t = \mathbf{H}\mathbf{x}_t + \boldsymbol{\epsilon}_t^z, \quad \boldsymbol{\epsilon}^z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\epsilon^z}) \tag{2.34}$$

with measurement matrix $\mathbf{H} \in \mathbb{R}^{n_z, n_x}$ and Gaussian measurement noise $\boldsymbol{\epsilon}_t^z$.

Following Sec. 2.1.3 now the equivalent belief MDP of the original LQG POMDP is derived.

Belief-MDP Due to the linear-Gaussian model parts (2.32),(2.34) here the belief $b(\mathbf{x}_t)$ is given by a Gaussian,

$$b(\mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x). \tag{2.35}$$

Specifically, the a-posterior mean $\boldsymbol{\mu}_t^x$ and a-posterior covariance $\boldsymbol{\Sigma}_t^x$ are obtained by means of the well-known Kalman filter,

$$\bar{\boldsymbol{\Sigma}}_{t+1}^x = \mathbf{A}_t\boldsymbol{\Sigma}_t^x\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} \tag{2.36}$$

$$\bar{\boldsymbol{\mu}}_{t+1}^x = \mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t \tag{2.37}$$

$$\mathbf{K}_{t+1} = \bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top(\mathbf{H}\bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top + \boldsymbol{\Sigma}^{\epsilon^z})^+ \tag{2.38}$$

$$\boldsymbol{\Sigma}_{t+1}^x = (\mathbf{I}^{n_x} - \mathbf{K}_{t+1}\mathbf{H})\bar{\boldsymbol{\Sigma}}_{t+1}^x \tag{2.39}$$

$$\boldsymbol{\mu}_{t+1}^x = \bar{\boldsymbol{\mu}}_{t+1}^x - \mathbf{K}_{t+1}(\mathbf{H}\bar{\boldsymbol{\mu}}_{t+1}^x - \mathbf{z}_{t+1}). \tag{2.40}$$

Note, that here the more common inverse of $(\mathbf{H}\bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top + \boldsymbol{\Sigma}^{\epsilon^z})$ is replaced with a pseudo-inverse. The reason is that in Cpt. 3-5 the matrix $\boldsymbol{\Sigma}_t^x$ is not necessarily invertible. In this case, the optimal Kalman gain $\mathbf{K}_{t+1}$ is obtained by means of the pseudo inverse. The Kalman update is summarized in Algo. 2.

Algorithm 2 Kalman Update

Require: $\boldsymbol{\Sigma}_t^x, \mathbf{A}, \mathbf{a}, \mathbf{B}, \boldsymbol{\Sigma}^{\epsilon^x}$

function KalmanUpdate($\boldsymbol{\Sigma}_t^x, \mathbf{A}, \mathbf{a}, \mathbf{B}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}, \boldsymbol{\Sigma}^{\epsilon^z}, \boldsymbol{\mu}_t^x, \mathbf{z}$)
  $\bar{\boldsymbol{\Sigma}}_{t+1}^x \leftarrow \mathbf{A}\boldsymbol{\Sigma}_t^x\mathbf{A}^\top + \boldsymbol{\Sigma}^{\epsilon^x}$
  $\mathbf{K}_{t+1} \leftarrow \bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top(\mathbf{H}\bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top + \boldsymbol{\Sigma}^{\epsilon^z})^+$
  $\boldsymbol{\Sigma}_{t+1}^x \leftarrow (\mathbf{I}^{n_x} - \mathbf{K}\mathbf{H})\bar{\boldsymbol{\Sigma}}_{t+1}^x$
  $\bar{\boldsymbol{\mu}}_{t+1}^x \leftarrow \mathbf{A}\boldsymbol{\mu}_t^x + \mathbf{B}\mathbf{u}_t + \mathbf{a}$
  $\boldsymbol{\mu}_{t+1}^x \leftarrow \bar{\boldsymbol{\mu}}_{t+1}^x - \mathbf{K}(\mathbf{H}\bar{\boldsymbol{\mu}}_{t+1}^x - \mathbf{z}_{t+1})$
  return $\boldsymbol{\Sigma}_{t+1}^x, \bar{\boldsymbol{\mu}}_{t+1}^x$
end function

Finally, the belief-MDP of LQG is formulated as follows: The reward function r^b is given by

$$r^b(b(\mathbf{x}_t), \mathbf{u}_t) = r^b(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x, \mathbf{u}_t) = \mathbb{E}\big[-\mathbf{x}_t^\top \mathbf{C}_x \mathbf{x}_t - \mathbf{u}_t^\top \mathbf{C}_u \mathbf{u}_t \big| \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x) \big] \tag{2.41}$$

$$= -[\boldsymbol{\mu}^x]_t^\top \mathbf{C}_x \boldsymbol{\mu}_t^x - \mathrm{tr}(\mathbf{C}_x \boldsymbol{\Sigma}_t^x) - \mathbf{u}_t^\top \mathbf{C}_u \mathbf{u}_t, \tag{2.42}$$

and the belief dynamics $\boldsymbol{\mathcal{P}}^b$ result from the Kalman filter update (2.36) according to

$$\mathcal{P}_t^b(b(\mathbf{x}_{t+1}) | b(\mathbf{x}_t), \mathbf{u}_t) = \mathcal{P}_t^b(\boldsymbol{\mu}_{t+1}^x, \boldsymbol{\Sigma}_{t+1}^x | \boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x, \mathbf{u}_t) \tag{2.43}$$

$$= p_t(\boldsymbol{\mu}_{t+1}^x | \boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x, \mathbf{u}_t) \cdot \mathcal{I}(\boldsymbol{\Sigma}_{t+1}^x | \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)) \tag{2.44}$$

$$= \mathcal{N}(\boldsymbol{\mu}_{t+1}^x | \mathbf{A}_t \boldsymbol{\mu}_t^x + \mathbf{B}_t \mathbf{u}_t + \mathbf{a}_t, \mathbf{A}_t \boldsymbol{\Sigma}_t^x \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)) \quad \text{(see A.1)} \tag{2.45}$$

$$\cdot \mathcal{I}(\boldsymbol{\Sigma}_{t+1}^x | \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)). \tag{2.46}$$

Notably, the dynamics of the a-posterior covariance $\boldsymbol{\Sigma}_t^x$ are independent of the controls $\mathbf{u}_t$. This aspect greatly facilitates computing optimal policies.

Optimal Policies As Kalman discovered already in 1960 [100], in LQGs the state-control function and the value function of the corresponding belief-MDP are given by

$$Q_t^\star(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x, \mathbf{u}_t) = [\mathbf{x}_t; \mathbf{u}_t]^\top \mathbf{M}_t^{Q^\star} [\mathbf{x}_t; \mathbf{u}_t] + \mathbf{m}_t^{Q^\star} [\mathbf{x}_t; \mathbf{u}_t] + m_t^{Q^\star,1} + m_t^{Q^\star,2}(\boldsymbol{\Sigma}_t^x) \tag{2.47}$$

$$V_t^\star(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x) = \mathbf{x}_t^\top \mathbf{M}_t^{V^\star} \mathbf{x}_t + \mathbf{m}_t^{V^\star} \mathbf{x}_t + m_t^{V^\star,1} + m_t^{V^\star,2}(\boldsymbol{\Sigma}_t^x), \tag{2.48}$$

which can be proven by evaluating the Bellman equations of the belief MDP. In this context, $\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}$, $m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}$ follow the same recursive relations as their LQR counterparts (2.12),(2.18), whereas the scalar-value functions $m_t^{Q^\star,2}(.), m_t^{V^\star,2}(.)$ are obtained according to

$$m_t^{Q^\star,2}(\boldsymbol{\Sigma}_t^x) = \mathrm{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^x) + \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^\star} \big(\mathbf{A}_t \boldsymbol{\Sigma}_t^x \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x) \big) \right) + m_t^{V^\star,2}\big(\boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)\big) \tag{2.49}$$

$$m_t^{V^\star,2}(\boldsymbol{\Sigma}_t^x) = m_t^{Q^\star,2}(\boldsymbol{\Sigma}_t^x). \tag{2.50}$$

That is $m_t^{Q^\star,2}(\boldsymbol{\Sigma}_t^x)$ takes into account the cost of uncertainty by means of term $\mathrm{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^x)$ and the potential variation of the future a-posterior mean $\boldsymbol{\mu}_t^x$ by means of term $\mathrm{tr}\left(\mathbf{M}_{t+1}^{V^\star} \big(\mathbf{A}_t \boldsymbol{\Sigma}_t^x \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x) \big) \right)$. As a first direct consequence, the optimal LQG policy is equivalent to the LQR policy (2.22) applied to the a-posterior mean $\boldsymbol{\mu}_t^x$:

$$\pi^\star(\mathbf{u}_t | \boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}_t^x) = \mathcal{I}(\mathbf{u}_t | \mathbf{u}_t^\star(\boldsymbol{\mu}_t^x)), \quad \mathbf{u}_t^\star(\boldsymbol{\mu}_t^x) = \mathbf{F}_t^\star \boldsymbol{\mu}_t^x + \mathbf{f}_t^\star \tag{2.51}$$

$$\mathbf{F}_t^\star := -\frac{1}{2} [\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1} \mathbf{M}_{t\,u,x}^{Q^\star}, \quad \mathbf{f}_t^\star := -\frac{1}{2} [\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1} \mathbf{m}_{t\,u}^{Q^\star}, \tag{2.52}$$

Second, in contrast to LQRs the full value function $V_t^\star(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x)$ and state-control function $Q_t^\star(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x, \mathbf{u}_t)$ are no longer of simple analytic form. This is because the functions $m_t^{Q^\star,2}(\boldsymbol{\Sigma}_t^x), m_t^{V^\star,2}(\boldsymbol{\Sigma}_t^x)$ feature nested Kalman updates and are therefore non-linear and non-quadratic. Fortunately, due to the deterministic nature of the dynamics of the a-posterior covariance both value function and state-control function can exactly be evaluated. This is possible by means of a forward-pass starting in $\boldsymbol{\Sigma}_0^x$ and traversing equation (2.49).

LQGs inherit the computational efficiency of and are consequently similar popular in practice [11].

2.2 Models of Rational Sequential Decision Making

In the previous section we reviewed *optimal* sequential decision making in MDP and POMDP. While this provides an excellent conceptual framework for normative behavior, it is less suited for modeling realistic human behavior. This is because human decision making deviates from optimal policies in manifold ways what has long been discussed [212]. In simple discrete choice and gambling tasks,

humans have famously been shown to be biased from optimality by false beliefs and framing [98]. In more complex real world task like driving, there is the additional problem of accurately modeling the task in the first place. For example, in driving we do not exactly know all muscular and perceptual constraints that are imposed on a human driver. As a consequence, the optimal policy might deviate from the observed real world behavior simply because of inaccuracy of the MDP/POMDP model. Finally, individuals can all act optimal under the constraints defined by the MDP/POMDP model and still show marked differences in their policies. This is possible when they pursue different objectives. For example, in lane keeping this could be individual trade-offs between steering effort and deviation from the lane center.

For these reasons, modeling human behavior for predictive purpose often requires to relax optimality. Here, approaches are desired that allow to model variations in behavior and deviation from optimal behavior while preserving the key aspect of rationality. That is, controls that result in high return are more likely to be applied by the agent than controls that result in low return.

In the following section we will review popular frameworks for modeling rational behavior in the aforementioned sense. Here, our goal will be coverage of those frameworks that are suited for the application context of this work or have previously been used to modeling real world behavior.

2.2.1 Boltzmann Policies

A first framework to implement the desired model of rational behavior is given by so-called *Boltzmann policies*. Here, the agent applies controls u_t according to the distribution

$$\pi_t^\tau(u_t|x_t) = \frac{\exp(\frac{1}{\tau}Q^\star(x_t,u_t))}{\int \exp(\frac{1}{\tau}Q^\star(x_t,u_t'))\,\mathrm{d}\,u_t'}, \tag{2.53}$$

where τ is referred to as the *temperature* and where $Q^\star$ is the optimal state-control function. Here, the likelihood of a control is a monotonic function of the optimal state-control function. Hence, choosing sub-optimal controls is less likely than choosing optimal controls according to their decreased value of the optimal state-control function. This property of the Boltzmann policy leverages an interpretation as rational behavior. While initially popularized in the context of exploration in reinforcement learning [231, 232, 99], the Boltzmann policy model found increasing popularity for modeling human real-world behavior [256, 168]. A favorable property of this behavior model is that existing OC and RL techniques can be reused to obtain the state-control function, while the temperature τ allows to adjust the amount of "spread" around the optimal policy. However, the approach comes with a conceptual weakness: The agent using the Boltzmann policy must be considered overly optimistic with respect to his own controls. Although, actually applying a sub-optimal policy the agent plans with a *optimal* policy what

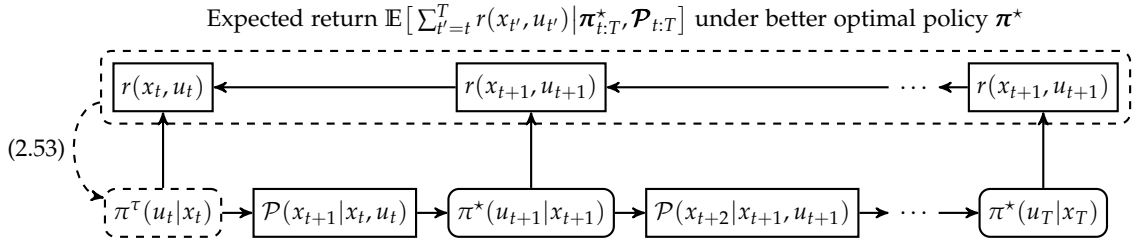

Figure 2.4: Illustration of the Boltzmann policy model. In the used stochastic policy $\pi^\tau(u_t|x_t)$ at times step t the likelihood of actions is related optimal state-control function $Q_t^\star(x_t,u_t)$. This quantity is the expected return when following the optimal policy $\pi_{t+1:T}^\star$.

is illustrated in Fig. 2.4. Specifically, it uses the optimal state-control function $Q_t^\star$ to decide on controls. Hence, the agent does not take into account potential failures to perform optimal. As a consequence, the controls most likely applied by the agent using the Boltzmann policy do not necessarily result in the highest return and otherwise.

2.2.2 The Maximum Causal Entropy Policy Model

As alternative to the Boltzmann policy model the *Maximum Causal Entropy* (MCE) policy model has been proposed [260, 257]. In contrast to the latter, here it is ensured that the most likely choice of controls according to the MCE policy $\tilde{\boldsymbol{\pi}}$ also results in highest expected returns. Specifically, the likelihood of actions *is* a monotonic function of the return when continuing to follow the MCE policy,

$$\tilde{\pi}_t(u_t|x_t) \propto \exp\Big(\mathbb{E}\big[\sum_{t'=t}^{T} r(x_{t'}, u_{t'})\big|\tilde{\boldsymbol{\pi}}_{t:T}, \boldsymbol{\mathcal{P}}_{t:T}\big]\Big), \tag{2.54}$$

as proven in [258], Theorem 6.10. As illustrated in Fig. 2.5 in the MCE model takes account of not performing optimal.

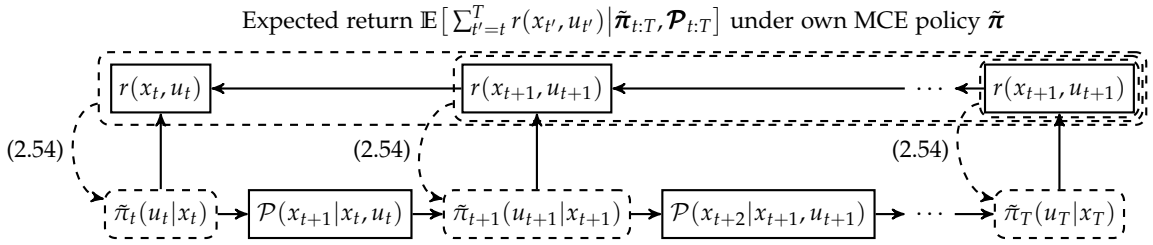

Figure 2.5: Illustration of the maximum causal entropy policy model. Here, the used stochastic policy $\tilde{\pi}(u_t|x_t)$ is related to the expected return when following policies $\tilde{\boldsymbol{\pi}}_{t+1:T}$ of the same policy model.

Markov Decision Processes and Partially Observable Markov Decision Processes

Although, the MCE model was developed for the special purpose of inverse optimal control, we will now review the policy model from the perspective of implementing sub-optimal yet rational behavior. It will be revisited in Cpt. 4.

Definition The maximum causal entropy policy $\tilde{\boldsymbol{\pi}}$ of an MDP with dynamics $\boldsymbol{\mathcal{P}}$, initial distribution p_0 and reward $r(x_t, u_t)$ is defined as the optimal solution of the optimization problem

$$\tilde{\boldsymbol{\pi}} := \arg\max_{\boldsymbol{\pi}} \mathcal{H}(\boldsymbol{\pi}) + \mathbb{E}\big[\sum_{t=0}^{T} r(x_t, u_t)\big|\boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\big]. \tag{2.55}$$

Here $\mathcal{H}(\boldsymbol{\pi})$ denotes the causal entropy of policy $\boldsymbol{\pi}$ given by

$$\mathcal{H}(\boldsymbol{\pi}) = -\sum_{t=0}^{T} \mathbb{E}\Big[\int \pi(u_t|x_t) \log \pi(u_t|x_t)\, \mathrm{d}\, u_t \Big| \boldsymbol{\pi}_{0:t-1}, \boldsymbol{\mathcal{P}}_{0:t-1}, p_0\Big]. \tag{2.56}$$

The role of the causal entropy term is to reward "broad" stochastic policies, i.e. where the probability mass is ideally equally distributed across controls.

Interestingly, the temperature parameter τ is absent in the MCE model. Instead, here the scale of the reward function $r(x_t, u_t)$ controls the spread of the control distribution. For example, multiplying the reward with a factor $\eta > 1$ puts more weight into obtaining high reward in the optimization problem (2.55),

$$\tilde{\boldsymbol{\pi}} := \arg\max_{\boldsymbol{\pi}} \mathcal{H}(\boldsymbol{\pi}) + \eta\mathbb{E}\big[\sum_{t=0}^{T} r(x_t, u_t)\big|\boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\big].$$

As a result $\tilde{\boldsymbol{\pi}}$ will be closer to the optimal policy $\boldsymbol{\pi}^\star$.

Policy Computation Using calculus of variation, [258] show that the maximizer of (2.55) can be obtained by means of the recursion

$$\tilde{Q}_t(x_t, u_t) = \begin{cases} r(x_t, u_t) + \mathbb{E}[\tilde{V}_{t+1}(x_{t+1}) | \mathcal{P}_t(x_{t+1}|x_t, u_t)] & \text{if } t < T \\ r(x_T, u_T) & \text{else} \end{cases} \tag{2.57}$$

$$\tilde{V}_t(x_t) = \log \int \exp(\tilde{Q}_t(x_t, u_t)) \, \mathrm{d}\, u_t =: \operatorname{softmax}_{u_t} \tilde{Q}_t(x_t, u_t) \tag{2.58}$$

$$\tilde{\pi}_t(u_t|x_t) = \exp(\tilde{Q}_t(x_t, u_t) - \tilde{V}_t(x_t)). \tag{2.59}$$

Due to similarity to the Bellman equations (2.5)-(2.7), the equations (2.57)-(2.59) are referred to as *soft* Bellman equations. Correspondingly, $\tilde{Q}_t(x_t, u_t)$ is termed soft state-control function and $\tilde{V}_t(x_t)$ is termed soft value function.

The MCE policy model is also well defined in POMDPs [258], where it can be obtained in a convenient form via an equivalent belief MDP similar to the optimal policy [38].

Similar to the classic Bellman equation the soft Bellman equations are often in-feasible for large discrete and continuous state spaces. This is because applying the soft Bellman equation requires full backup of the soft state-control. Therefore a variety of approximation techniques have been developed in recent years [31, 133, 85, 159].

Linear Quadratic Regulation and Linear Quadratic Gaussian Problems

Fortunately, computing MCE policies remains efficient for the special case of LQG[1][38] which we will review in this subsection. In the setting of linear affine dynamics with Gaussian noise and a linear Gaussian sensor model (2.32), (2.34), the soft state-control function and the soft value function are given by

$$\tilde{Q}_t(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x, \mathbf{u}_t) = [\boldsymbol{\mu}_t^x; \mathbf{u}_t]^\top \mathbf{M}_t^{\tilde{Q}} [\boldsymbol{\mu}_t^x; \mathbf{u}_t] + \mathbf{m}_t^{\tilde{Q}} [\boldsymbol{\mu}_t^x; \mathbf{u}_t] + m_t^{\tilde{Q},1} + m_t^{\tilde{Q},2}(\boldsymbol{\Sigma}_t^x) \tag{2.60}$$

$$\tilde{V}_t(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x) = [\boldsymbol{\mu}_t^x]^\top \mathbf{M}_t^{\tilde{V}} \boldsymbol{\mu}_t^x + \mathbf{m}_t^{\tilde{V}} \boldsymbol{\mu}_t^x + m_t^{\tilde{V},1} + m_t^{\tilde{V},2}(\boldsymbol{\Sigma}_t^x), \tag{2.61}$$

with negative definite matrices $\mathbf{M}_t^{\tilde{Q}}, \mathbf{M}_t^{\tilde{V}}$. As $\tilde{Q}_t$ results from $\tilde{V}_t$ according to the same equation as $Q_t^\star$ from $V_t^\star$ (compare (2.5) and (2.57)), this yields the elements $\mathbf{M}_t^{\tilde{Q}} \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, m_t^{\tilde{Q},2}$ as

$$\mathbf{M}_t^{\tilde{Q}} = \begin{cases} [\mathbf{A}_t \, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}} [\mathbf{A}_t \, \mathbf{B}_t] - \operatorname{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{if } t < T \\ -\operatorname{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{else} \end{cases} \tag{2.62}$$

$$\mathbf{m}_t^{\tilde{Q}} = \begin{cases} 2[\mathbf{A}_t \, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}} \mathbf{a}_t + [\mathbf{A}_t \, \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{\tilde{V}} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \tag{2.63}$$

$$m_t^{\tilde{Q},1} = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{\tilde{V}} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{\tilde{V}} + m_{t+1}^{\tilde{V},1} & \text{if } t < T \\ 0 & \text{else} \end{cases} \tag{2.64}$$

$$m_t^{\tilde{Q},2}(\boldsymbol{\Sigma}_t^x) = \begin{cases} \operatorname{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^x) + \operatorname{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}} \left(\mathbf{A}_t \boldsymbol{\Sigma}_t^x \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)\right)\right) + m_t^{\tilde{V},2}\left(\boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)\right) & \text{if } t < T \\ 0 & \text{else} \end{cases}. \tag{2.65}$$

$\tilde{V}_t$ is obtained from $\tilde{Q}_t$ by applying the softmax operator,

$$\tilde{V}_t(\boldsymbol{\mu}_t^x, \boldsymbol{\Sigma}^x) = \log\Big(\int \exp\left([\boldsymbol{\mu}_t^x; \mathbf{u}_t]^\top \mathbf{M}_t^{\tilde{Q}} [\boldsymbol{\mu}_t^x; \mathbf{u}_t] + \mathbf{m}_t^{\tilde{Q}} [\boldsymbol{\mu}_t^x; \mathbf{u}_t] + m_t^{\tilde{Q},1} + m_t^{\tilde{Q},2}(\boldsymbol{\Sigma}_t^x)\right) \mathrm{d}\, u_t \Big). \tag{2.66}$$

Following [258][2], [38] the expression (2.66) can be simplified to

[1] including LQR [261]
[2] Theorem 6.10

$$\mathbf{M}_t^{\tilde{V}} = \mathbf{M}_{t\,x,x}^{\tilde{Q}} - \mathbf{M}_{t\,x,u}^{\tilde{Q}}[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}\mathbf{M}_{t\,u,x}^{\tilde{Q}} \tag{2.67}$$

$$\mathbf{m}_t^{\tilde{V}} = \mathbf{m}_{t\,x}^{\tilde{Q}} - \mathbf{M}_{t\,x,u}^{Q^\star}[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}\mathbf{m}_{t\,u}^{\tilde{Q}} \tag{2.68}$$

$$m_t^{\tilde{V},1} = m_t^{\tilde{Q},1} - \frac{1}{4}[\mathbf{m}_{t\,u}^{\tilde{Q}}]^\top[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}\mathbf{m}_{t\,u}^{\tilde{Q}} + \frac{1}{2}\log(\det(\pi[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1})) \tag{2.69}$$

$$m_t^{\tilde{V},2} = m_t^{\tilde{Q},2}(\mathbf{\Sigma}_t^x). \tag{2.70}$$

Finally, applying Gaussian conditioning (see e.g. [175]) yields the MCE policy $\tilde{\pi}_t$ as

$$\begin{aligned}\tilde{\pi}_t(\mathbf{u}_t|\boldsymbol{\mu}_t^x) = \exp\Big([\boldsymbol{\mu}_t^x;\mathbf{u}_t]^\top\mathbf{M}_t^{\tilde{Q}}[\boldsymbol{\mu}_t^x;\mathbf{u}_t] + \mathbf{m}_t^{\tilde{Q}}[\boldsymbol{\mu}_t^x;\mathbf{u}_t] + m_t^{\tilde{Q},1} + m_t^{\tilde{Q},2}(\mathbf{\Sigma}_t^x)\\ - \left([\boldsymbol{\mu}_t^x]^\top\mathbf{M}_t^{\tilde{V}}\boldsymbol{\mu}_t^x + \mathbf{m}_t^{\tilde{V}}\boldsymbol{\mu}_t^x + m_t^{\tilde{V},1} + m_t^{\tilde{V},2}(\mathbf{\Sigma}_t^x)\right)\Big)\end{aligned} \tag{2.71}$$

$$= \mathcal{N}(\mathbf{u}_t|\tilde{\mathbf{F}}_t\boldsymbol{\mu}_t^x + \tilde{\mathbf{f}}_t, \mathbf{\Sigma}_t^{\mathbf{u}}) \tag{2.72}$$

$$\tilde{\mathbf{F}}_t := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}\mathbf{M}_{t\,u,x}^{\tilde{Q}}, \quad \tilde{\mathbf{f}}_t := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}\mathbf{m}_{t\,u}^{\tilde{Q}}, \quad \mathbf{\Sigma}_t^{\mathbf{u}} := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}. \tag{2.73}$$

Comparing the recursions for the MCE policy (2.62)-(2.72) and the optimal policy (2.12)-(2.22) there are only two main differences. First, $m_t^{\tilde{V},1}$ features the additional summand $\frac{1}{2}\log(\det(\pi[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}))$ which corresponds to the additional causal entropy term in the MCE objective (2.55). Second, the MCE policy is a stochastic conditional Gaussian policy in contrast to the deterministic (conditional Dirac) optimal policy. However, both policies have the same conditional mean, i.e. $\tilde{\mathbf{F}}_t = \mathbf{F}_t^\star \quad \tilde{\mathbf{f}}_t = \mathbf{f}_t^\star$, as $\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}$, and $\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}$, coincide. In Cpt. 3 computing MCE policies in deterministic LQRs is required to solve a more complex POMDP. These policies can be computed by means of Algo. 3.

Algorithm 3 MCE policy of LQR LQRMCE

Require: all

function LQRMCE($\mathbf{C}_x, \mathbf{C}_u, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}$)
  $\mathbf{M}_{T+1}^{\tilde{V}} \leftarrow \mathbf{0}^{n_x,n_x}$
  $\mathbf{m}_{T+1}^{\tilde{V}} \leftarrow \mathbf{0}^{1,n_x}$
  $m_{T+1}^{\tilde{V},1} \leftarrow 0$
  for $t = T, \dots, 0$ **do**
    $\mathbf{M}_t^{\tilde{Q}} \leftarrow [\mathbf{A}_t\,\mathbf{B}_t]^\top\mathbf{M}_{t+1}^{\tilde{V}}[\mathbf{A}_t\,\mathbf{B}_t] - \mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u)$
    $\mathbf{m}_t^{\tilde{Q}} \leftarrow 2[\mathbf{A}_t\,\mathbf{B}_t]^\top\mathbf{M}_{t+1}^{\tilde{V}}\mathbf{a}_t + [\mathbf{A}_t\,\mathbf{B}_t]^\top\mathbf{m}_{t+1}^{\tilde{V}}$
    $\mathbf{U} = [\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}$
    $m_t^{\tilde{Q},1} \leftarrow \mathbf{a}_t^\top\mathbf{M}_{t+1}^{\tilde{V}}\mathbf{a}_t + 2\mathbf{a}_t^\top\mathbf{m}_{t+1}^{\tilde{V}} + m_{t+1}^{\tilde{V},1}$
    $\mathbf{M}_t^{\tilde{V}} \leftarrow \mathbf{M}_{t\,x,x}^{\tilde{Q}} - \mathbf{M}_{t\,x,u}^{\tilde{Q}}\mathbf{U}\,\mathbf{M}_{t\,u,x}^{\tilde{Q}}$
    $\mathbf{m}_t^{\tilde{V}} \leftarrow \mathbf{m}_{t\,x}^{\tilde{Q}} - \mathbf{M}_{t\,x,u}^{\tilde{Q}}\mathbf{U}\,\mathbf{m}_{t\,u}^{\tilde{Q}}$
    $m_t^{\tilde{V},1} \leftarrow -\frac{1}{4}[\mathbf{m}_{t\,u}^{\tilde{Q}}]^\top\mathbf{U}\,\mathbf{m}_{t\,u}^{\tilde{Q}} + m_t^{\tilde{Q},1} + \frac{1}{2}\log(\det(\pi[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}))$
    $\tilde{\mathbf{F}}_t \leftarrow -\frac{1}{2}\mathbf{U}\,\mathbf{M}_{t\,u,x}^{\tilde{Q}}$
    $\tilde{\mathbf{f}}_t \leftarrow -\frac{1}{2}\mathbf{U}\,\mathbf{m}_{t\,u}^{\tilde{Q}}$
    $\mathbf{\Sigma}_t^{\mathbf{u}} \leftarrow -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{\tilde{Q}}]^{-1}$
  end for
  return $\left(\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \mathbf{\Sigma}_t^{\mathbf{u}}\right)_{t=0:T}$
end function

2.3 Conclusion

This chapter introduced MDPs and POMDPs as frameworks for optimal decision making. Here, POMDPs also allow to consider aspects of sensing and perception which will be relevant for modeling appropriate glance behavior. In addition to this, the fundamental solution technique using the Bellman equation was introduced and applied to the classes of linear quadratic regulation and linear quadratic Gaussian problems. Finally, we discussed two approaches namely Boltzmann and maximum causal entropy policies that allow to model sub-optimal yet rational behavior which is more realistic for human real-world decision making.

3 Appropriate Glance Behavior in the Joint Task of Driving and Secondary Task Interaction

To assess driver behavior with respect to the situational context, one needs to specify appropriate behavior. That is, a *normative* model of behavior is required. In this chapter we first derive a joint partially observable Markov decision processes model of the driving situation (Sec. 3.3.1), the driver's sensing of the driving situation (Sec. 3.3.2) and the potentially distracting secondary task (Sec. 3.3.3). Given this POMDP appropriate glance behavior can precisely be defined by means of a tolerance on deviation from rational policies therein (Sec. 3.4). This allows to effectively support the driver by means of warning system. Furthermore, computing both optimal and rational policies in the joint task POMDP is addressed leading to several algorithmic solution approaches (Sec. 3.5). Finally, (Sec. 3.6) analyses the realism of the computed glance behavior and aspects of computational feasibility which are both highly relevant for development of a real-time distraction warning system.

Parts of this chapters have previously been published in [203, 202].

3.1 Introduction

In the review of the state of the art in research (Sec. 1.1), we unveiled a gap between the work of different research communities. Specifically, human factors research and cognitive science found convincing evidence for human situationally adaptive glance behavior in driving. However, the state-of-the-art algorithmic approaches to assessment of attention do not consider the situational context. Some kinds of adaptation of drivers is undesirable, e.g. adapting longer glances off the road due to overtrust in an imperfect partially automated driving system, but most of it is intuitively reasonable. This is the case, for example in adapting shorter off-road glances at higher speeds to maintain an acceptable lane position [207]. Hence, ideally distraction assessment establishes a mathematical relation between glance patterns and objectively defined risk and performance measures in the given driving situation. For example, the vehicle's increased deviation from the lane center when steered by the driver who is not looking at the road is such an objective measure. As pointed out by Sheridan in his opinion paper [209], the desired relation can be established taking a control/decision theoretic perspective to driver distraction. That is, to model the glance behavior in interaction with a secondary task in driving in the formalism of partially observable Markov decision processes introduced in the previous chapter. Given a suitable POMDP model of the joint task of driving and secondary task engagement, we can compute optimal or rational policies therein. These can thereafter be used to numerically define appropriate behavior. Moreover, situationally specific appropriate glance behavior can be obtained. This is possible by modeling the variation among driving situations by means of a parametric POMDP structure. Once the POMDPs situation parameters are known in action, the corresponding policies can be computed resulting in adaption to the specific circumstances.

In the context of this thesis, we will consider modeling appropriate glance behavior in the exemplary manual driving task lane keeping in presence of a secondary task. In the course of modeling both tasks in a POMDP, a desirable goal is to precisely address the vehicle dynamics, characteristics of driver's sensing and the secondary task. As a consequence, these aspects will be taken into account in the computed appropriate glance behavior. However, in general solving POMDPs has a very high complexity [171]. Consequently, we must exercise caution in the model development to ensure that the POMDP permits tractable solution. As computing appropriated glance behavior is used in a real-time distraction warning system therefore sometimes only the fundamental aspects of the tasks can be addressed in the POMDP model. Furthermore, highly efficient solution approaches must be developed.

3.2 Related Work

Lane keeping is the most fundamental task in driving on public streets, roads and motorways. Except for special occasions the vehicle needs to stay in its lane to prevent collisions with static objects and other traffic participants. Consequently, there has long been research interest in modelling this driving task. In the last two decades more and more work has focused on development of lane tracking and automatic lane keeping algorithms for assisted and automated driving [151]. Control theoretic modeling of human manual control in lane keeping can be dated back at least as far as the early 60s as pointed out in the comprehensive overviews [150, 178]. From early on, explicit elements of driver sensing and perception have been incorporated e.g. in [113]. Recently, the optimal fit of the parameters of these models to experimental data have been found to be sensitive to various types of distraction [78].

Posing lane keeping as a linear quadratic Gaussian POMDP has first been proposed in [149]. The approach has been extended to incorporate neuro-muscular aspects [42] and non-linear vehicle dynamics [103].

Baron and Kleinman were the first to extend LQG models for manual control with respect to switching of the focus of attention among a small disrete set of options [19, 20, 109]. Here, they merged previous control theoretic models with those of human information gathering in manual control [207]. To the best of our knowledge, the approach has found a single previous application to modeling manual lane keeping in the context of behavior under visual occlusion [26]. Baron and Kleinman's original model was further extended in [173] to incorporate secondary tasks in discrete variables, similar as considered in this thesis.

Moreover, some authors have also considered more complex POMDP models of human real-world behavior: Modeling reaching movements using linear quadratic optimal control with multiplicative and additive noise was proposed in [240]. Here, a locally optimal linear filter and a globally optimal policy was obtained using coordinate descent. [220, 191] considered a POMDP model for walkway navigation including avoiding obstacles and collecting targets. This was computationally approached by decomposing the task into sub-tasks and applying an arbitration heuristic. The task of catching balls under sensor and control constraints was modeled as a nonlinear POMDP by [23]. [57] addressed hand eye coordination in reaching. In both works an approximate solution was obtained by solving deterministic substitute MDPs in the belief space [56, 244].

Finally, secondary task interaction and its influence on lane keeping performance have been modeled in the human computer interaction community [34, 35, 91, 90, 121]. However, in those works only crude models of vehicle dynamics and manual control thereof were used. Furthermore, no efficient algorithmic approaches for computing appropriate behavior were considered.

In this thesis we employ the model approach proposed by [19] and consider the discrete binary option x_t^z that denotes whether the gaze of the driver is on the road or the gaze is off the road. This is a feasible approach for the scenario where the driver switches gaze between the forward road scenery and a display relevant for a secondary task. [26] applied the model of [19] for a simulation study of lane-keeping under fixed glance behavior comprising by a interval where the driver is looking at the road and an interval were the driver's vision is fully occluded. In contrast we consider computing optimal and rational policies. Furthermore, the secondary task causing aversion of gaze from the road and the driver's sensing characteristics are addressed in the model. [173] used a heuristic approach to compute gaze switching policies with uncertain performance. However, this has the disadvantage that the performance of heuristically obtained gaze switching policies can interact with the specific model parametrization. In contrast, this work considers exact computation of optimal and rational policies. Throughout this chapter, we use a pragmatic modeling paradigm in regards to application in a real-time warning system.

3.3 Modeling the Joint Task of Driving and Secondary Task Interaction

We first consider development of a suitable partially observable Markov decision process model of the joint task of driving and secondary task interaction. Importantly, our ultimate goal is to compute glance strategies in real-time in a vehicle. Therefore, physical realism, biological plausibility and granularity

of the secondary task model must be balanced against practical considerations. Here it is crucial that, first, the POMDP model admits efficient solution. Second, the states of the model must be measurable in an instrumented vehicle. In the following, we first consider the primary task of vehicle control and derive both a kinematic model of the task as well as the objective of the task in form of a reward function (Sec. 3.3.1). Thereafter, a practical model of the driver's sensor characteristics is introduced (Sec. 3.3.2). Finally, we address modeling a visually demanding secondary task with objectives given by an additional reward function (Sec. 3.3.3) and summarize the joint task model in (Sec. 3.3.4).

3.3.1 The Primary Task of Manual Lane Keeping

As noted in the introduction, this thesis considers modeling glance behavior in the task of lane keeping as an exemplar driving task. Here, aspects of the driving situation, i.e. the track topology and the driving speed, the vehicle dynamics, i.e. the lateral dynamics, the driver's control input, i.e. steering velocity, and finally the task's objective must be considered. The task of lane keeping will be referred to as the *primary task* throughout this thesis. Correspondingly, states, control etc. related to the driving task will also be referred to as primary task states, primary task controls etc.

Situation and Vehicle Dynamics

The first step towards modeling manual lane keeping is to consider the dynamics of the driving situation and the driven vehicle. Generally, the dynamics of an automobile are highly nonlinear and complex due to the characteristic of tires and aerodynamics [93]. Fortunately, simplified models are available that are realistic in almost all driving situations encountered in real traffic.

Kinematic Model One simple model of the vehicle and the situational dynamics that has been utilized in automatic lane keeping systems [188, 189] is the kinematic model which is derived in the following. Given a signed distance y_t to the lane center line l_c, the orientation with respect to l_c, i.e. the angle ϕ_t

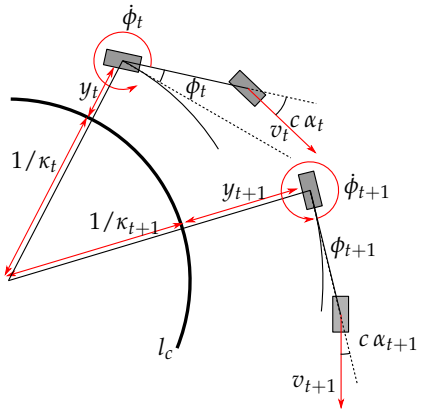

Figure 3.1: Illustration of the variables of the kinematic model

between the vehicle's longitudinal axis and the tangent to l_c and the vehicle's absolute velocity v_t, the derivative of y_t, $\dot{y}_t$ is given by

$$\dot{y}_t = v_t \sin(\phi_t) \tag{3.1}$$

The vehicles orientation in the lane ϕ_t is influenced by both the curvature of the lane κ_t and the angular velocity induced by the orientation of the front wheels wrt. the longitudinal axis, the so-called effective steering angle β_t. We assume a constant transmission ratio c_1 of the steering wheel, that is $\beta_t = c_1\alpha_t$ with the steering angle α_t. When the vehicle is moving on a circle arc of radius $1/\kappa_t$ with speed v_t, its angular velocity $\hat{\dot{\phi}}_t$ is

$$\hat{\dot{\phi}}_t = \kappa_t v_t \tag{3.2}$$

and its orientation wrt. the tangent of the arc is zero $\varphi_t = 0$. Hence, if the vehicle keeps a straight course instead, i.e. $\alpha_t = 0$, the change of orientation $\dot{\varphi}_t$ of the vehicle wrt. to l_c is

$$\dot{\varphi}_t = -\hat{\dot{\phi}}_t = -\kappa_t v_t. \tag{3.3}$$

An effective steering angle β_t causes a velocity $v_t^l = v_t \sin(\beta_t)$ in direction of the vehicle's lateral axis. This results in an angular velocity $\bar{\dot{\phi}}_t$ the so-called *yaw-rate* of

$$\bar{\dot{\phi}}_t = c_2 v_t^l = c_2 v_t \sin(\beta_t) = c_2 v_t \sin(c_1 \alpha_t), \tag{3.4}$$

where c_1 is a constant related to the distance between front and rear wheel axes. Hence, the vehicles orientation wrt. to a straight line, i.e. $\kappa_t = 0$ changes according to

$$\dot{\phi}_t = \bar{\dot{\phi}}_t = c_2 v_t^l = c_2 v_t \sin(\beta_t) = c_2 v_t \sin(c_1 \alpha_t). \tag{3.5}$$

Altogether, the vehicle's change of orientation with respect to the lane center line is given by

$$\dot{\phi}_t = c_2 v_t \sin(c_1 \alpha_t) - \kappa_t v_t. \tag{3.6}$$

Finally, the driver changes the steering angle by means of its velocity $\dot{\alpha}_t$. The introduced variables of the kinematic model are summarized in the Tab. 3.1.

Tabular 3.1: Variables of Kinematic Vehicle Model

Symbols	Definitions	Units
y	lateral position wrt. lane center line l_c	m
ϕ	orientation angle between tangent of lane center line l_c and vehicle's longitudinal axis	rad
$\dot{\phi}$	yaw-rate	rad/s
β	effective steering angle	rad
α	steering angle	rad
v	vehicle's absolute velocity	m/s
κ	curvature of lane	1/m
c_1	steering wheel transmission ratio	
c_2	front wheel angle to yaw-rate constant	

Linear Approximations In the overall majority of driving situations angles ϕ_t and β_t are rather small, which allows to use linear approximation for the nonlinear trigonometric functional relations (3.1)-(3.6). For example, Fig. 3.1 shows the distribution of the speed perpendicular to the lane $v_t \sin(\phi_t)$ and the distribution of absolute and relative approximation error on the data of driving experiment II (Cpt. 5.6). As can be seen in the left plot, the spread of the distribution of the signed approximation error $p(v_t\phi_t -$

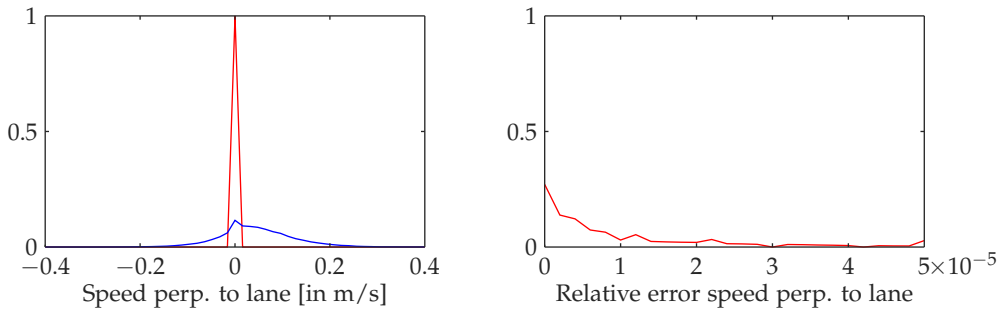

Figure 3.2: Errors induced by linear approximation of (3.1) on data of experiment II. Distribution of $v_t \sin(\phi_t)$ and $v_t\phi_t - v_t \sin(\phi_t)$ **(left)**, distribution of the relative error $\frac{|v_t\phi_t - v_t \sin(\phi_t)|}{|v_t \sin(\phi_t)|}$ **(right)**

$v_t \sin(\phi_t))$ is significantly smaller than the spread of the distribution of the speed perpendicular to the

lane $v_t \sin(\phi_t)$. Furthermore, the right plot shows that the relative approximation error $\frac{|v_t\phi_t - v_t\sin(\phi_t)|}{|v_t\sin(\phi_t)|}$ is low in almost all cases $p(\frac{|v_t\phi_t - v_t\sin(\phi_t)|}{|v_t\sin(\phi_t)|} < 2.69 \times 10^{-5}) \geq 0.9$ and the mean relative approximation error is at 1.04×10^{-5}.

Similar results hold true for the trigonometric relation between steering angle and yaw-rate $\bar{\dot{\phi}}$ (3.4). We first used least-squares fitting to obtain the constants c_1, c_2 using the data of experiment II. Thereafter the approximation error is evaluated. The distribution of $c_2 v_t \sin(c_1 \alpha_t)$, the approximation error and the relative approximation error are depicted in Fig. 3.3. Here, similar as in the previous cases

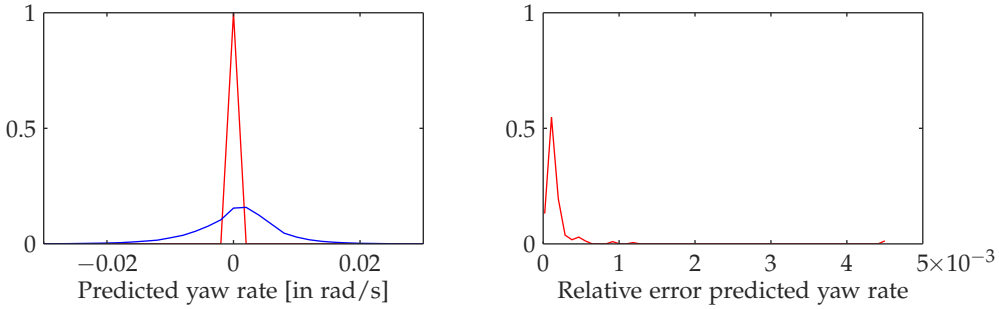

Figure 3.3: Errors induced by linear approximation of (3.4) on data of experiment II. Distribution of $c_2 v_t \sin(c_1 \alpha_t)$ and $c_2 v_t \sin(c_1 \alpha_t) - c v_t \alpha_t$ **(left)**, distribution of the relative error $\frac{|c_2 v_t \sin(c_1 \alpha_t) - c v_t \alpha_t|}{|c_2 v_t \sin(c_1 \alpha_t)|}$ **(right)**

the approximation error turns out to be neglectable: In most of the data the relative approximation error is small $p(\frac{|c_2 v_t \sin(c_1 \alpha_t) - c v_t \alpha_t|}{|c_2 v_t \sin(c_1 \alpha_t)|} < 2.76 \times 10^{-4}) \geq 0.9$ and the mean relative approximation error is of 2.15×10^{-4}.

Consequently, the following linear approximation

$$\begin{bmatrix} \dot{y}_t \\ \dot{\phi}_t \end{bmatrix} = \begin{bmatrix} v_t \phi_t \\ \bar{c}_1 v_t \alpha_t - \kappa_t v_t \end{bmatrix}. \tag{3.7}$$

of the original kinematic model were employed to model the dynamics of the vehicle and its position and orientation in lane throughout this thesis. To obtain a discrete time version of the above system of differential equation these were integrated at 25 Hz.

Primary Task Dynamics After considering the kinematics of lane keeping, the dynamics of the primary task can be formulated. Summarized, these are given by the linear affine dynamics in primary task states $\mathbf{x}_t^{\mathrm{p}} = [y_t\, \dot{y}_t\, \phi_t\, \alpha_t]^\top$ and primary task control $u_t^{\mathrm{p}} = \dot{\alpha}$

$$\mathbf{x}_t^{\mathrm{p}} = \mathbf{A}(v_t)\mathbf{x}_t^{\mathrm{p}} + \mathbf{B}(v_t)u_t^{\mathrm{p}} + \mathbf{a}(v_t, \kappa_t) + \boldsymbol{\epsilon}_t^{\mathrm{p}}, \quad \Delta t = 1/(25\ \mathrm{Hz}), \quad \boldsymbol{\epsilon}_t^{\mathrm{p}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\epsilon^{\mathrm{p}}}) \tag{3.8}$$

Importantly, the noise in the dynamics of the steering angle α_t has zero variance, i.e. $\boldsymbol{\Sigma}^{\epsilon^{\mathrm{p}}}_{\alpha,\alpha} := \boldsymbol{\Sigma}^{\epsilon^{\mathrm{p}}}_{4,4} = 0$. This because the control u_t^{p} is the derivative of the steering angle α_t and hence both quantities are deterministically related.

In our kinematic model of lane keeping external parameters the vehicle's velocity $\boldsymbol{v}_{0:T}$ and the lane curvature $\boldsymbol{\kappa}_{0:T}$ are present. These parameters are explicitly allowed to be time dependent. Although the vehicle's velocity is controlled by the driver, in this thesis it is considered also an external parameter. This is because empirically we observed that the vehicle's velocity is weakly controlled by the driver once he or she has started engaging into a secondary task. Instead, the velocity appears to be reduced by the driver in an open-loop fashion right before engagement [154]. Both the vehicle's velocity and the track curvature are not simply disturbances of the primary task model. Instead, both quantities can significantly alter the kinematic model. Consequently, different parameter value require different steering policies and different glance behavior (see Sec. 3.5). In this thesis it is assumed that the vehicles velocity and the lane curvature already describe the variety of encountered driving situations in lane-

keeping. Hence, if the specific policy for the parameter values is computed, situational adaptability is obtained.

Note that the states and controls present in the kinematic model are all directly available in series vehicles: Velocity, yaw-rate and steering angle are used to realize electronic stability control. The other variables can be estimated by means of a lane tracking camera system as described in [188, 189] which is also available in modern vehicles. Consequently, also the parameters of the vehicle can be estimated online.

Task Objective

Previously, a model of the vehicle dynamics controllable by the driver by means of the steering angle velocity was derived. In addition to that the POMDP model requires a reward that measures the performance in the primary task. For this purpose, we employ the quadratic reward function

$$r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) = \theta_1 (y_t)^2 + \theta_2 (\dot{y}_t)^2 + \theta_3 (\alpha_t)^2 + \theta_4 (u_t^{\mathrm{p}})^2. \tag{3.9}$$

In this context, the quantities $\theta_1, \theta_2, \theta_3, \theta_4 < 0$ are parameters that weight lane keeping performance against the steering effort necessary to obtain it. The rationale behind the individual terms in the reward function is discussed in the following.

1. Term $\theta_1 (y_t)^2$ penalizes deviation from the lane center, which is an obvious objective in lane keeping. It has been used in previous work [149, 26, 43] and is also present in the *root mean squared lane deviation* $\sqrt{\mathbb{E}[(y_t)^2]}$ frequently used in human factors research in distraction [252](Cpt. 7, *Measuring the Effects of Driver Distraction*).

2. Similar as in [26] the term $\theta_2 (\dot{y}_t)^2$ is employed because it relates to the impact energy in cases of a collision due to lane departure. Furthermore, empirically drivers tend to focus on keeping their current position in lane if it is not to close to the lane borders instead of seeking zero deviation from the lane center [69]. For these reasons a main metric in human factors research is the standard deviation of the lane position $\sqrt{\mathbb{E}[(y_t)^2] - \mathbb{E}[y_t]^2}$ [252](Cpt. 7) that is more closely related to the velocity in direction of the lane borders in driving.

3. Term $\theta_3 (u_t^{\mathrm{p}})^2$ contributes to modeling the driver's steering effort. A non-zero steering angle induces an approximately proportional counter-directed force on the driver's arms [93]. Hence, it can be assumed that ideally the steering angles are as small as possible.

4. Term $\theta_4 (\dot{\alpha}_t)^2$ serves two main purposes: Sudden changes of the steering angle are stressful for the driver and must therefore be considered in the steering effort. Furthermore, the steering angle is the control input modality of the driver in our model of lane keeping. Hence, a penalty imposed on its squared derivative, i.e. the squared steering wheel velocity, enforces low frequency control input. As suggested in [176] this approach can be used to model neuro-muscular delays present in human operators and was also applied in [26]. Finally, the term is also used as a standard metric in human factors research [252](Cpt. 7).

In the preceding list we introduced and motivated the individual terms for the primary task reward function. Several reference to previous work that employed similar objectives were given. In addition to the reward terms, computing glance policies in application requires *numerically* specified parameters $\boldsymbol{\theta}$ of the reward function. In [149, 26, 43] the parameters of the individual parts of the reward function are reported, but as we use a different combination of terms these parameters are not applicable. Furthermore, neither of these works gives a derivation of the applied weighting coefficients. Instead, the values seem to have been obtained by guessing. As will be shown in Sec. 3.5 the reward of lateral vehicle control significantly influences the computed appropriate glance behavior. Therefore, instead of guessing the parameters $\boldsymbol{\theta}$ are better empirically obtained. This issue will be addressed in Cpt. 4 where we derive new techniques for inference of these parameters from behavioral data.

3.3.2 Driver's Sensor Characteristics

As already noted in the introduction, vision is the predominant sensor modality that humans rely on in automobile driving. However, human's vision is subject to several limitations. Most important, its performance shows a drastic decrease from the most central part of gaze, the *fovea*, to the periphery [36, 61, 226]. This is the case for the capability of recognizing visual patterns. For example, in one classic experiment [249] visual acuity was measured by the distance to the eye at which individual cells of a wire grid were recognizable. These distances at different angular eccentricities from the fovea were shown to follow a heavy-tailed function which is depicted in Fig. 3.4.

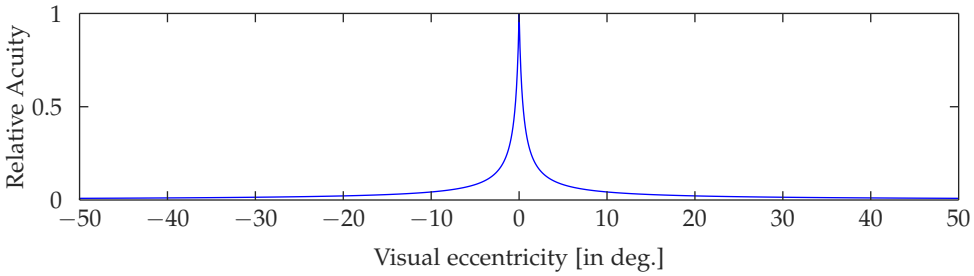

Figure 3.4: Measured relative visual acuity dependent on angular eccentricity from the fovea. Values correspond to the fraction of the distance at which grid cells were recognizable in relation to viewing the grid at the fovea (eccentricity of 0 degree). Redrawn from [249].

Correspondingly, the region where the sensor capabilities are highest and sufficient for visually demanding activities, such as e.g. reading, spans only a few degrees.

In lane keeping, the driver's gaze was found to be concentrated on road corners and line markings [124]. Hence, it has been concluded that these regions contain visual key features for perception of self-movement [71] and for anticipatory steering control [124]. With respect to the states and parameters of the model of lane keeping, directing the fovea towards lane marking and road corners allows to sense both the in lane and orientation in lane as well as the curvature of the lane.

Besides the high acuity vision in the fovea, humans are able to detect a reduced amount of visual stimuli in peripheral regions [17]. As [124] correctly hypothesized, peripheral vision significantly contributes to sensing in driving which has been experimentally shown in [230, 123].

Typical visually demanding secondary task drivers are engaging in, e.g. interaction with the vehicle's infotainment system, require to gaze at objects in the vehicle interior. Here, in most cases the visual angle between the direction of the road scenery and the fovea exceeds 5 degrees. As a consequence of the decreased visual acuity, the visual information from the road scenery obtained by the driver drops as illustrated in Fig. 3.5.

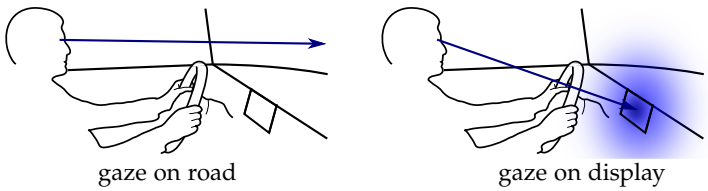

Figure 3.5: Illustration of the visual acuity in driving dependent on gaze direction. The distribution of acuity is illustrated as blue region with opacity decreasing with the deviation from the fovea.

Consequently, the performance of lane keeping control [230] and lane keeping control [123] decreases as a monotonic function of the angular deviation of gaze from the forward road viewing direction.

Sensor Model and Dynamics Despite the aforementioned sensorial limitations humans are very successful in utilizing vision to perform a huge variety of tasks. The reason is that humans apply active vision, i.e. use policies to plan when and where to gather sensorial information necessary for task completion [61]. The required eye movements are commonly classified into *fixations*, *smooth pursuits*

and *saccades*. Fixations occur when the human looks at a static object with almost no change in the gaze angle. If the human visually tracks a moving object, continuously changing the gaze angle, this is classified as smooth pursuit. Finally, switching gaze from one to another object is conducted by saccades. These are rapid eye movement with angular velocities of several hundred degrees per second [131]. Although, it is possible to explicitly model these movement of gaze as in [56], in this thesis a simpler model is used. Similar as in [20] we consider the binary *sensor state* x_t^z,

$$x_t^z \in \{0 := \text{gaze on road}, 1 := \text{gaze off road}\}, \tag{3.10}$$

that can be switched by a binary *sensor control* u_t^z

$$u_t^z \in \{0 := \text{keep current sensor state}, 1 := \text{switch sensor state}\}, \tag{3.11}$$

$$x_{t+1}^z = x_t^z \oplus u_t^z \tag{3.12}$$

where $\oplus$ denotes the logical xor operator.

The rationale behind this is to simplify computing optimal/rational glance policies as will be shown in Sec. 3.5. Furthermore, in lane keeping eye movements consist of primarily saccades and fixations. Fixations are well represented in the discrete states. In addition to that, *ongoing* saccades account for a very small proportion of sample instances at the model frequency of 25 Hz (3.8) and hence can be approximated by switches between the sensor states.

Given a sensor state x_t^z, the sensing of the road scenery is modeled by a linear Gaussian sensor model

$$\mathbf{z}_t \sim \mathcal{N}(\mathbf{H}, \mathbf{\Sigma}^{\epsilon^z}(x_t^z)), \tag{3.13}$$

where there is a static sensor matrix $\mathbf{H}$ and sensor noise covariance $\mathbf{\Sigma}^{\epsilon^z}$ that depends on the sensor state. Specifically, the parameterization

$$\mathbf{H}(x_t^z) = \text{diag}(1, 0, 1, 1) \tag{3.14}$$

$$\mathbf{\Sigma}^{\epsilon^z}(x_t^z) = \text{diag}((\sigma_y)^2(x_t^z), 0, (\sigma_\phi)^2(x_t^z), 0) \tag{3.15}$$

is employed in this work.

Assumptions The underlying model assumptions are explained in the following. First, it is assumed that only states geometrically related to the road scenery can be sensed by vision, while their derivatives, are obtained by temporal differentiation. Second, the steering angle is assumed to be perfectly sensable. Surely, this is an unrealistic assumption. However, the POMDP formalism assumes that the applied controls are known to the agent. Consequently, the driver has full knowledge of the steering angle, as it is deterministically given by its velocity which is the driver's control. The effects of this model assumption will be compensated by sub-optimal choice of controls u_t^p.

Linear Gaussian sensor models must be considered a rather crude approximation to human sensing and cannot directly be related to the physiology of the visual organs. Indeed, in many works on signal detection, non-Gaussian sensor models haven been applied [110, 225, 4]. Nevertheless, in almost all POMDP approaches to modeling human real-world reviewed in Sec. 3.2 linear Gaussian sensor models have been used. This is because they render formulation of the belief computationally tractable and allow to at least approximately solve the POMDPs. Furthermore, we additionally need to specify the parameters of the sensor model. Even for this simple linear Gaussian model no realistic values are known in the literature. In more complex sensor models we can expect this issue to be even more pronounced. Hence, we address estimating these sensor models separately in Cpt. 5. It will be shown that in the case of linear Gaussian model an efficient and exact inference procedure can be obtained.

Reward For the purpose of computing of appropriate glance behavior, a second reward function is employed. Similar as in [19, 20, 109], a penalty on gaze switches

$$r(u_t^z) = \theta_5 u_t^z, \tag{3.16}$$

with parameter $\theta_5 > 0$ is imposed. Although, we forwent considering saccadic eye movements in the sensor states their characteristics are addressed in the reward function. Saccades come with a masking phenomenon that causes momentary reduction of sensor capabilities [131]. Hence, a small number of gaze switches to avoid missing sensory information is desired in real-world behavior. Additionally, pursuit of minimal muscular effort as in the case of steering inputs can be assumed to be relevant for the driver. These aspects are both incorporated in the presented reward function.

3.3.3 Secondary Task Interaction

Engagement in secondary tasks during the primary task of driving can cause distraction that ultimately leads to a crash as discussed in Cpt. 1. However, drivers want to and do actually engage into various secondary tasks while driving. Hence, the interaction with a potentially distracting secondary task must explicitly be considered in the POMDP for the purpose of computing appropriate glance policies. In this context, of course only those tasks are relevant that require to avert gaze from the road, i.e. require $x_t^{\mathsf{z}} = 0$ for successful interaction.

MDP Model Formally, this can be modeled by an MDP in interaction states x_t^{i} in a state space S^{i} and in interaction controls u_t^{i} in a control space U^{i} with dynamics $\mathcal{P}^{\mathsf{i}}$ that depend non-trivially on both the sensor state x_t^{z} and the sensor control u_t^{z}

$$\mathcal{P}^{\mathsf{i}}(x_{t+1}^{\mathsf{i}} | x_t^{\mathsf{z}}, u_t^{\mathsf{z}}; x_t^{\mathsf{i}}, u_t^{\mathsf{i}}). \tag{3.17}$$

In addition to that a reward function on the secondary task states and secondary task controls

$$r(x_t^{\mathsf{i}}, u_t^{\mathsf{i}}) = \boldsymbol{\theta}_6^{\top} \varphi(x_t^{\mathsf{i}}, u_t^{\mathsf{i}}) \tag{3.18}$$

is given.

Previous work on human dual-tasking in driving found significant influences of task characteristics on the decision to switch the tasks. For example, "natural" sub-task boundaries such as blocks in the representation of US telephone numbers [196, 35] or confirmatory button presses [128] favor gaze switches. Furthermore, task objectives specified by reward functions affect interleaving strategies [90]. Using the MDP in interaction states and control it is possible to explicitly model these aspects. This is demonstrated with two exemplar secondary task models.

Exemplar Secondary Task Consider the following secondary task in states $\mathbf{x}_t^{\mathsf{i}} = [n_t \, f_t \, m_t]^{\top}$, $n_t \in \{1, 2\}$, $f_t \in \{0, 1\}$, $m_t \in \{0, 1\}$ and controls $u_t^{\mathsf{i}} \in \{0, 1\}$: Random numbers 1 and 2 are generate and displayed on a screen in the vehicle interior, e.g. on the infotainment display, denoted by display state n_t. If the driver averts his or her gaze from the road $x_t^{\mathsf{z}} = 0$ he or she can read the displayed number (see Fig. 3.4). This is modeled by setting the memorization state $m_t = 1$. The number can be typed by pressing buttons u_t^{i}. The processes of reading and pressing is indicated by $f_t \in \{0 := \text{not finished yet}, 1 := \text{finished}\}$ with a finishing probability of p_f. The pressed button and the displayed number are compared. If both coincide a new number is uniformly sampled $n_t \sim \mathcal{U}$. In case of incorrect button press the old number n_t remains. The driver can also press the buttons while looking at the road $x_t^{\mathsf{z}} = 1$: If he or she has memorized the last visible number $m_t = 1$ he or she can type the memorized number. If the typed number was correct the next number is generated, but the driver is not aware of its value $m_t = 0$. That is, effectively any button press has a chance of 0.5 to be correct. The driver cannot check whether the button press was successful or not when his or her gaze is on the road. Hence, from the perspective of the driver this is equivalent to n_t being re-sampled after every button press.

Finally, a suitable reward function for this task is a reward for correct button presses and a penalty for wrong button presses

$$r(\mathbf{x}_t^{\mathsf{i}}, u_t^{\mathsf{i}}) = 2\mathbb{I}_{n_t = u_t^{\mathsf{i}}}(\mathbf{x}_t^{\mathsf{i}}, u_t^{\mathsf{i}}) - 1. \tag{3.19}$$

The final MDP of the secondary task interaction is depicted in Fig. 3.6.

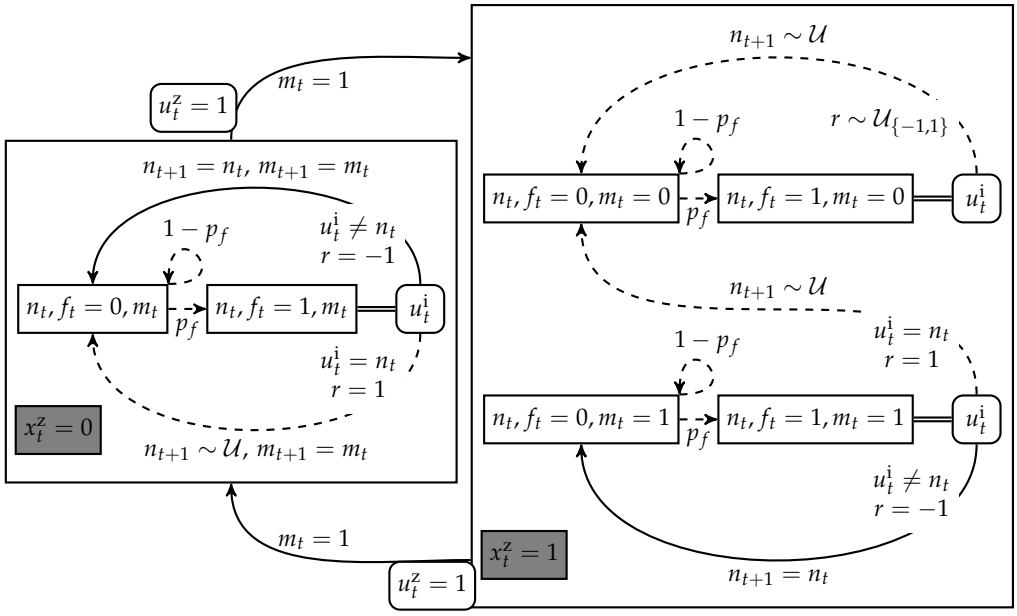

Figure 3.6: Illustration of MDP model of secondary task interaction. Rectangles denote the components of the interaction state x_t^i. These are the displayed number n_t, finish state f_t and memorization state m_t. Rectangles with rounded corners denote the interaction control u_t^i and sensor controls x_t^z. Arrows depict the dynamics. Here, solid arrows indicate deterministic transitions, dashed arrows denote stochastic transitions. Changes of states are explicitly specified, only if not obvious from the depicted successor state.

Already, in this simple task interesting interaction policies can emerge. First, the dynamics of f_t may induce natural task boundaries. This is the case if the probability of transiting to $f_t = 1$, p_f is low and therefore a significant amount of time steps is required until success. Second, if incorrect typing does not result in a high penalty, drivers might choose to keep the gaze on the road to improve visual sensing while following a uniformly random typing policy. This is the case, especially, if random typing is faster $p_f(m_t = 1) > p_f(m_t = 0)$ than reading and pressing the correct button.

Simple Secondary Task Model A very simple alternative to the previous model is given by the following MDP. The secondary task state is defined as the sensor state

$$x^i = x_t^z \in \{0 := \text{gaze on road}, 1 := \text{gaze off road}\}, \tag{3.20}$$

which is combined with a reward $r(x_t^i) = 1 - x_t^i$. This model effectively assumes constant utility per time step gazing off road, which is a crude yet reasonable approximation to many common secondary tasks. Note, in this model the driver obtains immediately reward from the secondary task interaction after averting his or her gaze from the road. This is in contrast to the more elaborate model, where some time steps elapse until the button is pressed. Consequently, glances off the road can be significantly shorter in the simple model and no natural task boundaries are present.

3.3.4 Overview of the POMDP Model

In the previous sections we introduced a kinematic model and a reward of the driving task, the model of driver's vision and gaze switching as well as a model and a reward of the visually demanding secondary task. Putting together the individual parts we can now formulate the POMDP model of the joint task of driving while engaging in a visually demanding secondary task. Formally this is an POMDP in states x_t, controls u_t, sensory measurements z_t, reward $r(x_t, u_t)$, dynamics $\mathcal{P}_t(x_{t+1}|x_t, u_t)$ and sensor model $p^{\mathrm{z}}(z_t|x_t)$ given as

$$r(x_t, u_t) = \theta_1(y_t)^2 + \theta_2(\dot{y}_t)^2 + \theta_3(\alpha_t)^2 + \theta_4(u_t^{\mathrm{p}})^2 + \theta_5 u_t^{\mathrm{z}} + \boldsymbol{\theta}_6^\top \varphi(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \tag{3.21}$$

$$\mathcal{P}_t(x_{t+1|}x_t, u_t) \text{ def. by } \begin{cases} \mathbf{x}_{t+1}^{\mathrm{p}} & = \mathbf{A}(v_t)\mathbf{x}_t^{\mathrm{p}} + \mathbf{B}(v_t)u_t^{\mathrm{p}} + \mathbf{a}(v_t, \kappa_t) + \boldsymbol{\epsilon}_t^{\mathrm{p}} \\ x_{t+1}^{\mathrm{z}} & = x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}} \\ & \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \end{cases} \tag{3.22}$$

$$p^{\mathrm{z}}(z_t|x_t) \text{ def. by } \left\{ \mathbf{z}_t \quad = \mathbf{H}(x_t^{\mathrm{z}})\mathbf{x}_t^{\mathrm{p}} + \boldsymbol{\epsilon}_t^{z}(x_t^{\mathrm{z}}). \right. \tag{3.23}$$

As in the cases of LQGs, we can formulate the equivalent belief MDP of the POMDP model using the Kalman filter (2.36). This results in the a-posterior mean $\boldsymbol{\mu}_t^{\mathrm{p}}$ and the a-posterior covariance $\boldsymbol{\Sigma}_t^{\mathrm{p}}$ of the primary task state $\mathbf{x}_t^{\mathrm{p}}$. Notably, in this case the belief is dependent on both the external influences v_t, κ_t and the sensor state x_t^{z}. The interplay of the different components of the belief MDP and the policy is depicted in Fig. 3.7.

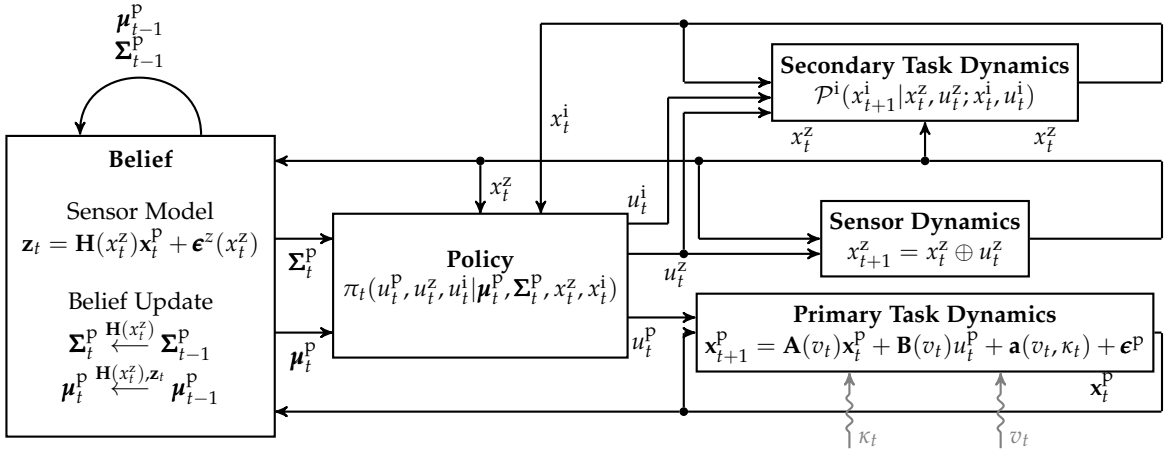

Figure 3.7: Illustration of the POMDP model of the joint task. The model is given in form of the equivalent belief MDP. It consists of the primary task dynamics, the sensor dynamics and the secondary task dynamics and is subject to the two external influences κ_t and v_t.

3.4 Appropriate Glance Behavior

In the previous section a POMDP of the joint task of driving and secondary task interaction was presented. This model features both reward parameters $\boldsymbol{\theta}$ and external parameters driving speed $\mathbf{v}_{0:T}$ and lane curvature $\boldsymbol{\kappa}_{0:T}$. In this context, the external parameters model the variety of possible driving situations encountered in lane keeping. Consequently, rational and optimal policies computed for a specific POMDP instance are adapted to the modeled driving situation. For example, the policy for gaze switching and the resulting durations of glances are adapted to the driving speed.

In the following we will use this important property of rational and optimal policies to define situationally appropriate glance behavior. As this normative definition takes into account an explicit model of the joint task of driving and secondary task interaction it provides a framework for sensitive assessment of driver attention.

We define the Eyes-Off Duration (EOD) d_t as the time steps passed since the driver's gaze was *on the road* for the last time

$$d_t := \min_{\{k:\, k\geq 0,\, x^z_{t-k}=1\}} (k). \tag{3.24}$$

Consider the joint task POMDP with reward function r, dynamics $\boldsymbol{\mathcal{P}}_{0:t}$ and sensor model p^z (Sec. 3.3.4) resulting from a specific realization of the driving speed $\mathbf{v}_{0:T}$ and lane curvature $\boldsymbol{\kappa}_{0:T}$. That is, a POMDP model of the specific current driving situation. Let $p(d_t^{\boldsymbol{\pi},\boldsymbol{\mathcal{P}},p^z})$ denote the distribution of EODs resulting from a policy $\boldsymbol{\pi}_{0:t}$ and the situation specific POMDP. Given these quantities, we can ask for the probability $p(d_t^{\boldsymbol{\pi},\boldsymbol{\mathcal{P}},p^z} < d_t)$ that the eyes-off duration the current EOD d_t of the driver exceeds the EOD $d_t^{\boldsymbol{\pi},\boldsymbol{\mathcal{P}},p^z}$ under a policy $\boldsymbol{\pi}_{0:t}$. This can also be formulated as "How likely is it that an agent applying the policy has already returned his gaze to the road?". In the following we introduce a mathematical definition of appropriate glance behavior based on the probability that an agent following an optimal policy $\pi^\star$ or rational policy $\tilde{\pi}$ has already returned his gaze to the road.

Definition of Appropriate Glance Behavior *Appropriate Glance Behavior* (AGB) is the glance behavior, where the probability of the driver's eyes-off duration d_t exceeding the eyes-off duration $d_t^{\boldsymbol{\pi}^\star,r,\boldsymbol{\mathcal{P}},p^z}$ of the specific optimal policy $\boldsymbol{\pi}^\star_{0:T}(r;\boldsymbol{\mathcal{P}})$, or the $d_t^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z}$ of the specific maximum causal entropy policy $\tilde{\boldsymbol{\pi}}_{0:T}(r;\boldsymbol{\mathcal{P}})$ of the joint task POMDP model instance $r, \boldsymbol{\mathcal{P}}_{0:t}, p^z$ of the driving situation, is below a threshold τ_{AGB}

$$p(d_t^{\boldsymbol{\pi}^\star,r,\boldsymbol{\mathcal{P}},p^z} < d_t) < \tau_{\text{AGB}} \text{ or } p(d_t^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z} < d_t) < \tau_{\text{AGB}}.$$

Among the possible deviations from both rational policies, we only consider excessive eyes-off durations which are related to decreased *primary task* performance. This is because the targeted warning system should not intervene in cases the driver's behavior is found to be more cautious, i.e. if he or she shows shorter EOD, than necessary. The probability threshold τ_{AGB} in our definition is a tuning parameter which needs to be set in accordance to the urgency of the employed type of warning in a distraction warning system. An exemplar choice of this parameter for a prototypical distraction warning system is presented in Sec. 6.4.4.

Recently, Kircher and Ahlström proposed the framework of *Minimum Required Attention* (MiRA) [106]. In this framework the driver is considered to be attentive when he or she sampled sufficient information to meet the demands of the driving task given the current situation. This approach shows many similarities to our mathematical definition. However, MiRA was presented purely conceptional and no algorithmic approaches to implement this framework in a warning systems were given. In contrast, this is possible with our precise formal definition. Our approach considers both the primary task and the secondary task that compete for the visual attention of the driver. Hence, under a reasonable choice of parameters $\boldsymbol{\theta}$ we obtain a policy for returning gaze back to road that trades off decreased performance in the primary task against increased utility from the secondary task. For the special case of the simple secondary task model, Sec. 3.5.2 shows that the optimal gaze switching policy implicitly

defines a threshold on loss of primary task performance. The problem of finding suitable parameters $\boldsymbol{\theta}$ will separately be addressed in Cpt. 4

3.5 Computation of Appropriate Glance Behavior

Using the previous definition of appropriate glance behavior enables situation specific driver attention assessment. To implement this definition, it is required to compute the specific optimal or rational policies wrt. the current driving situation. This will be addressed in the following section. Ultimately, the definition of appropriate glance behavior shall be used in a real-time warning system. Hence, we will especially focus on efficiently computing policies. For this purpose, we first analyze the model and classify its problem class. Next, the optimal value and state-control function are derived and solution approaches are considered (Sec. 3.5.2). Thereafter, computing policies in the maximum causal entropy policy model is addressed (Sec. 3.5.3). In contrast, to other work on POMDP models of glance behavior this work will focus on techniques that are exact. Inexact solution approaches are reviewed in Sec. 3.5.2 and the issues of these methods wrt. defining appropriate glance behavior are discussed.

3.5.1 Classification of Problem Class

As a first step towards computing rational policies, we classify the problem class of the POMDP model of secondary task interaction while driving which we summarized in Sec. 3.3.4. This enables comparison with other POMDP models and allows to transfer solution techniques. Optimal policies in the joint task POMDP are mathematically defined as the solution of the optimization problem

$$\max_{\boldsymbol{\pi}_{0:T}} \quad \mathbb{E}\Big[\sum_{t=0}^{T} -[\mathbf{x}_t^{\mathrm{p}}]^\top \mathbf{C}_x[\mathbf{x}_t^{\mathrm{p}}] - [u_t^{\mathrm{p}}]^\top \mathbf{C}_u[u_t^{\mathrm{p}}] + r(u_t^{\mathrm{z}}) + r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p^{\mathrm{z}}, p_0\Big] \tag{3.25}$$

$$\mathcal{P}_t(x_{t+1|}x_t, u_t) \text{ def. by } \begin{cases} \mathbf{x}_t^{\mathrm{p}} & = \mathbf{A}_t\mathbf{x}_t^{\mathrm{p}} + \mathbf{B}_t u_t^{\mathrm{p}} + \mathbf{a}_t + \boldsymbol{\epsilon}_t^{\mathrm{p}} \\ x_{t+1}^{\mathrm{z}} & = x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}} \\ & \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \end{cases} \tag{3.26}$$

$$p^{\mathrm{z}}(z_t|x_t) \text{ def. by } \Big\{ \mathbf{z}_t \quad = \mathbf{H}(x_t^{\mathrm{z}})\mathbf{x}_t^{\mathrm{p}} + \boldsymbol{\epsilon}_t^{z}(x_t^{\mathrm{z}}). \tag{3.27}$$

To the best of our knowledge there is no specific POMDP subclass that contains the general joint task model. However, in the case of the simple secondary task model considered in Sec. 3.3.3,

$$\max_{\boldsymbol{\pi}_{0:T}} \quad \mathbb{E}\Big[\sum_{t=0}^{T} -[\mathbf{x}_t^{\mathrm{p}}]^\top \mathbf{C}_x[\mathbf{x}_t^{\mathrm{p}}] - [u_t^{\mathrm{p}}]^\top \mathbf{C}_u[u_t^{\mathrm{p}}] + r(u_t^{\mathrm{z}}) + r(x_t^{\mathrm{z}}) \Big| \boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p^{\mathrm{z}}, p_0\Big] \tag{3.28}$$

$$\mathcal{P}_t(x_{t+1|}x_t, u_t) \text{ def. by } \begin{cases} \mathbf{x}_t^{\mathrm{p}} & = \mathbf{A}_t\mathbf{x}_t^{\mathrm{p}} + \mathbf{B}_t u_t^{\mathrm{p}} + \mathbf{a}_t + \boldsymbol{\epsilon}_t^{\mathrm{p}} \\ x_{t+1}^{\mathrm{z}} & = x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}} \end{cases} \tag{3.29}$$

$$p^{\mathrm{z}}(z_t|x_t) \text{ def. by } \Big\{ \mathbf{z}_t \quad = \mathbf{H}(x_t^{\mathrm{z}})\mathbf{x}_t^{\mathrm{p}} + \boldsymbol{\epsilon}_t^{z}(x_t^{\mathrm{z}}) \tag{3.30}$$

the POMDP is an instance of so-called *optimal measurement scheduling/optimal sensor scheduling* problems [152]. In these problems the objective is to find an optimal policy for selecting sensors as to maximize state estimation performance. Like this thesis, the seminal work of Meier and colleagues [152] considered sensor scheduling to maximize the resulting control performance in a linear quadratic Gaussian problem, which we will denote as Sensor scheduling Linear Quadratic Gaussian (SLQG) problems. More recently also sensor scheduling in hidden Markov models has been considered [115]. Other authors investigated different objectives for estimation performance or additional constraints in linear Gaussian problems [158, 92]. Notably, [19] already suggested sensor scheduling in LQGs as a normative model for monitoring of several displays in manual control.

Unfortunately, solving SLQGs is significantly more challenging than ordinary LQGs. In the following we will now consider obtaining optimal and rational policies in the class of the joint task POMDP. This will be done by means of the corresponding (soft) Bellman-equations.

3.5.2 Optimal Policies

In the POMDP model considered in this work the optimal policy can be characterized via its optimal value function $V_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{\Sigma}_t^\mathrm{p}, x_t^\mathrm{z}, x_t^\mathrm{i})$ and optimal state-control function $Q_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{\Sigma}_t^\mathrm{p}, x_t^\mathrm{z}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i})$. In this context, let $\boldsymbol{\Sigma}_t^\mathrm{p}(\boldsymbol{x}^\mathbf{z}_{0:t})$ be the a-posterior covariance of the primary task state which is fully determined from the history of sensor states $\boldsymbol{x}^\mathbf{z}_{0:t} = [x_0^\mathrm{z}\, x_1^\mathrm{z}\, \dots\, x_t^\mathrm{z}]$ and the initial covariance $\boldsymbol{\Sigma}_0^\mathrm{p}$. Now we can reformulate the value function as $V_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i})$ and the state-control function as $Q_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i})$.

Bellman Equations The Bellman equations are evaluated by first separating the primary task parts from the remaining variables. This allows to employ the techniques introduced earlier in the context of LQR (Sec. 2.1.2) and LQG (Sec. 2.1.4). Finally, $V_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i})$ and $Q_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i})$ are given in the form of

$$Q_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i}) = [\boldsymbol{\mu}_t^\mathrm{p}; u_t^\mathrm{p}]^\top \mathbf{M}_t^{Q^\star}[\boldsymbol{\mu}_t^\mathrm{p}; u_t^\mathrm{p}] + \mathbf{m}_t^{Q^\star}[\boldsymbol{\mu}_t^\mathrm{p}; u_t^\mathrm{p}] + m_t^{Q^\star,1}(\boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i}) \tag{3.31}$$

$$V_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}) = [\boldsymbol{\mu}_t^\mathrm{p}]^\top \mathbf{M}_t^{V^\star}[\boldsymbol{\mu}_t^\mathrm{p}] + \mathbf{m}_t^{V^\star}[\boldsymbol{\mu}_t^\mathrm{p}] + m_t^{V^\star,1}(\boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}). \tag{3.32}$$

Specifically, for the terms present in the state-control function $Q_t^\star(\boldsymbol{\mu}_t^\mathrm{p}, \boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i})$ it holds

$$\mathbf{M}_t^{Q^\star} = \begin{cases} [\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star}[\mathbf{A}_t\, \mathbf{B}_t] - \mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{if } t < T \\ -\mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{else} \end{cases} \tag{3.33}$$

$$\mathbf{m}_t^{Q^\star} = \begin{cases} 2[\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star}\mathbf{a}_t + [\mathbf{A}_t\, \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{V^\star} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \tag{3.34}$$

$$m_t^{Q^\star,1}(\boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}, u_t^\mathrm{p}, u_t^\mathrm{z}, u_t^\mathrm{i}) = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{V^\star}\mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{V^\star} \\ \quad + \mathrm{tr}(-\mathbf{C}_x\boldsymbol{\Sigma}_t^\mathrm{p}(\boldsymbol{x}^\mathbf{z}_{0:t})) \\ \quad + \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^\star}\left(\mathbf{A}_t\boldsymbol{\Sigma}_t^\mathrm{p}(\boldsymbol{x}^\mathbf{z}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{e^x} - \boldsymbol{\Sigma}_{t+1}^\mathrm{p}([\boldsymbol{x}^\mathbf{z}_{0:t}\, x_t^\mathrm{z} \oplus u_t^\mathrm{z}])\right)\right) \\ \quad + r(u_t^\mathrm{z}) + r(x_t^\mathrm{i}, u_t^\mathrm{i}) \\ \quad + \mathbb{E}\left[m_{t+1}^{V^\star,1}([\boldsymbol{x}^\mathbf{z}_{0:t}\, x_t^\mathrm{z} \oplus u_t^\mathrm{z}], x_{t+1}^\mathrm{i}) \middle| \mathcal{P}^\mathrm{i}(x_{t+1}^\mathrm{i} | x_t^\mathrm{z}, u_t^\mathrm{z}; x_t^\mathrm{i}, u_t^\mathrm{i})\right] & \text{if } t < T \\ -\mathrm{tr}(\mathbf{C}_x\boldsymbol{\Sigma}_t^\mathrm{p}(\boldsymbol{x}^\mathbf{z}_{0:T})) + r(u_T^\mathrm{z}) + r(x_T^\mathrm{i}, u_T^\mathrm{i}) & \text{else} \end{cases} \tag{3.35}$$

Furthermore, the quantities $\mathbf{M}_t^{V^\star}\mathbf{m}_t^{V^\star}, m_t^{V^\star,1}(\boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i})$ (see Sec. 2.1.2 and Sec. 2.1.4) are given by

$$\mathbf{M}_t^{V^\star} = \mathbf{M}_{t\,x,x}^{Q^\star} - \mathbf{M}_{t\,x,u}^{Q^\star}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{M}_{t\,u,x}^{Q^\star} \tag{3.36}$$

$$\mathbf{m}_t^{V^\star} = \mathbf{m}_{t\,x}^{Q^\star} - \mathbf{M}_{t\,x,u}^{Q^\star}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{m}_{t\,u}^{Q^\star} \tag{3.37}$$

$$\begin{aligned} m_t^{V^\star,1}(\boldsymbol{x}^\mathbf{z}_{0:t}, x_t^\mathrm{i}) = {} & -\frac{1}{4}[\mathbf{m}_{t\,u}^{Q^\star}]^\top[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{m}_{t\,u}^{Q^\star} + \mathrm{tr}(-\mathbf{C}_x\boldsymbol{\Sigma}_t^\mathrm{p}(\boldsymbol{x}^\mathbf{z}_{0:t})) + \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^\star}\left(\mathbf{A}_t\boldsymbol{\Sigma}_t^\mathrm{p}(\boldsymbol{x}^\mathbf{z}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{e^x}\right)\right) \\ & + \max_{u_t^\mathrm{z}, u_t^\mathrm{i}}\Big(r(u_t^\mathrm{z}) - \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^\star}\boldsymbol{\Sigma}_{t+1}^\mathrm{p}([\boldsymbol{x}^\mathbf{z}_{0:t}\, x_t^\mathrm{z} \oplus u_t^\mathrm{z}])\right) \\ & + r(x_t^\mathrm{i}, u_t^\mathrm{i}) + \mathbb{E}\left[m_{t+1}^{V^\star,1}([\boldsymbol{x}^\mathbf{z}_{0:t}\, x_t^\mathrm{z} \oplus u_t^\mathrm{z}], x_{t+1}^\mathrm{i}) \middle| \mathcal{P}^\mathrm{i}(x_{t+1}^\mathrm{i} | x_t^\mathrm{z}, u_t^\mathrm{z}; x_t^\mathrm{i}, u_t^\mathrm{i})\right]\Big). \end{aligned} \tag{3.38}$$

Characterizing the Bellman Equations Analyzing the structure of value function and state-control function, it turns out both functions factorize. The functions comprise of summands that either depend on the primary task variables $\boldsymbol{\mu}^\mathrm{p}, u_t^\mathrm{p}$ or on variables the sensor dynamics $\boldsymbol{x}^\mathbf{z}_{0:t}, u_t^\mathrm{z}$ and the secondary task $x_t^\mathrm{i}, u_t^\mathrm{i}$. As a consequence, like in LQG, the optimal policy for the primary task control u_t^p is a linear feedback controller dependent only on the a-posterior mean $\boldsymbol{\mu}_t^\mathrm{p}$ of the primary task state $\mathbf{x}_t^\mathrm{p}$:

$$\pi^\star(u_t^\mathrm{p} | \boldsymbol{\mu}_t^\mathrm{p}) = \mathcal{I}(u_t^\mathrm{p} | [u_t^\mathrm{p}]^\star(\boldsymbol{\mu}_t^\mathrm{p})), \quad [u_t^\mathrm{p}]^\star(\boldsymbol{\mu}_t^\mathrm{p}) = \mathbf{F}_t^\star\boldsymbol{\mu}_t^\mathrm{p} + \mathbf{f}_t^\star \tag{3.39}$$

$$\mathbf{F}_t^\star := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{M}_{t\,u,x}^{Q^\star}, \quad \mathbf{f}_t^\star := -\frac{1}{2}[\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1}\mathbf{m}_{t\,u}^{Q^\star}. \tag{3.40}$$

In contrast, the policy for choice of the sensor and secondary task control depends on all aspects of the joint task. Specifically, the policy takes into account, the cost of switching $r(u_t^z)$, the reward of the secondary task $r(x_t^i, u_t^i)$ and the expected primary task control performance under the primary state uncertainty resulting from all possible future sequences of sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{t:T}$. The latter is a result of the both terms $\mathrm{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}))$ and $\mathrm{tr}\left(\mathbf{M}_{t+1}^{V^\star}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z])\right)\right)$ involved in the value function.

The part of the value function that depends on the sensor state and secondary task state (3.38) and the corresponding part of the state-control function (3.35) cannot further be factorized. As the involved states $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^i$ and controls u_t^z, u_t^i are discrete, those functions can be represented by a multi-dimensional array. Let uss consider the size of this array. The state $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ is a sequence of t sensor states and therefore an element of the t-fold product of the sensor state space $\bigotimes_{i=1}^t \mathsf{S}^z$. The sensor control u_t^z is an element of U^z, x_t^i is an element of S^i and u_t^i is an element of U^i. As a result, $m_t^{Q^\star,1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^i, u_t^p, u_t^z, u_t^i)$ corresponds to an array with

$$\left|\left(\bigotimes_{i=1}^{t} \mathsf{S}^z\right) \times \mathsf{U}^z \times \mathsf{S}^i \times \mathsf{U}^i\right| = 2^t\, 2\, |\mathsf{S}^i|\, |\mathsf{U}^i| \tag{3.41}$$

elements. In the array size the factor $\bigotimes_{i=1}^t \mathsf{S}^z$ is problematic as it grows exponentially with the time step t or rather the length of the planning horizon T. The growth of this space is illustrated in Fig. 3.8.

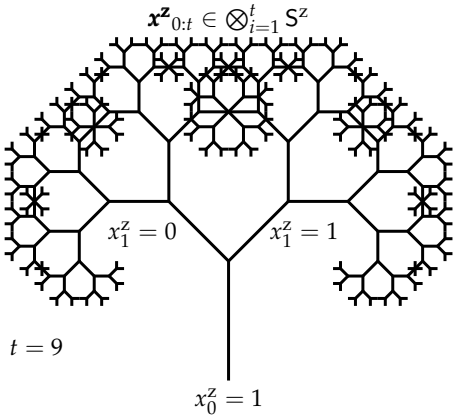

Figure 3.8: The space $\bigotimes_{i=1}^t \mathsf{S}^z$ of sensor state sequences $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ for $t = 9$ illustrated as an Pythagorean binary tree.

Already in case of a modest planning horizon $T = 25$ that corresponds to one second in the model frequency of 25 Hz, the size of the sensor state space is 2^{25} and above 10 millions. This would impose large memory and computational requirements on a computer system, which would render it impossible to use this approach online. Consequently, further approaches in addition to Bellman equations are required to efficiently compute optimal policies.

Sensor Model Restriction

The computational burden of the Bellman equation can be reduced by restricting the POMDP in such a way that only a small subset of $\bigotimes_{i=1}^t \mathsf{S}^z$ needs to be considered. If for example the POMDP model can be restricted such that it suffices to consider the EOD d_t then the maximum array size of $m_t^{Q^\star,1}$ would be $T\, 2\, |\mathsf{S}^i|\, |\mathsf{U}^i|$. Clearly, this would tremendously facilitate computation while still appropriate glance behavior could be specified by means of the definition in Sec. 3.4. As we will see, this desired property can be obtained by a sensor model restriction.

Previously in this thesis, the driver's sensing has been modeled by means of $\mathbf{H}(x_t^z) = \mathrm{diag}(1, 0, 1, 1)$, $\boldsymbol{\Sigma}^{\epsilon^z}(x_t^z) = \mathrm{diag}((\sigma_y)^2(x_t^z), 0, (\sigma_\phi)^2(x_t^z), 0)$. Here, typically the sensorial noise when gazing at the road $(\sigma_y)^2(0), (\sigma_\phi)^2(0)$ is significantly lower than the noise when not looking at the road $(\sigma_y)^2(1), (\sigma_\phi)^2(1)$. Consider a sequence of sensor states where the driver's gaze is first off the road, then he returns

his gaze to the road followed by finally averting gaze. The reason why generally the full sequence of sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ needs to be considered in the state-control function, is that uncertainty monotonically decreases after the driver has returned his or her gaze back to road. This relationship is depicted in Fig. 3.9. The parameters of the sensor noise $(\sigma_y)^2(0), (\sigma_\phi)^2(0)$ for gaze on the road were chosen that the steady-state belief of the lane position has a 0.96 confidence interval of reasonable 0.3 m. The sensor model parameters for the sensor state gaze off the road were set in a way that corresponds to the driver not receiving any new information of the vehicles position and orientation in lane.

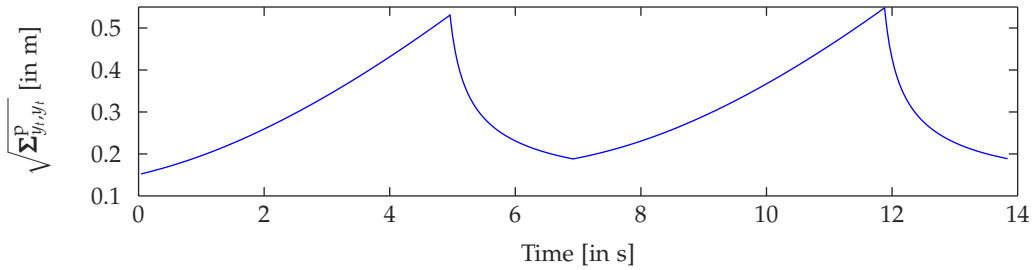

Figure 3.9: Resulting uncertainty in the belief for a sequence $x^{\mathrm{z}}_{[0,5]\,\mathrm{s}} = 1$, $x^{\mathrm{z}}_{[5,7]\,\mathrm{s}} = 0$, $x^{\mathrm{z}}_{[7,11.9]\,\mathrm{s}} = 1$, $x^{\mathrm{z}}_{[11.9,14]\,\mathrm{s}} = 0$ at 80 m/h . The blue line indicates the standard deviation of the belief of the lane position y_t, i.e. $\sqrt{\Sigma^{\mathrm{p}}_{y_t,y_t}}$. Noise in primary task dynamics was fit to experimental data, sensor noise was set to $(\sigma_y)^2(0) = 0.64, (\sigma_\phi)^2(0) = 0.01, (\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty$.

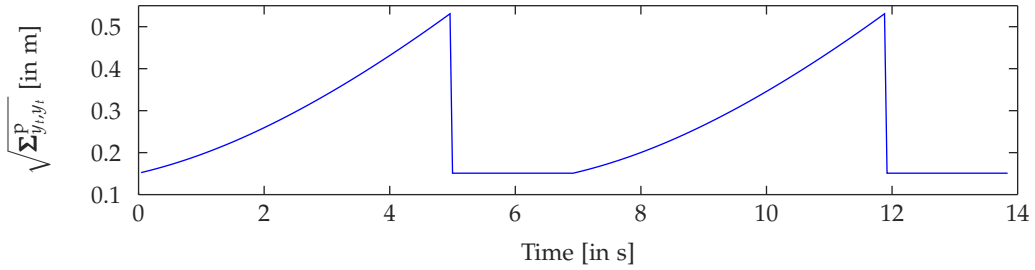

Figure 3.10: Resulting uncertainty in the belief using the immediate full information heuristic. Gaze switches and parameters as in Fig. 3.9.

As can be seen, for these sensor model parameters a gaze on the road for 2 s (5 s $-$ 7 s) does not suffice to decrease uncertainty to the previous level at 0 s. Consequently, the uncertainty for the same amount of time glancing off the road reaches a higher level at 12 s than previously obtained at 5 s. As a result, the time spent gazing at the road, the so-called *viewing time*, positively correlates with the previous time of gaze aversion under optimal and rational gaze switching policies. In the evaluation Sec. 3.6.1, we will revisit this property and investigate if this is also present in the glance behavior in real driving.

Viewing time is independent of the duration of the previous off-road glance, if all available information is immediately obtained when the gaze is returned to the road. This is the case, if the driver perfectly senses the vehicle's position in lane and its orientation $\boldsymbol{\Sigma}^{\epsilon^z}(x^z_t) = \mathrm{diag}((\sigma_y)^2(0), 0, (\sigma_\phi)^2(0), 0)$, $(\sigma_\phi)^2(0), (\sigma_y)^2(0) = 0$. Here, the covariance of the belief $\boldsymbol{\Sigma}^{\mathrm{p}}_t$ shrinks to zero making any previous sensor states obsolete. Interestingly, the occurrence of the belief in the Bellman equations 3.35 allows for an additional heuristic. We can simply set $\boldsymbol{\Sigma}^{\mathrm{p}}_t(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})$ to an arbitrary constant value if $x^z_t = 0$ and the Bellman equations are still well defined. Fig. 3.10 shows the dynamics of the belief when it is assumed that the belief convariance immediately jumps the steady-state covariance $\hat{\boldsymbol{\Sigma}}^{\mathrm{p}}_t$ once the driver's gaze is returned to the road. In this context, we use the same gaze switch series, driving speed and noise parameters of Fig. 3.9. Important to mention, If applying this heuristic it is not guaranteed that there

actually exists any sensor model that results in the same jump of the belief covariance. Consequently, this belief MDP might not have a POMDP equivalent.

This said, we can formulate Algo. 4 to obtain the optimal policy for the joint task of lane keeping and secondary task interaction under the sensor model restriction. This algorithm takes as input the POMDP model as well as the initial state x_0^z with a corresponding covariance of the primary task belief $\mathbf{\Sigma}_0^{\mathrm{p}}$ and the alternate steady state covariance for gaze on road $\hat{\mathbf{\Sigma}}_0^{\mathrm{p}}$.

Algorithm 4 Optimal Solution of Joint Task Model under Sensor Model Restriction SRopt

function SROPT($\mathbf{C}_x, \mathbf{C}_u, r(u_t^z), r(x_t^i, u_t^i), (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B})_{t=0:T}, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \mathbf{\Sigma}^{\epsilon^z}(x_t^z), \mathcal{P}^i, x_0^z, \mathbf{\Sigma}_0^p, \hat{\mathbf{\Sigma}}_0^p$)
  $\left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_t^\star\right)_{t=0:T} \leftarrow$ LQROPT$(\mathbf{C}_x, \mathbf{C}_u, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T})$ ▷ Algo. 1
  for $t = 0 : T$ **do** ▷ Pre-compute belief for gaze on road
    $\hat{\mathbf{\Sigma}}_{t+1}^p \leftarrow$ KALMANUPDATE$(\hat{\mathbf{\Sigma}}_t^p, \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(0), \mathbf{\Sigma}^{\epsilon^z}(0))$ ▷ Algo. 2
  end for
  for $t = 0 : T$ **do** ▷ Forward pass for belief for gaze off road
    if $t = 0$ **then**
      $\mathbf{\Sigma}^1 \leftarrow \mathbf{\Sigma}_0^p$
      $d_0 \leftarrow x_0^z$
    else
      $\mathbf{\Sigma}^1 \leftarrow \hat{\mathbf{\Sigma}}_t^p$
      $d_0 \leftarrow 0$
    end if
    for $d_t = d_0 : T - t - 1$ **do** ▷ Roll-out off road belief until max. duration
      $t' \leftarrow t + d_t - d_0 - 1$
      if $t' > 0$ **then**
        $\mathbf{\Sigma}^1 \leftarrow$ KALMANUPDATE$(\mathbf{\Sigma}^1, \mathbf{A}_{t'}, \mathbf{a}_{t'}, \mathbf{B}_{t'}, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(1), \mathbf{\Sigma}^{\epsilon^z}(1))$
      end if
      $\mathbf{\Sigma}^2 \leftarrow$ KALMANUPDATE$(\mathbf{\Sigma}^1, \mathbf{A}_{t'+1}, \mathbf{a}_{t'+1}, \mathbf{B}_{t'+1})$
      $\mathbf{\Sigma}^3 \leftarrow$ KALMANUPDATE$(\mathbf{\Sigma}^1, \mathbf{A}_{t'+1}, \mathbf{a}_{t'+1}, \mathbf{B}_{t'+1}, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(1), \mathbf{\Sigma}^{\epsilon^z}(1))$
      if $d = 0$ **then**
        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{Q^\star,1}(d_t, x_t^i, 0, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \hat{\mathbf{\Sigma}}_{t'+1}^p) + \operatorname{tr}(\mathbf{M}_{t'+2}^{V^\star}(\mathbf{\Sigma}^2 - \hat{\mathbf{\Sigma}}_{t'+2}^p)) + r(u_t^z = 0)$
        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{Q^\star,1}(d_t, x_t^i, 1, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \hat{\mathbf{\Sigma}}_{t'+1}^p) + \operatorname{tr}(\mathbf{M}_{t'+2}^{V^\star}(\mathbf{\Sigma}^2 - \mathbf{\Sigma}^3)) + r(u_t^z = 1)$
      else
        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{Q^\star,1}(d_t, x_t^i, 0, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \mathbf{\Sigma}^1) + \operatorname{tr}(\mathbf{M}_{t'+2}^{V^\star}(\mathbf{\Sigma}^2 - \mathbf{\Sigma}^3)) + r(u_t^z = 0)$
        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{Q^\star,1}(d_t, x_t^i, 1, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \mathbf{\Sigma}^1) + \operatorname{tr}(\mathbf{M}_{t'+2}^{V^\star}(\mathbf{\Sigma}^2 - \hat{\mathbf{\Sigma}}_{t'+2}^p)) + r(u_t^z = 1)$
      end if
    end for
  end for
  for $t = T : 0$ **do** ▷ Use Bellman equations to obtain remaining part
    $\forall_{d_t \in \{0:T\}, x_t^i, u_t^z, u_t^i}\ m_t^{Q^\star,1}(d_t, x_t^i, x_t^z, u_t^i) \leftarrow^+ r(x_t^i, u_t^i)$
    for $d_t = 0 : \min(t+1, T)$ **do**
      if $d_t = 0$ **then**
        $d'(u_t^z = 0) \leftarrow 0,\ d'(u_t^z = 1) \leftarrow 1$
        $\forall_{x_t^i, u_t^z, u_t^i}\ m_t^{Q^\star,1}(d_t, x_t^i, u_t^z, u_t^i) \leftarrow^+ \mathbb{E}[m_{t+1}^{V^\star,1}(d'(u_t^z), x_{t+1}^i) | \mathcal{P}_t^i(x_{t+1}^i | 0, u_t^z; x_t^i, u_t^i)]$
      else
        $d'(u_t^z = 0) \leftarrow d_t + 1,\ d'(u_t^z = 1) \leftarrow 0$
        $\forall_{x_t^i, u_t^z, u_t^i}\ m_t^{Q^\star,1}(d_t, x_t^i, u_t^z, u_t^i) \leftarrow^+ \mathbb{E}[m_{t+1}^{V^\star,1}(d'(u_t^z), x_{t+1}^i) | \mathcal{P}_t^i(x_{t+1}^i | 1, u_t^z; x_t^i, u_t^i)]$
      end if
      $\forall_{x_t^i} m_t^{V^\star,1}(d_t, x_t^i) \leftarrow \max_{u_t^z, u_t^i}(m_t^{Q^\star,1}(d_t, x_t^i, u_t^z, u_t^i))$
      $\forall_{x_t^i} \pi_t^\star(d_t, x_t^i) \leftarrow \arg\max_{u_t^z, u_t^i}(m_t^{Q^\star,1}(d_t, x_t^i, u_t^z, u_t^i))$
    end for
  end for
  return $\left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_T^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_T^\star, \pi_t^\star(d_t, x_t^i)\right)_{t=0:T}$
end function

Secondary Task Model Restriction

Let us return to the joint task model obtained from the simple secondary task model (3.28)-(3.30). As noted before this is an instance of sensor scheduling in linear quadratic Gaussian problems. Although, this POMDP class has already been investigated in the early 60s [152, 19], it received little attention for several decades due to the in-feasibility of solution by means of the Bellman equations. Recently, SLQGs returned into the focus of control theoretic research because tractable solution approaches were discovered. Considering the Bellman equations (3.35)-(3.35) wrt. this special case, simplifications can be made. Specifically, it holds:

$$m_t^{Q^\star,1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, u_t^z) = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{V^\star} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{V^\star} \\ \quad + \operatorname{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \\ \quad + \operatorname{tr}\left(\mathbf{M}_{t+1}^{V^\star}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z])\right)\right) \\ \quad + r(u_t^z) + r(x_t^z) + m_{t+1}^{V^\star,1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z]) & \text{if } t < T \\ 0 & \text{else} \end{cases} \tag{3.42}$$

$$\begin{aligned} m_t^{V^\star,1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) = &-\frac{1}{4}[\mathbf{m}_{t\,u}^{Q^\star}]^\top [\mathbf{M}_{t\,u,u}^{Q^\star}]^{-1} \mathbf{m}_{t\,u}^{Q^\star} + \operatorname{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \operatorname{tr}\left(\mathbf{M}_{t+1}^{V^\star}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x}\right)\right) \\ &+ \max_{u_t^z}\left(r(u_t^z) - \operatorname{tr}\left(\mathbf{M}_{t+1}^{V^\star}\boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z])\right) + r(x_t^z) + m_{t+1}^{V^\star,1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z])\right). \end{aligned} \tag{3.43}$$

In contrast to the general joint task POMDP, here no stochastic element is present in the parts of the Bellman equation dependent on the sensor state x_t^z. Consequently, the optimal policy wrt. u_t^z depends only on the initial sensor state x_t^z as well as its associated covariance $\boldsymbol{\Sigma}_0^{\mathrm{p}}$. That is, the policy is given by an optimal sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star$ that can be obtained by solving the deterministic optimization problem

$$\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star = \arg\max_{\boldsymbol{x}^{\mathbf{z}}{}_{0:T}(\boldsymbol{u}^{\mathbf{z}}{}_{0:T})} \sum_{t=0}^{T} \operatorname{tr}\left(\mathbf{M}_t^{V^\star}\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\right) + r(u_t^z) + r(x_t^z) \tag{3.44}$$

as shown by reordering of terms in [152].

Tolerance on Loss of Primary Task Performance Before we consider solution approaches, we wish to illustrate the link between the minimum required attention (MiRA) framework [106] and optimal policies. Specifically, we will show that our definition of appropriate glance behavior allows to formally implement MiRA. That is, the decrease in lane keeping performance of the computed glance behavior compared to fully attentive driving is uniformly bounded over all driving situations. For this purpose consider the sensor state sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^1$ where the driver continuously gazes on the road, i.e $x_t^z{}^1 = 1$. Let us assume that the *optimal* primary task policy $u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}^{\mathrm{p}})$ is applied. If we compare the return of the sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^1$ to the optimal sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star$ for the rewards $r(x_t^z) = \theta_5(x_t^z - 1)$, $\theta_5 \geq 0$ and $r(u_t^z) = \theta_6(u_t^z)$, $\theta_6 \leq 0$ it holds

$$\mathbb{E}[\sum_{t=0}^{T} r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) | u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}_t^{\mathrm{p}}), \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^1, \boldsymbol{\mathcal{P}}] = \sum_{t=0}^{T} \operatorname{tr}\left(\mathbf{M}_t^{V^\star}\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^1)\right) + r(u_t^z{}^1) + r(x_t^z{}^1) \tag{3.45}$$

$$\underbrace{\leq}_{\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star \text{ is optimal}} \sum_{t=0}^{T} \operatorname{tr}\left(\mathbf{M}_t^{V^\star}\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^\star)\right) + r(u_t^{z\,\star}) + r(x_t^{z\,\star}) \tag{3.46}$$

$$= \mathbb{E}[\sum_{t=0}^{T} r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) + \theta_5(x_t^{\mathrm{z}} - 1) + \theta_6(u_t^{\mathrm{z}}) | u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}_t^{\mathrm{p}}), \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}, \boldsymbol{\mathcal{P}}] \quad (3.47)$$

$$\underbrace{\leq}_{\theta_6 \leq 0} \mathbb{E}[\sum_{t=0}^{T} r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) + \theta_5(x_t^{\mathrm{z}} - 1) | u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}_t^{\mathrm{p}}), \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}, \boldsymbol{\mathcal{P}}] \quad (3.48)$$

$$\leq \mathbb{E}[\sum_{t=0}^{T} r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) | u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}_t^{\mathrm{p}}), \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}, \boldsymbol{\mathcal{P}}] + T\,\theta_5. \quad (3.49)$$

This results in the following bound on expected loss of primary task performance

$$\mathbb{E}[\sum_{t=0}^{T} r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) | u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}_t^{\mathrm{p}}), \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{1}, \boldsymbol{\mathcal{P}}] - \mathbb{E}[\sum_{t=0}^{T} r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) | u_t^{\mathrm{p}\,\star}(\boldsymbol{\mu}_t^{\mathrm{p}}), \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}, \boldsymbol{\mathcal{P}}] \leq T\,\theta_5. \quad (3.50)$$

Consequently, the glance behavior defined by the optimal sequence of sensor states ensures sufficient attention that the driver's expected lane keeping performance does not fall below the expected lane keeping performance of a fully attentive driver minus the tolerance $T\,\theta_5$. Note, that the previous analysis of course includes the variants of the joint task POMDP were both the sensor model is restricted and the simple reward is used. This shows that the proposed mathematical framework of appropriate glance behavior allows to precisely implement the concept proposed in the MiRA framework.

Properties of Kalman Belief Update A direct way to solve the optimization problem (3.44) is to use brute-force search to find $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}$. Although applied in early works [152, 19, 20, 109] this approach is too time-consuming to be applicable in a real-time distraction warning system.

A more efficient technique was developed in [9, 246]. For this purpose, consider the belief update according to the Kalman filter given a fixed sensor model (2.36). Intuitively, an update of a belief $b^1(\mathbf{x}_t^{\mathrm{p}})$ "more uncertain" than a second belief $b^2(\mathbf{x}_t^{\mathrm{p}})$ cannot a result in a belief $b^1(\mathbf{x}_{t+1}^{\mathrm{p}})$ "more certain" than the belief $b^2(\mathbf{x}_{t+1}^{\mathrm{p}})$ resulting from updating $b^2(\mathbf{x}_t^{\mathrm{p}})$. This is because the same amount of new information is added. Let $\mathbf{X} \succeq \mathbf{Y}$ denote that the difference $\mathbf{X} - \mathbf{Y}$ of the matrices $\mathbf{X}, \mathbf{Y}$ is positive definite. In [246] it is shown that the intuitive observation holds true for the covariances $\boldsymbol{\Sigma}_t^1, \boldsymbol{\Sigma}_t^2$ of Gaussian beliefs in the mathematical formalization of

$$\forall \boldsymbol{\Sigma}_t^1 \succeq 0,\, \boldsymbol{\Sigma}_t^2 \succeq 0,\, \mathbf{C}_k \succeq 0,\, \lambda \in [0,1],\, k \in \mathbb{N}:$$

$$\boldsymbol{\Sigma}_t^1 \succeq \boldsymbol{\Sigma}_t^2 \Rightarrow \mathbf{KU}^k(\boldsymbol{\Sigma}_t^1) \succeq \mathbf{KU}^k(\boldsymbol{\Sigma}_t^2) \quad (3.51)$$

$$\lambda \mathbf{KU}^k(\boldsymbol{\Sigma}_t^1) + (1-\lambda)\mathbf{KU}^k(\boldsymbol{\Sigma}_t^1) \preceq \mathbf{KU}^k(\lambda\boldsymbol{\Sigma}_t^1 + (1-\lambda)\boldsymbol{\Sigma}_t^1), \quad (3.52)$$

where $\mathbf{KU}^k$ denotes an arbitrary fix series of k-fold application of the Kalman belief updates. Here, covariance $\boldsymbol{\Sigma}_t^1$ is considered "more uncertain" than covariance $\boldsymbol{\Sigma}_t^2$ if it can be written as the sum $\boldsymbol{\Sigma}_t^1 = \boldsymbol{\Sigma}_t^2 + \boldsymbol{\Sigma}_t^3$ of $\boldsymbol{\Sigma}_t^2$ and a third covariance matrix $\boldsymbol{\Sigma}_t^3$. That is, if the first Gaussian belief can be obtained by adding independent Gaussian noise to the second belief.

As a generalization of [246], that has not yet been considered in the literature, these properties of the Kalman update can also be exploited to efficiently optimize SLQGs (3.44). To do so, first note that for any negative definite $\mathbf{M} \prec 0$ and any matrix $\mathbf{X}$ it holds

$$\exists\, \mathbf{L}_M \colon \mathbf{M} = -\mathbf{L}_M \mathbf{L}_M^\top \text{ and } \operatorname{tr}(\mathbf{M}\,\mathbf{X}) = -\operatorname{tr}(\mathbf{L}_M^\top \mathbf{X}\,\mathbf{L}_M) \quad (3.53)$$

by means of basic linear algebra. Furthermore for any covariance matrix $\boldsymbol{\Sigma}$ the mapping $\mathbf{L}_M^\top \boldsymbol{\Sigma}\,\mathbf{L}_M$ can be interpreted as a Kalman belief update wrt. a specific dynamics and sensor model. Hence, from the basic property $\mathbf{X} \succeq 0 \Rightarrow \operatorname{tr}(\mathbf{X}) \geq 0$ we obtain

$$\forall \boldsymbol{\Sigma}_t^1 \succeq 0, \boldsymbol{\Sigma}_t^2 \succeq 0,\, \boldsymbol{\Sigma}_t^1 \succeq \boldsymbol{\Sigma}_t^2,\, \mathbf{M} \prec 0\colon \quad (3.54)$$

$$\operatorname{tr}(\mathbf{M}\boldsymbol{\Sigma}_t^1) = -\operatorname{tr}(\mathbf{L}_M^\top \boldsymbol{\Sigma}_t^1\,\mathbf{L}_M) \leq -\operatorname{tr}(\mathbf{L}_M^\top \boldsymbol{\Sigma}_t^2\,\mathbf{L}_M) = \operatorname{tr}(\mathbf{M}\boldsymbol{\Sigma}_t^2). \quad (3.55)$$

In combination with the properties of the Kalman belief update presented in [246] we finally arrive at

$$\begin{aligned}\forall \boldsymbol{\Sigma}_t^1 \succeq 0, \boldsymbol{\Sigma}_t^2 \succeq 0, t \in \{0:T\}\, k \in \{0:T-t\}:\\ \boldsymbol{\Sigma}_t^1 \succeq \boldsymbol{\Sigma}_t^2 \Rightarrow \mathbf{KU}^k(\boldsymbol{\Sigma}_t^1) \succeq \mathbf{KU}^k(\boldsymbol{\Sigma}_t^2) \Rightarrow \operatorname{tr}(\mathbf{M}_{k+t}^{V^\star}\mathbf{KU}^k(\boldsymbol{\Sigma}_t^1)) \leq \operatorname{tr}(\mathbf{M}_{k+t}^{V^\star}\mathbf{KU}^k(\boldsymbol{\Sigma}_t^1))\\ \Rightarrow \sum_{k=0}^{T} \operatorname{tr}(\mathbf{M}_{k+t}^{V^\star}\mathbf{KU}^k(\boldsymbol{\Sigma}_t^1)) \leq \sum_{k=0}^{T} \operatorname{tr}(\mathbf{M}_{k+t}^{V^\star}\mathbf{KU}^k(\boldsymbol{\Sigma}_t^2))\end{aligned} \tag{3.56}$$

as a consequence of (3.51)-(3.52) and by the fact that all $\mathbf{M}_{k+t}^{V^\star}$ are negative definite.

This means, that if $\boldsymbol{\Sigma}_t^1 \succeq \boldsymbol{\Sigma}_t^2$ the future rewards $t:T$ according to (3.44) of $\boldsymbol{\Sigma}_t^1$ will be less or equal than those of $\boldsymbol{\Sigma}_t^2$ for any fixed sequence of future sensor controls $\boldsymbol{u}^{\mathbf{z}}{}_{t:T}$. This is because the rewards associated with sensor states and controls will be the same.

Search Tree Pruning Assume candidate sensor states sequences $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{i}$ with associated accumulated rewards $R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{i})$ and belief co-variances $\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{i})$

$$\begin{aligned}\mathsf{C} = \{&(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{1}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{1}), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{1})), (\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{2}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{2}), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{2})),\\ &\ldots, (\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}))\}\end{aligned} \tag{3.57}$$

are given at time step t. If it holds for a candidate sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}$

$$\exists \boldsymbol{\delta} \in \Delta(n) := \{\delta_k \geq 0, \sum_{k=1}^{n} \delta_k = 1\}: \tag{3.58}$$

$$\operatorname{blk}\left(\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}), -R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1})\right) \succeq \sum_{k}^{n} \delta_k \operatorname{blk}(\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{k}), -R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{k})) \tag{3.59}$$

then by (3.51), (3.52) and (3.56) it follows that $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}$ needs not to be considered in the search for the maximizer: Any sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{n+1}$ that contains $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}$ will not result in higher reward than the best of those sequence that contain any of the other candidates $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{k<n+1}$. Hence, the properties of the Kalman belief update can be exploited to prune the search tree. This is illustrated at an one-dimensional example in Fig. 3.11.

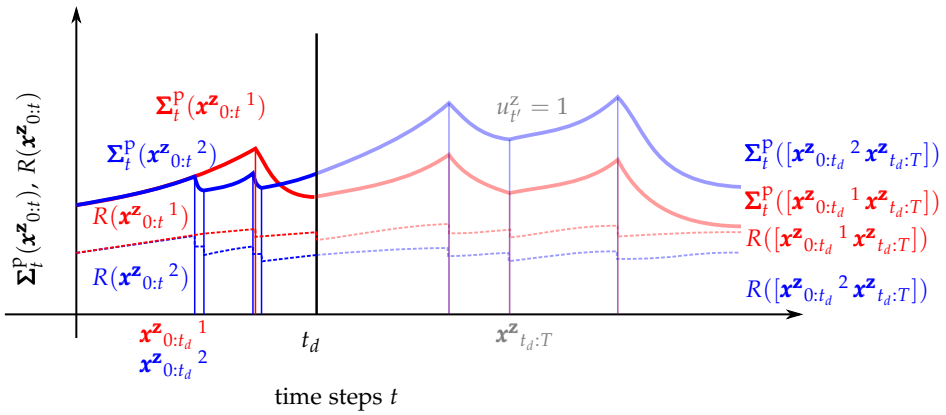

Figure 3.11: One dimensional example of the search tree pruning approach. Given are two sequences of sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{1}$ and $\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{2}$ until the current time step t_d. The changes in the sensor state are denoted by vertical solid lines. The evolution of the covariances are denoted by solid lines and the evolution of the accumulated rewards are denoted in dashed lines. It holds both $\boldsymbol{\Sigma}_{t_d}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{1}) \preceq \boldsymbol{\Sigma}_{t_d}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{2})$ and $R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{2}) < R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{1})$. For any succeeding sequence $\boldsymbol{x}^{\mathbf{z}}{}_{t_d:T}$ the property of the Kalman belief update leads to $R([\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{1}\, \boldsymbol{x}^{\mathbf{z}}{}_{t_d:T}]) > R([\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{2}\, \boldsymbol{x}^{\mathbf{z}}{}_{t_d:T}])$. As a result all branches of sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{2}$ can be pruned from the search tree.

The constraint satisfactory problem (3.58) for deciding whether the sequence ${\boldsymbol{x}^{\mathbf{z}}}_{0:t}{}^{k<n+1}$ can be neglected for solution is termed *algebraic redundancy check*. [246] suggested to conduct the algebraic redundancy check by standard convex programming techniques as [73, 74]. However, we developed a new simple gradient based technique that allows for more efficient solution. For this purpose, define

$$\boldsymbol{\Sigma}^j := \text{blk}\left(\boldsymbol{\Sigma}_t^{\text{p}}({\boldsymbol{x}^{\mathbf{z}}}_{0:t}{}^j), -R({\boldsymbol{x}^{\mathbf{z}}}_{0:t}{}^j)\right) \tag{3.60}$$

Now consider the following convex optimization problem

$$\min \quad \Big(\lambda_{\max}\big(\sum_{k=1}^{n} \delta_k \boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^{n+1}\big)\Big)_+ \tag{3.61}$$

$$\text{s.t. } \boldsymbol{\delta} \in \Delta(n), \tag{3.62}$$

where $\lambda_{\max}(\mathbf{X})$ denotes the maximum eigenvalue of a matrix $\mathbf{X}$ and where $(x)_+ = \mathbb{I}_{x>0}(x)\, x$. If it holds true $\lambda_{\max}\big(\sum_k^n \delta_k^\star \boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^{n+1}\big) \leq 0$ for the optimal solution $\boldsymbol{\delta}^\star$, then obviously condition (3.58) is met. Otherwise, any other $\boldsymbol{\delta}$ does also not fulfill condition (3.58).

To solve optimization problem (3.61), consider the gradient of the objective which is given as

$$\nabla_{\delta_k}\Big(\lambda_{\max}\big(\sum_k^n \delta_k \boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^{n+1}\big)\Big)_+|_{\boldsymbol{\delta}=\boldsymbol{\delta}^i} = \begin{cases} \mathbf{v}_{\lambda_{\max}}^\top \boldsymbol{\Sigma}^k \mathbf{v}_{\lambda_{\max}} & \text{if } \lambda_{\max}\big(\sum_k^n \delta_k^i \boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^{n+1}\big) \geq 0 \\ 0 & \text{else} \end{cases} \tag{3.63}$$

where $\mathbf{v}_{\lambda_{\max}}$ is any eigenvector associated with the current maximum eigenvalue $\lambda_{\max}\big(\sum_k^n \delta_k^i \boldsymbol{\Sigma}^k - \boldsymbol{\Sigma}^{n+1}\big)$ normalized to unit L2-norm, i.e. $1 = \|\mathbf{v}\|_{\text{L2}} := \sqrt{\sum_{k=1}^n v_k^2}$ [32]. Furthermore, projecting onto $\Delta(n)$ can efficiently be performed by the algorithm of [39]. Hence, we can try to minimize the optimization problem 3.61 by cycling between a gradient descent update of the current iterate $\boldsymbol{\delta}^i$ and a projecting the iterate on the simplex. This projected gradient technique is guaranteed to converge to a global optimum $\boldsymbol{\delta}^\star$. This is because it is an instance of convex proximal gradient descent [172]. We formulate this procedure in Algo. 5.

Algorithm 5 Algebraic Redundancy Check By Projected Gradient Descent AlgRed

function AlgRed($\{\mathbf{\Sigma}^k\}^{k=1:n}, \mathbf{\Sigma}^{n+1}$)
Require: Stepsize η, tolerance ε
  $\boldsymbol{\delta} \leftarrow$ Sample($\Delta(n)$) ▷ sample a random initial $\boldsymbol{\delta}$ from the simplex
  while not converged $\boldsymbol{\delta}$ **do**
    $\mathbf{D} \leftarrow \sum_k^n \delta_k \mathbf{\Sigma}^k - \mathbf{\Sigma}^{n+1}$
    $\lambda_{\max}, \mathbf{v}_{\lambda_{\max}} \leftarrow$ GetMaxEigen($\mathbf{D}$) ▷ standard linear algebra routine
    if $\lambda_{\max} < \varepsilon$ **then**
      break
    end if
    $\mathbf{v}_{\lambda_{\max}} \leftarrow \mathbf{v}_{\lambda_{\max}} / \|\mathbf{v}_{\lambda_{\max}}\|_{\mathrm{L2}}$
    $\forall_{k=1:n} d_k \leftarrow \mathbf{v}_{\lambda_{\max}}^\top \mathbf{\Sigma}^k \mathbf{v}_{\lambda_{\max}}$
    $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} - \eta \mathbf{d}$
    $\boldsymbol{\delta} \leftarrow$ ProjSimplex($\boldsymbol{\delta}, n$)
  end while
  return $\lambda_{\max}, \boldsymbol{\delta}$
end function

function ProjSimplex($\boldsymbol{\delta}, n$) ▷ Algorithm from [39]
  $\hat{\boldsymbol{\delta}} \leftarrow$ SortDescend($\boldsymbol{\delta}$) ▷ sort elements in descending order
  $s \leftarrow 0$
  $s_{\max} \leftarrow 0$
  flag $\leftarrow 0$
  for $i = 1 : n-1$ **do**
    $s \leftarrow s + \hat{\delta}_i$
    $s_{\max} \leftarrow (s-1)/i$
    if $s_{\max} \geq \hat{\delta}_{i+1}$ **then**
      flag $\leftarrow 1$
      break
    end if
  end for
  if flag $= 1$ **then**
    $s_{\max} \leftarrow (s + \hat{\delta}_n - 1)/n$
  end if
  $\boldsymbol{\delta} \leftarrow \max(\boldsymbol{\delta} - s_{\max}, \mathbf{0})$
  return $\boldsymbol{\delta}$
end function

Solving the SLQG Problem Following [246], we can now formulate an efficient solution approach for SLQGs. The algorithm starts with an initial candidate set C that contains the initial sensor state and its associated covariance as well as accumulated reward. At every time step, first the candidate set is expanded by applying both possible sensor controls and the corresponding beliefs and accumulated rewards are computed. Thereafter, the previously introduced technique is used to reject all candidates that cannot lead to improved return. The algorithm is outline in Algo. 6.

Algorithm 6 Optimal Solution of Joint Task Model under Secondary Task Model Restriction STRopt

function STROPT($\mathbf{C}_x, \mathbf{C}_u, r(u_t^z), r(x_t^z), (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z), x_0^z, \boldsymbol{\Sigma}_0^{\mathrm{p}}$)
Require: Tolerance ε
  $\left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_t^\star\right)_{t=0:T} \leftarrow$ LQROPT$(\mathbf{C}_x, \mathbf{C}_u, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{0:T})$ ▷ Algo. 1
  $\mathsf{C} \leftarrow \{(\boldsymbol{x}^{\mathbf{z}}{}_{0:0}{}^1 = x_0^z, \boldsymbol{\Sigma}_0^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:0}{}^1) = \boldsymbol{\Sigma}_0^{\mathrm{p}}, R(\boldsymbol{x}^{\mathbf{z}}{}_{0:0}{}^1) = -r(x_0^z))\}$ ▷ Initialize candidate set
  for $t = 1 : T$ **do**
    for element $e_i \in \mathsf{C}$ **do** ▷ Pass through candidate set according to order
      for $u_{t-1}^z = 0 : 1$ **do**
        $x_t^z \leftarrow x_{t-1}^z{}^i \oplus u_{t-1}^z$
        $\boldsymbol{x}^{\mathbf{z}}{}_{0:t} \leftarrow [\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i \; x_t^z]$
        $\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) \leftarrow$ KALMANUPDATE$(\boldsymbol{\Sigma}_{t-1}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i), \mathbf{A}_{t-1}, \mathbf{a}_{t-1}, \mathbf{B}_{t-1}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z))$
        $R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) \leftarrow R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i) - (r(x_t^z) + r(u_{t-1}^z) + \mathrm{tr}\left(\mathbf{M}_t^{V^\star} \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\right))$
        $\mathsf{C} \leftarrow$ INSERT$\Big((\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})), \mathsf{C}\Big)$ ▷ Insert into candidate set
      end for
      $\mathsf{C} \leftarrow$ DELETE$\Big((\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i)), \mathsf{C}\Big)$ ▷ Delete from candidate set
    end for
    $\mathsf{C} \leftarrow$ SORTASCEND$_{R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})}(\mathsf{C})$
    for element $e_i \in \mathsf{C}$, $i > 1$ **do** ▷ Pass through candidate set according to order
      $\lambda_{\max} \leftarrow$ ALGRED$\Big(\{\mathrm{blk}\left(\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^j), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^j)\right)\}^{j<i-1}, \mathrm{blk}\left(\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t-1}{}^i)\right)\Big)$
      ▷ Check for algebraic redundancy Algo. 5
      if $\lambda_{\max} < \varepsilon$ **then**
        $\mathsf{C} \leftarrow$ DELETE$\Big((\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^i, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^i), R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^i)), \mathsf{C}\Big)$
      end if
    end for
    $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^\star \leftarrow \arg\min_{\boldsymbol{x}^{\mathbf{z}}{}_{0:t} \in \mathsf{C}}(R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}))$ ▷ Pick sequence with max. cummulated reward from candidate set
  end for
  $m_0^{V^\star,1}(x_0^z) \leftarrow 0$
  for t=0:T **do**
    $m_0^{V^\star,1}(x_0^z) \leftarrow^+ r(x_t^z{}^\star) + r(u_t^z{}^\star)$
  end for
  return $\left((\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, \mathbf{F}_t^\star, \mathbf{f}_t^\star)_{t=0:T}, \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star, m_0^{V^\star,1}(x_0^z)\right)$
end function

Approximate and Heuristic Alternatives

Previously, we considered exact approaches for computing optimal policies in the joint task POMDP. Unfortunately, solving general SLQGs exactly using STRopt can be very demanding. We demonstrate this later in the evaluation in Sec. 3.6.2. Here, the main reason for the high computational demand is the size of the candidate set C_t. As an alternative, it is possible to compute an approximate solution by applying more aggressive pruning. For example, we could relax the tolerance in the algebraic redundancy check by a parameter ε

$$\exists \boldsymbol{\delta} \in \Delta(n) := \{\delta_k \geq 0, \sum_{k=1}^{n} \delta_k = 1\}: \tag{3.64}$$

$$\operatorname{blk}\left(\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1}), -R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{n+1})\right) + \varepsilon\mathbf{I} \succeq \sum_{k}^{n} \delta_k \operatorname{blk}(\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{k}), -R(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{k})). \tag{3.65}$$

Considering Algo. 6, note that the element with the current highest accumulated reward is always added to the candidate set. Hence, increasing ε makes the algorithm greedier and discards all candidates that do not have a significantly smaller covariance than the highest accumulated reward alternative. As shown in [246] this can tremendously decrease the candidate set. Furthermore, the authors prove that the loss in reward when using the relaxation parameter ε does not exceed a certain bound that is linear in ε. However, it must be noted that this bound is not uniform and can strongly depend on the model parameters. In regards to our application this means that using a fixed ε may result in loss of reward that depends on the external variables that parametrize the possible driving situations. Hence, a thorough evaluation of approximation quality over the likely model variations is required, which is considered an important issue for future research.

While relaxed tolerance can decrease computational demand by reducing the candidate set it never the less requires conducting the relaxed redundancy checks (3.64). As we will show in the analysis of the computational demands in Sec. 3.6.2 the redundancy checks can require significant computations. Therefore, one might wish to get rid of redundancy checks entirely. For example, one could consider to greedily pick the candidate with the highest accumulated reward. This can be interpreted as relaxed redundancy check using a high ε. Interestingly, this greedy algorithm has a bounded loss of performance if $-\log\det(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}))$ is used as a reward of the covariance [92]. Furthermore, a very similar greedy approach is used in [173] and in the works [220, 191, 192] for approximate solution of more complex models of normative gaze switching. However, even in the case of SLQGs in general no performance guarantees can be obtained [92]. Consequently, it needs evaluations of the heuristics with respect to their performance in application for computation of situationally appropriate glance behavior.

3.5.3 Maximum Causal Entropy Policies

In previous subsection, computing optimal policies for the joint task was considered. That is, we derived approaches to compute the optimal choice of sensor controls, i.e. gaze switches, as well as the optimal choice of the primary task controls, i.e. changes of the steering angle, and the control of the secondary task. To realize a distraction warning system based on the definition of appropriate glance behavior a gaze switch policy which results in eyes-off durations of high reward is desired. That is, we seek a gaze switch policy that obtains a very good trade-off between primary task performance and performance in the secondary task. As a consequence of the analysis of Sec. 3.5.2 this ensures bounded loss of vehicle control performances compared to fully attentive driving.

In contrast, the viewing time is not considered in the definition of appropriate glance Sec. 3.4. Hence, in this case a more realistic stochastic policy can also be used. Furthermore, note that the overall goal of this thesis is an improved distraction warning system. That is, we can only hope to assist the driver in applying a better glance strategy but can neither improve his or her vehicle control policy nor the strategy for fulfilling the secondary task. As a consequence, the human steering and typing policies must be taken into account. Although a good match of manual control data with optimal control models could be obtained, in many cases the fit was not perfect and delays [149, 42] or noise [27] needed to be introduced to increase realism. Therefore, if we compute glance behavior with respect to

optimal steering policies there is the risk that an overly optimistic assumption on the drivers steering performance is made. Consequently, the computed policy for gaze switches can be less suitable for realistic drivers e.g. as it results in too long off road glances because of assuming the absence of steering errors. Summarized, it can be desirable to take into account potential imperfect driver policies in computing rational glance policies. Fortunately, this can be achieved using the maximum causal entropy model of rational behavior. As stated previously, the maximum causal entropy policy $\tilde{\pi}_t(u_t|x_t)$ fulfills

$$\tilde{\pi}_t(u_t|x_t) \propto \exp\Big(\mathbb{E}\big[\sum_{t'=t}^{T} r(x_{t'}, u_{t'})\big|\tilde{\boldsymbol{\pi}}_{t:T}, \boldsymbol{\mathcal{P}}_{t:T}\big]\Big).$$

Hence, the gaze switching control $(u_t^z)^\dagger$ that results in the maximum reward under the expected future imperfect driver behavior can be obtain by means of

$$(u_t^z)^\dagger = \arg\max_{u_t^z} \log\left(\tilde{\pi}_t(u_t^z|\boldsymbol{\mu}^{\mathrm{p}}, \boldsymbol{\Sigma}^{\mathrm{p}}, x_t^{\mathrm{z}}, x_t^{\mathrm{i}})\right) = \arg\max_{u_t^z} \Big(\mathbb{E}\big[\sum_{t'=t}^{T} r(x_{t'}, u_{t'})\big|u_t^z, \tilde{\boldsymbol{\pi}}_{t:T}, \boldsymbol{\mathcal{P}}_{t:T}\big]\Big). \tag{3.66}$$

Consequently, the MCE policy model is suitable to obtain appropriate glance behavior for realistic driver behavior.

Soft Bellman Equations To compute the MCE policy we consider the soft value function $\tilde{V}_t(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}})$ and the soft state-control function $\tilde{Q}_t(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^z, u_t^{\mathrm{i}})$ analogously to the optimal policy. Here, the primary task states and controls are separated allowing similar treatment as in the maximum causal entropy policy in LQG (see Sec. 2.2.2). This yields the following terms

$$\tilde{Q}_t(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^z, u_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^z, u_t^{\mathrm{i}}) \tag{3.67}$$

$$\tilde{V}_t(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}}[\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{V}}[\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{\tilde{V},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}). \tag{3.68}$$

Here, the summands involved in $\tilde{Q}_t(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^z, u_t^{\mathrm{i}})$ are

$$\mathbf{M}_t^{\tilde{Q}} = \begin{cases} [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}}[\mathbf{A}_t\ \mathbf{B}_t] - \mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{if } t < T \\ -\mathrm{blk}(\mathbf{C}_x, \mathbf{C}_u) & \text{else} \end{cases} \tag{3.69}$$

$$\mathbf{m}_t^{\tilde{Q}} = \begin{cases} 2[\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}}\mathbf{a}_t + [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{\tilde{V}} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \tag{3.70}$$

$$m_t^{\tilde{Q},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^z, u_t^{\mathrm{i}}) = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{\tilde{V}}\mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{\tilde{V}} \\ \quad + \mathrm{tr}(-\mathbf{C}_x \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \\ \quad + \mathrm{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^z])\right)\right) \\ \quad + r(u_t^z) + r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \\ \quad + \mathbb{E}\left[m_{t+1}^{\tilde{V},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^z], x_{t+1}^{\mathrm{i}})\middle|\mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^{\mathrm{z}}, u_t^z; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right] & \text{if } t < T \\ -\mathrm{tr}(\mathbf{C}_x \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T})) + r(u_T^{\mathrm{z}}) + r(x_T^{\mathrm{i}}, u_T^{\mathrm{i}}) & \text{else} \end{cases}. \tag{3.71}$$

As in previously considered POMDPs and MDPs $\tilde{Q}_t, Q_t^\star$ have the same structure. In contrast for the parts of the soft value function $\mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}})$ it holds

$$\mathbf{M}_t^{\tilde{V}} = \mathbf{M}^{\tilde{Q}}_{t\,x,x} - \mathbf{M}^{\tilde{Q}}_{t\,x,u}[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1}\mathbf{M}^{\tilde{Q}}_{t\,u,x} \tag{3.72}$$

$$\mathbf{m}_t^{\tilde{V}} = \mathbf{m}^{\tilde{Q}}_{t\,x} - \mathbf{M}^{\tilde{Q}}_{t\,x,u}[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1}\mathbf{m}^{\tilde{Q}}_{t\,u} \tag{3.73}$$

$$\begin{aligned} m_t^{\tilde{V},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = &-\frac{1}{4}[\mathbf{m}^{\tilde{Q}}_{t\,u}]^\top[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1}\mathbf{m}^{\tilde{Q}}_{t\,u} + \frac{1}{2}\log(\det(\pi[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1})) \\ &+ \operatorname{tr}(-\mathbf{C}_x\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \operatorname{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}}\left(\mathbf{A}_t\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x}\right)\right) \\ &+ \operatorname{softmax}_{u_t^z, u_t^{\mathrm{i}}}\left(r(u_t^z) - \operatorname{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}}\boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z])\right)\right. \\ &\left. + r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) + \mathbb{E}\left[m_{t+1}^{\tilde{V},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^z \oplus u_t^z], x_{t+1}^{\mathrm{i}})\middle|\mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^z, u_t^z; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]\right). \end{aligned} \tag{3.74}$$

Consequently, in the joint task model the MCE policy admits the same factorization as the optimal policy and the primary task policy (see (2.71)) is obtained as

$$\tilde{\pi}_t(u_t^{\mathrm{p}}|\boldsymbol{\mu}_t^{\mathrm{p}}) = \mathcal{N}(u_t^{\mathrm{p}}|\tilde{\mathbf{F}}_t\boldsymbol{\mu}_t^{\mathrm{p}} + \tilde{\mathbf{f}}_t, \boldsymbol{\Sigma}_t^{u^{\mathrm{p}}}) \tag{3.75}$$

$$\tilde{\mathbf{F}}_t := -\frac{1}{2}[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1}\mathbf{M}^{\tilde{Q}}_{t\,u,x}, \quad \tilde{\mathbf{f}}_t := -\frac{1}{2}[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1}\mathbf{m}^{\tilde{Q}}_{t\,u}, \quad \boldsymbol{\Sigma}_t^{u^{\mathrm{p}}} := -\frac{1}{2}[\mathbf{M}^{\tilde{Q}}_{t\,u,u}]^{-1}. \tag{3.76}$$

Furthermore, we also face the issue of intractability of the soft Bellman equations in the general case. In the MCE policy model the exponentially growing state space of $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ is even more problematic. In the conditional distribution defined by the MCE policy $\tilde{\pi}$ any sensor control that results in finite return has a likelihood greater than zero. As a consequence, we can expect that a greater proportion of the state space $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ will be visited when following the MCE policy. Therefore, we will revisit the techniques that were presented for computing the optimal policy of the joint task model. Here we will investigate whether these are applicable to obtain the MCE policies.

Sensor Model Restriction

Fortunately, assuming immediate saturation of information once gaze returns to the road results in tractable computation of the maximum causal entropy policy. We can also replace the sequence of past sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ with the EOD d_t and obtain Algo. 7 that is very similar to Algo. 4:

Algorithm 7 MCE Policy of Joint Task Model under Sensor Model Restriction SRMCE

function SRMCE($\mathbf{C}_x, \mathbf{C}_u, r(u_t^z), r(x_t^i, u_t^i), (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \mathbf{\Sigma}^{\epsilon^z}(x_t^z), \mathcal{P}^i, x_0^z, \mathbf{\Sigma}_0^p, \hat{\mathbf{\Sigma}}_0^p$)

  $\left[\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \mathbf{\Sigma}_t^{u^p}\right)_{t=0:T} \leftarrow$ LQRMCE$(\mathbf{C}_x, \mathbf{C}_u, \left(\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t\right)_{t=0:T})$ ▷ Algo. 3

  for $t = 0 : T$ **do** ▷ Pre-compute belief for gaze on road

    $\hat{\mathbf{\Sigma}}_{t+1}^p \leftarrow$ KalmanUpdate$(\hat{\mathbf{\Sigma}}_t^p, \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(0), \mathbf{\Sigma}^{\epsilon^z}(0))$ ▷ Algo. 2

  end for

  for $t = 0 : T$ **do** ▷ Forward pass for belief for gaze off road

    if $t = 0$ **then**

      $\mathbf{\Sigma}^1 \leftarrow \mathbf{\Sigma}_0^p$

      $d_0 \leftarrow x_0^z$

    else

      $\mathbf{\Sigma}^1 \leftarrow \hat{\mathbf{\Sigma}}_t^p$

      $d_0 \leftarrow 0$

    end if

    for $d_t = d_0 : T - t - 1$ **do** ▷ Roll-out off road belief until max. duration

      $t' \leftarrow t + d_t - d_0 - 1$

      if $t' > 0$ **then**

        $\mathbf{\Sigma}^1 \leftarrow$ KalmanUpdate$(\mathbf{\Sigma}^1, \mathbf{A}_{t'}, \mathbf{a}_{t'}, \mathbf{B}_{t'}, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(1), \mathbf{\Sigma}^{\epsilon^z}(1))$

      end if

      $\mathbf{\Sigma}^2 \leftarrow$ KalmanUpdate$(\mathbf{\Sigma}^1, \mathbf{A}_{t'+1}, \mathbf{a}_{t'+1}, \mathbf{B}_{t'+1})$

      $\mathbf{\Sigma}^3 \leftarrow$ KalmanUpdate$(\mathbf{\Sigma}^1, \mathbf{A}_{t'+1}, \mathbf{a}_{t'+1}, \mathbf{B}_{t'+1}, \mathbf{\Sigma}^{\epsilon^x}, \mathbf{H}(1), \mathbf{\Sigma}^{\epsilon^z}(1))$

      if $d_t = 0$ **then**

        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{\tilde{Q},1}(d_t, x_t^i, 0, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \hat{\mathbf{\Sigma}}_{t'+1}^p) + \operatorname{tr}(\mathbf{M}_{t'+2}^{\tilde{V}}(\mathbf{\Sigma}^2 - \hat{\mathbf{\Sigma}}_{t'+2}^p)) + r(u_t^z = 0)$

        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{\tilde{Q},1}(d_t, x_t^i, 1, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \hat{\mathbf{\Sigma}}_{t'+1}^p) + \operatorname{tr}(\mathbf{M}_{t'+2}^{\tilde{V}}(\mathbf{\Sigma}^2 - \mathbf{\Sigma}^3)) + r(u_t^z = 1)$

      else

        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{\tilde{Q},1}(d_t, x_t^i, 0, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \mathbf{\Sigma}^1) + \operatorname{tr}(\mathbf{M}_{t'+2}^{\tilde{V}}(\mathbf{\Sigma}^2 - \mathbf{\Sigma}^3)) + r(u_t^z = 0)$

        $\forall_{x_t^i, u_t^i}\ m_{t'+1}^{\tilde{Q},1}(d_t, x_t^i, 1, u_t^i) \leftarrow -\operatorname{tr}(\mathbf{C}_x \mathbf{\Sigma}^1) + \operatorname{tr}(\mathbf{M}_{t'+2}^{\tilde{V}}(\mathbf{\Sigma}^2 - \hat{\mathbf{\Sigma}}_{t'+2}^p)) + r(u_t^z = 1)$

      end if

    end for

  end for

  for $t = T : 0$ **do** ▷ Use soft Bellman equations to obtain remaining part

    $\forall_{d_t \in \{0:T\}, x^i, u^z, u^i}\ m_t^{\tilde{Q},1}(d_t, x^i, x^z, u^i) \leftarrow^+ r(x^i, u^i)$

    for $d_t = 0 : \min(t+1, T)$ **do**

      if $d_t = 0$ **then**

        $d'(u^z = 0) \leftarrow 0,\ d'(u^z = 1) \leftarrow 1$

        $\forall_{x_t^i, u_t^z, u_t^i}\ m_t^{\tilde{Q},1}(d_t, x_t^i, u_t^z, u_t^i) \leftarrow^+ +\mathbb{E}[m_{t+1}^{\tilde{V},1}(d'(u_t^z), x_{t+1}^i) | \mathcal{P}_t^i(x_{t+1}^i | 0, u_t^z; x_t^i, u_t^i)]$

      else

        $d'(u_t^z = 0) \leftarrow d + 1,\ d'(u_t^z = 1) \leftarrow 0$

        $\forall_{x_t^i, u_t^z, u_t^i}\ m_t^{\tilde{Q},1}(d_t, x_t^i, u_t^z, u_t^i) \leftarrow^+ \mathbb{E}[m_{t+1}^{\tilde{V},1}(d'(u_t^z), x_{t+1}^i) | \mathcal{P}_t^i(x_{t+1}^i | 1, u_t^z; x_t^i, u_t^i)]$

      end if

      $\forall_{x_t^i} m_t^{\tilde{V},1}(d_t, x_t^i) \leftarrow \operatorname{softmax}_{u_t^z, u_t^i}(m_t^{\tilde{Q},1}(d_t, x_t^i, u_t^z, u_t^i))$

      $\forall_{x_t^i} \tilde{\pi}_t(u_t^z, u_t^i | d_t, x_t^i) \leftarrow \exp\left(m_t^{\tilde{Q},1}(d_t, x_t^i, u_t^z, u_t^i) - m_t^{\tilde{V},1}(d_t, x_t^i)\right)$

    end for

  end for

  return $\left(\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \mathbf{\Sigma}_t^{u^p}, \tilde{\pi}_t(u^z, u^i | d, x_t^i)\right)_{t=0:T}$

end function

Secondary Task Model Restriction

Restricting the secondary task model to the simple model, allows to apply recent solution techniques for sensor scheduling in linear quadratic Gaussian problems as shown in Sec. 3.5.2. Here, solution exploited properties of the Kalman belief update to prune sub-optimal branches of the search tree at every time step. Unfortunately, this technique is not applicable for computing the MCE policy. In a MCE policy model also suboptimal behavior has a non-zero probability. Consequently, branches of the search tree identified as being suboptimal cannot simply be neglected in this policy model. Still, due to connection between the maximum likelihood policy and the optimal policy (2.54) pruning can be used to obtain the most likely sequence of sensor states $(\boldsymbol{x}^{\mathbf{z}}{}_{0:T})^{\dagger}$.

3.6 Evaluation

In the first section we presented and discussed several potential models for secondary task engagement during lane keeping. Thereafter, approaches for obtaining gaze switching policies were derived exploiting factorization properties. Furthermore, all tractable algorithms for policy computation came with further restrictions on either the sensor model or the secondary task model. As the computed policies shall be used in a distraction warning system, it is necessary to critically review and validate the structural properties of the policies. In addition to, that we also need to address the question whether the developed algorithms are indeed computationally feasible for usage in an online warning system.

3.6.1 Realism of Computed Appropriate Glance Behavior

An obvious objective for the computed appropriate glance behavior is high realism with respect to characteristics of drivers' behavior. Previously, we have already discussed the assumptions and approximations made in both the vehicle and the sensor model. Furthermore, we addressed how an exemplar secondary task can be represented as a sub-MDP in the joint task model. Given these models as well as the reward function the optimal policy (3.39) and the maximum causal entropy policy (3.75) were obtained in special analytic forms. Here, all policies factorized into a policy for the sensor control, i.e. gaze switch, that was independent of the primary task states. Furthermore, under restriction of the sensor model the policy of the sensor control was only dependent on eyes-off duration d_t. Note, that these properties are independent from the specific numerical values of the reward parameters or the parameters of the dynamics. Hence, in this subsection we will validate these two fundamental properties of the computed policies with respect to experimental data. This serves the purpose of checking whether the aspects of the joint task important for the driver have been incorporated in the POMDP model.

The Quadratic Primary Task Reward

In this context, let us first consider the policy factorization into two independent policies for the primary task and the secondary task plus the sensor model present in all the considered approaches 3.39, 3.75. Under this policy factorization the rational policy for the gaze switches depends on the external influences $\mathbf{v}_{0:T}, \boldsymbol{\kappa}_{0:T}$ but importantly does *not* depend on the current primary task state $\mathbf{x}_t^{\mathrm{p}}$. That is, deciding to avert gaze is independent on whether the vehicle is close to the lane borders or in the center of the lane. This is intuitively not very plausible but note that this factorization was no result of an approximation technique. Instead it is directly related to the combination of the quadratic reward model and the linear-affine kinematic model of the primary task. Using a different reward function

can result in optimal policies for gaze switches that are dependent on the primary task state $\mathbf{x}_t^{\mathrm{p}}$. This is demonstrated at the small toy POMDP,

$$\mathbf{x}_{t+1}^{\mathrm{p}} = 1.01\,\mathbf{x}_t^{\mathrm{p}} + u_t^{\mathrm{p}} + \boldsymbol{\epsilon}^{\mathrm{p}}, \quad \boldsymbol{\epsilon}^{\mathrm{p}} \sim \mathcal{N}(0, 0.2) \tag{3.77}$$

$$z_t = x_t^{\mathrm{z}}\,\mathbf{x}_t^{\mathrm{p}} \tag{3.78}$$

$$r(x_t^{\mathrm{z}}, u_t^{\mathrm{z}}) = \begin{bmatrix} -0.1 & -0.1 \\ 0.2 & -0.1 \end{bmatrix} \tag{3.79}$$

$$r(u_t^{\mathrm{p}}) = -0.25\,(u_t^{\mathrm{p}})^2 \tag{3.80}$$

$$\text{for either a quadratic reward } r(\mathbf{x}_t^{\mathrm{p}}) = -0.2\,(\mathbf{x}_t^{\mathrm{p}})^2 \tag{3.81}$$

$$\text{or a indicator reward } r(\mathbf{x}_t^{\mathrm{p}}) = \mathbb{I}_{|\mathbf{x}_t^{\mathrm{p}}|<0.8}(\mathbf{x}_t^{\mathrm{p}}). \tag{3.82}$$

For numerical solution the space of the primary task state and the primary task control were discretized from $[-5, 5]$ at a resolution of 0.1. Furthermore, the eyes-off duration was restricted to a maximum of 6 time steps. For single purpose of this toy example the infinite horizon solution wrt. a discount of $\gamma = 0.9$ was computed. We refer to [181] for the precise definition and the properties of that form of solution. In the remaining part of this thesis we will always consider *finite horizon* problems. Hence, we omitted introducing those problems.

As in previous cases, the value functions and policies wrt. the a-posterior mean of the primary task state $\boldsymbol{\mu}_t^{\mathrm{p}}$ and EOD d_t are considered. Fig. 3.12, Fig. 3.13 and Fig. 3.14 depict the solution wrt. the quadratic reward function on the left and the solution wrt. the indicator reward function on the right.

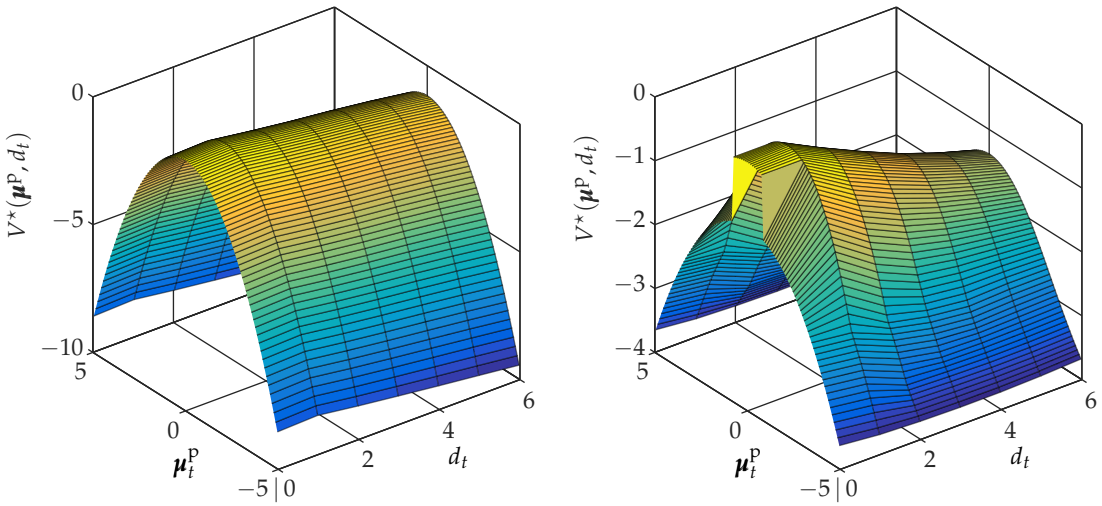

Figure 3.12: Value functions $V^\star(\boldsymbol{\mu}^{\mathrm{p}}, d_t)$ for the discretized infinite-horizon POMDP with quadratic reward on the primary task state $\mathbf{x}_t^{\mathrm{p}}$ (left) and with indicator reward on the primary task state $\mathbf{x}_t^{\mathrm{p}}$ (right).

As can be seen in the left part of Fig. 3.12, the quadratic reward function produces a quadratic value function. Here, the width of the quadratic function wrt. $\boldsymbol{\mu}_t^{\mathrm{p}}$ remains static, while a constant offset depended on d_t is added. Fig. 3.14 shows that the policy for the primary task control is a linear function of the a-posterior mean $\boldsymbol{\mu}_t^{\mathrm{p}}$ and is independent from the sensor state x_t^{z}. Furthermore, the policy for the sensor control u_t^{z} is independent of the expected primary task state $\boldsymbol{\mu}_t^{\mathrm{p}}$ as depicted in Fig. 3.13. In contrast, the width of the value function for the indicator reward wrt. $\boldsymbol{\mu}_t^{\mathrm{p}}$ depends on the EOD d_t as can be seen in the left part of Fig. 3.12. Consequently, the optimal policy wrt. to the sensor control u_t^{z} depends on the expected primary task state $\boldsymbol{\mu}_t^{\mathrm{p}}$ which is shown in Fig. 3.13.

Hence, if the primary task model and the sensor model is accepted, the policy factorization can directly be attributed to a quadratic reward. Consequently, if human gaze switch policy does not show this factorization this suggest the presence of a different reward function of the primary task states. For example, an additional indicator term could be more appropriate.

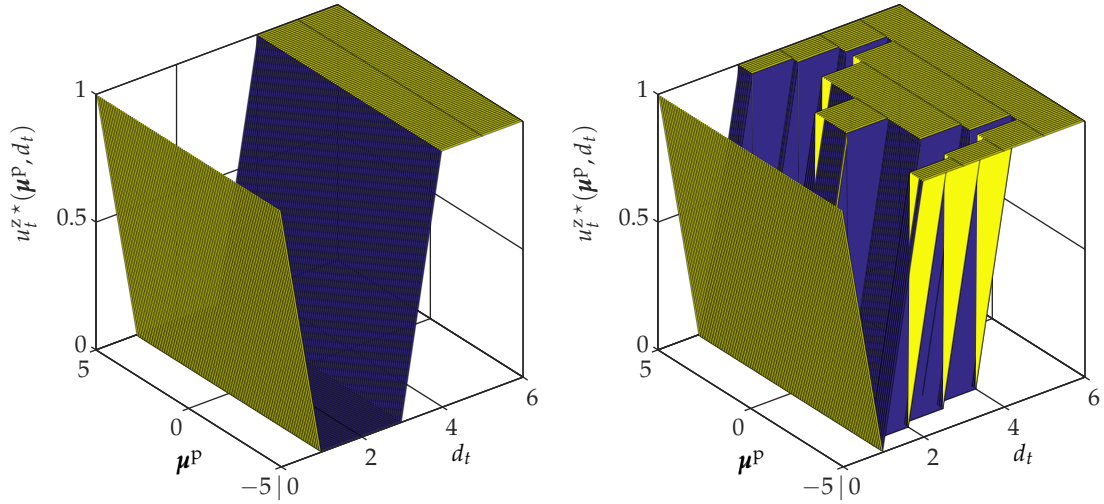

Figure 3.13: Sensor control switch policy wrt. $u_t^{z\star}(\boldsymbol{\mu}^{\mathrm{p}}, d_t)$ for the discretized infinite-horizon POMDP with quadratic reward on the primary task state $\mathbf{x}_t^{\mathrm{p}}$ (left) and with indicator reward on the primary task state $\mathbf{x}_t^{\mathrm{p}}$ (right).

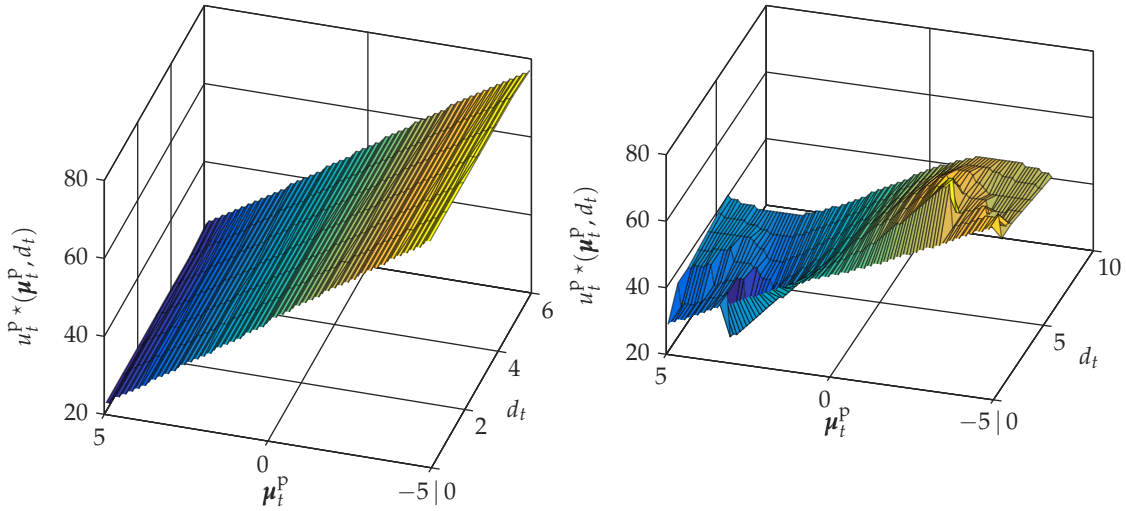

Figure 3.14: Primary task policy $u_t^{\mathrm{p}\star}(\boldsymbol{\mu}^{\mathrm{p}}, d_t)$ for the discretized infinite-horizon POMDP with quadratic reward on the primary task state $\mathbf{x}_t^{\mathrm{p}}$ (left) and with indicator reward on the primary task state $\mathbf{x}_t^{\mathrm{p}}$ (right).

Intuitively, one would expect the driver's policy to be dependent on the primary task states, especially the position in lane y_t. However, we are not aware of any work that empirically investigated this hypothesis. Therefore, we used the experimental data obtained throughout this work to test this hypothesis. For this purpose we analyzed the drivers' gaze switch behavior in the data of experiment II. Specifically, for periods where the drivers engaged in a secondary task that required to gaze off the road, we considered the lane positions. In the intervals the drivers had their gaze on the road the absolute value of the lane position in the middle of the interval $|y_{t^{\mathrm{m}}}|$ was compared to the absolute value at the end of the interval $|y_{t^{\mathrm{e}}}|$ when the driver averted his or her gaze.

The distribution of the absolute values of the lane positions $|y_{t^{\mathrm{m}}}|, |y_{t^{\mathrm{e}}}|$ for both time points as well as the distribution of the difference of the absolute values $|y_{t^{\mathrm{m}}}| - |y_{t^{\mathrm{e}}}|$ are depicted in Fig. 3.15.

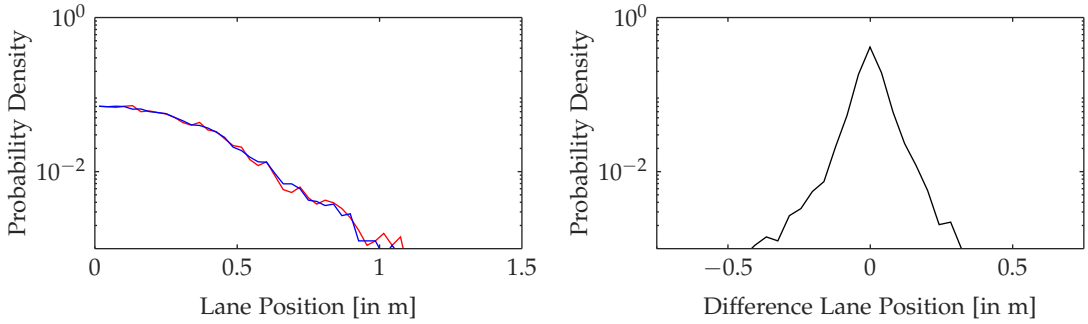

Figure 3.15: Distribution of the absolute value of the lane positions $|y_{t^m}|, |y_{t^e}|$ before (time t^m) and at gaze switch (time t^e) (left plot) and difference of lane positions $|y_{t^m}| - |y_{t^e}|$ (right plot).

As can be seen in the plots, the distributions for both time points are very similar. This is supported by a median difference that differed not significantly from zero $p_{\text{test}} = 0.64$ according to a signed-rank test.

Hence, in the lane keeping scenarios considered in this thesis the policy factorization induced by the choice of rewards and dynamics does not conflict with empirical data. Specifically, the hypothesis that deciding to switch gaze is independent of the absolute lane position could not statistically be rejected. However, we wish to note that in the experimental data underlying this analysis the vehicle was always well inside the lane. This can be seen considering the fact that the lane boundaries were at $-1.75, 1.75$ meters, whereas the probability mass of the lane positions was inside ± 1 m. Similar distributions of the lane position were also present in the other driving experiments of this thesis. Consequently, it is possible that driver's gaze switching policies depend on the lane position when the vehicle is close to the lane borders. This is an aspect that should be investigated in future work.

The Sensor Model Restriction

Among the two considered further restrictions on the joint task POMDP, restricting the sensor model is especially important. This is because it allows efficiently and exactly compute both the optimal and the maximum causal entropy policy. Furthermore, full flexibility for modeling the secondary task is provided. In the approach of restricting the sensor model, the key aspect was that we could replace the sequence of sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ with the eyes-off duration d_t. This is because immediately all available information is received when the driver returns his or her gaze to the road. Consequently, the sensor control policy for averting gaze is independent of the duration of the current glance on the road which we will refer to as the *viewing time*.

In contrast to the applied policy model interactions between viewing-time, eyes-off duration and driving behavior have been found in the literature. The influence of the viewing time in lane keeping have previously been investigated in occlusion experiments [207, 69]. Here, special glasses where used to fully occlude the driver's vision. In these experiments occlusion time and viewing time as well as driving speed were varied. Under occlusion times of 1 s to 9 s and deliberately chosen driving speed, [207] reported that a viewing time of 0.25 s was the minimum practical time. Furthermore, a viewing time above 0.5 s did not further increase the chosen driving speed and it was concluded that 0.5 s suffices to obtain all available information necessary for lane keeping. [69] (Cpt. 7.3) investigated the relation of deliberately chosen occlusion time at a fixed driving speed and a fixed viewing time. While median occlusion and viewing time strongly correlated at 20 km/h (increase of occlusion time from 4.5 s to 6.5 s at viewing times 0.25 s and 4.00 s) turned out to be small at speeds 60, 100 km/h. Summarized, occlusion experiments indicate that there is a minimum viewing time and that viewing time can depend on occlusion time especially at low speeds.

In contrast to complete occlusion of vision as in these driving experiments, in many secondary tasks drivers can to a small extent sense the vehicle's position and orientation in lane. This was for example shown in [230]: In the presented driving experiment the distance driven without lane departure negatively correlated with the amount of angular deviation of gaze from the forward road

scenery. Furthermore, the secondary task structure can also influence the viewing time, which has already been discussed previously in this thesis. Hence, it is not clear if the observations made wrt. viewing time generalize to naturalistic interaction with secondary tasks while driving. Therefore, we analyzed the interaction of the duration of glances off the road with the succeeding viewing time. This was done using the data of experiment II which is presented in detail later in this thesis (Sec. 5.6). The experiment had a 3×3 design considering a secondary task displayed at three different positions and three driving speeds $80, 90, 110$ km/h.

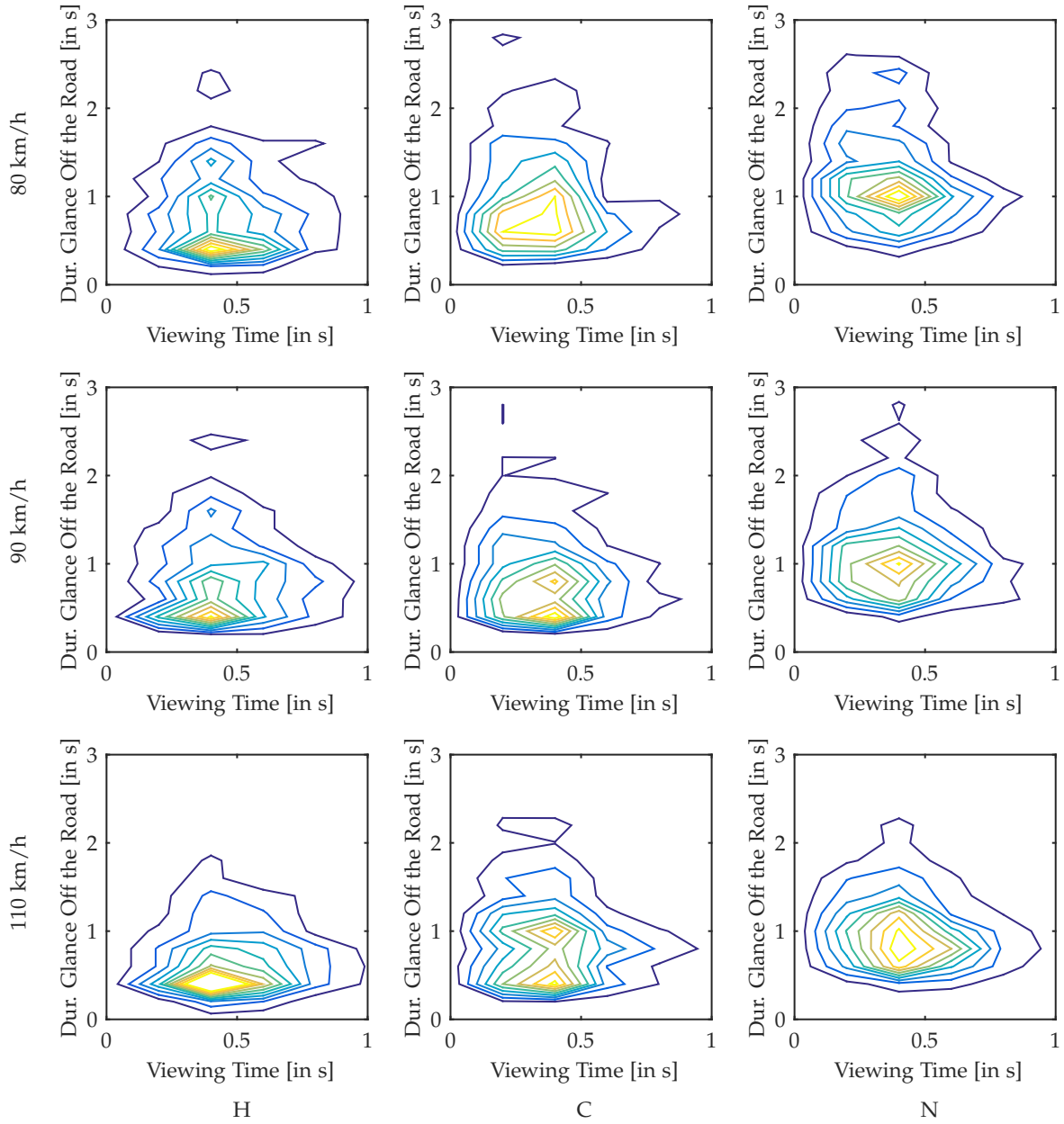

Figure 3.16: Joint distribution of the duration of glances off the road and succeeding glances on the road in the conditions of experiment II. Columns correspond to the different secondary tasks. Rows correspond to the different driving speeds. Plots shows the empirical quantile levels as obtained from a discretization in 0.2 s steps.

The joint distribution of the duration of glances off the road and the viewing time of the successive glances on the road is depicted in Fig. 3.16. As can be seen from the plots, in contrast to the occlusion experiments both variables seem to be rather uncorrelated. A test on Pearson's correlation coefficient

revealed that the null hypothesis of no positive correlation of the duration of glances of the road and the succeeding viewing time could not be rejected for any combination of experimental conditions. The results of the correlation analysis are summarized in Tab. 3.2.

Tabular 3.2: Pearsons's correlation coefficient *rho* between the duration of glances off the road and succeeding viewing time

Driving Speeds	**Display Positions**		
	H	C	N
80 km/h	$\rho = -0.07$, $p_{\text{test}} = 0.94$	$\rho = -0.05$, $p_{\text{test}} = 0.89$	$\rho = -0.02$, $p_{\text{test}} = 0.69$
90 km/h	$\rho = -0.14$, $p_{\text{test}} = 0.99$	$\rho = -0.07$, $p_{\text{test}} = 0.97$	$\rho = -0.05$, $p_{\text{test}} = 0.93$
110 km/h	$\rho = -0.17$, $p_{\text{test}} = 0.99$	$\rho = -0.09$, $p_{\text{test}} = 0.99$	$\rho = -0.10,$, $p_{\text{test}} = 0.99$

In contrast to previous occlusion experiments, we were not able to find any positive correlation between the duration of glances off the road and the succeeding viewing time. With respect to glance policies that depend only in eyes-off duration d_t data shows no conflicting evidence. As stated before, we believe that this is most likely the case because of the influence of the specific secondary task and because in secondary task engagement drivers can to a small extend sense the road scenery.

3.6.2 Computational Feasibility

In previous subsection we validated certain structural properties of the computed optimal and maximum causal entropy policies with respect to experimental data. Fortunately, no significant conflicting evidence could be found. Consequently, we are free in the choice of the developed algorithms to obtain appropriate glance behavior considering the realism of their structural properties. This is important, as besides realism of glance policies fast computation is of high importance in their application in a real time warning system. Therefore, the present subsection discusses aspects computational feasibility.

Computational Demand

As the first step towards assessing the computational feasibility, we analyze the total computational demand resulting from the operations involved in SRopt (Algo. 4) and STRopt (Algo. 6). For the purpose of comparability, the application to the joint task POMDP using the simple secondary task model is considered. In this case, STRopt computes an optimal solution for the general case (Sec. 3.5.2), while applying SRopt comes with a further restriction on the sensor model (Sec. 3.5.2).

Theoretical Analysis We first theoretically analyze the computational demand of SRopt and STRopt by considering the computation flow in both algorithms. In both cases first the same deterministic LQR problem is solved. Thereafter, the algorithms differ in the way how they compute the policy regarding sensor states and controls.

SRopt (Algo. 4) first enumerates all possible beliefs. This is done in the following way. Starting in $d_t = 0$ with the steady state covariance $\hat{\mathbf{\Sigma}}_t^{\text{p}}$ incrementally the $T - t$ covariances $\mathbf{\Sigma}_{t+k}^{\text{p}}(d_{t+k} = k\Delta t)$, where $\Delta t = 0.04$ s according to the model frequency, are computed. Hence, in total $T(T+1)/2$ Kalman belief updates are required. In the second part of this algorithm the final policy is obtained using the Bellman equation. This requires finding the binary sensor control that maximizes the state-control function for every d_t, which can be done in parallel. Therefore, its computational demand is neglect-able in comparison to the Kalman belief updates under moderate T and small loop overhead.

STRopt (Algo. 6) iteratively expands the candidates set C_t requiring $2|\mathsf{C}_t|$ Kalman belief updates. Thereafter, the algebraic redundancy check is conducted. The redundancy is done in iterative fashion by incrementally testing new candidates with respect to the previously accepted ones. In total $2|\mathsf{C}_t| - 1$ algebraic redundancy checks are required. In this context, every iteration i of the projected gradient Algo. 5 in the redundancy check requires an eigenvalue computation. We use i_t^{avg} to denote the average number required for the redundancy checks for the expansion of candidate set C_t.

For the purpose of comparison, we define the computational demand of the SRopt and STRopt as the number of required Kalman belief updates and eigenvalue computations $C(\text{Algo. } 4)$ and $C(\text{Algo. } 6)$.

This quantity is used as an estimate of the total computational demand. In this context, the assumption is made that computing the Kalman belief update and computing eigenvalues are approximately equally demanding and that these are the dominating factors in both algorithms. Considering the number of Kalman belief updates and eigenvalue computations in both approaches we thus obtain an estimate of the total computational demand as

$$C(\text{SRopt}) = T\,(T+1)/2 \quad \text{vs.} \quad C(\text{STRopt}) = \sum_{t=0}^{T} (2|\mathsf{C}_t| + (2|\mathsf{C}_t| - 1)\, i_t^{\text{avg}}). \tag{3.83}$$

As a direct consequence the increased demand when not restricting the sensor model is strongly dependent on the growth of $|\mathsf{C}_t|$ and the average number of iterations i_t^{avg} of the projected gradient approach in the redundancy check. In the following both quantities will empirically be analyzed.

Empirical Analysis Although the projected gradient algorithm is known to obtain linear convergence, there is in general no bound on the number of iterations needed to numerical convergence. Similarly, the dynamic behavior of the candidate set C_t has not been subject to detailed research in the literature. Therefore, we studied these variables empirically with respect to the SLQGs present in our application.

For this purpose, 50 random SLQG problems of 76 time steps (corresponding to 3 s) were sampled from the driving situations encountered in the third secondary task of experiment II (Sec. 5.6). These driving situations differed in their initial value and in the velocity profile $\mathbf{v}_{1:76}$ as well as the lane curvature profile $\boldsymbol{\kappa}_{0:76}$. In the evaluation the sensor model parameters estimated in the numerical experiment of Cpt. 5 were used. Furthermore, we set the reward parameters $\boldsymbol{\theta}_{1:5}$ to those estimated in Cpt. 5 but increased the reward parameter θ_6 for the simple secondary task reward $\mathbb{I}_{x^z=0}(x_t^z)$ by a factor of 40. As those reward parameters were specifically estimated for the MCE policy, the scaling was necessary to obtain similar eyes-off durations. Finally, the tolerance of the algebraic redundancy check was set to 10^{-6}. The projected gradient descent used a constant step size of 1, as well as a relative tolerance of 10^{-4}.

We report on the average number of projected gradient iterations per time step and on the size of the candidate set. The results of the empirical evaluation are depicted in Fig. 3.17.

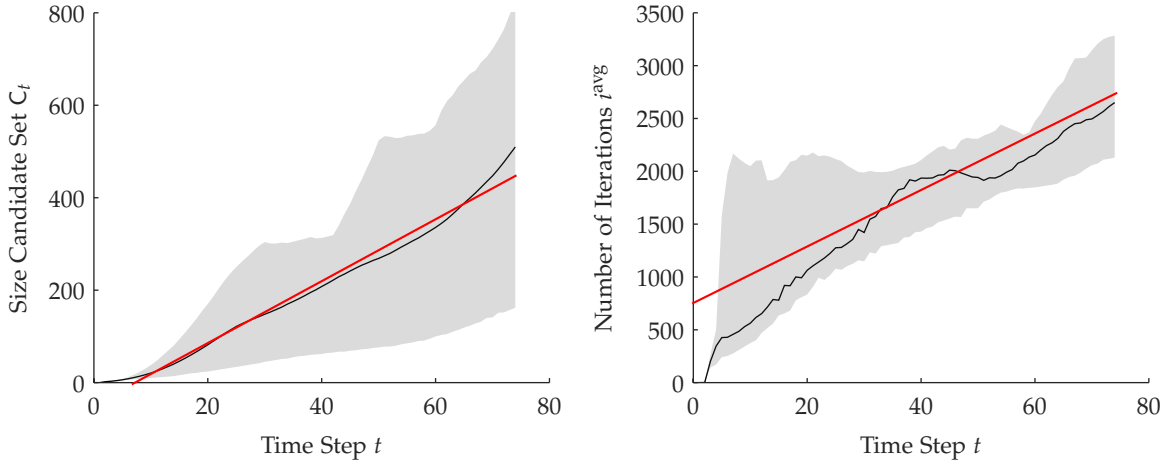

Figure 3.17: Results of the empirical evaluation of the pruning based solution of the joint task model STRopt (Algo. 6). Plot depicts the size of the candidate set C_t (left) and the average number of projected gradient iterations per time step (right). Solid black line denotes the median and shaded area the [0.25,0.75] confidence band over 50 randomly sampled SLQGs. The red line depicts the least squares fit of a linear affine function to the data.

As shown in the figure, the number of elements in the candidate showed an approximate linear growth up to a median number of elements of 500. Note that the size of the candidate set is not unrealistically large. [246] reported a candidate sets of 113 elements for a two dimensional toy example. In that work neither the dynamics nor the reward of the belief covariance were time-dependent. This is in

contrast to the SLQGs considered in this thesis (see (3.44)). Time dependency is likely to increase the size of candidate set set. This is because in this cases many different sensor sequences can result in similar accumulated rewards which prohibits strong pruning. Comparing only the mean total number of Kalman belief updates required in STopt and STRopt we obtain

$$C(\text{SRopt}) = (72 \times 73)/2 = 2628 \quad \text{vs.} \quad \mathbb{E}[C(\text{STRopt})] \geq 2.91 \times 10^4.$$

That is, STRopt required a number of Kalman belief updates that was higher than 10-times the number of updates required in SRopt. The size of the candidate set does not only result in a larger number of Kalman updates. Additionally, this number also strongly affects the contribution of the projected gradient descent for the algebraic redundancy check to the computational effort. As can be seen on the right part of Fig. 3.17 the number of required iterations grew up to a median of 2500. As every iteration of the projected gradient descent requires an eigenvalue computation of a matrix in $\mathbb{R}^{n_x+1,n_x+1}$ this part of STRopt actually dominated the overall computational demand. Under the assumption that eigenvalue computation has a similar demand as a Kalman belief update we can add the required numbers of iteration for checking the candidate sets C_t and obtain a total expected demand of

$$\mathbb{E}[C(\text{STRopt})] = 2.91 \times 10^4 + 6.90 \times 10^7. \tag{3.84}$$

Summarized, empirically the number of Kalman belief updates and eigenvalue computations required in STRopt is approximately a factor 10^4 of the number of Kalman belief updates that are conducted in SRopt. Hence, not restricting the sensor model of the joint task POMDP results in a tremendous increase in computation demand which may be in-feasible for a real-time distraction warning system.

CPU Times

To validate the analyis of computational demand a comparison of CPU times was conducted. This was done using MATLAB implementations of SRopt, of STRopt using the Projected Gradient descent for algebraic redundancy check STRopt + PG and of STRop using CVX [74] to solve 3.61 denote as STRopt + CVX. In this context, we considered CVX as alternative solver for the redundancy check to verify the efficiency of PG for solution. Note that the solvers underlying CVX have also been used in [246].

We report CPU times from a machine with a i7-3740 QM, 2.70 GHz processor and 16.0 GB RAM. For the purpose of this evaluation all high level parallelism was disabled. The results of the evaluation are depicted in Fig. 3.18 and its main statistics are summarized in Tab. 3.3.

Tabular 3.3: Statistics of the CPU time for the Solution approaches for SRopt and STRopt

Methods	**Statistics**		
	0.25 Quantile	0.50 Quantile	0.75 Quantile
SRopt	0.83×10^0 s	0.88×10^0 s	0.95×10^0 s
SRopt + PG	0.09×10^4 s	0.48×10^4 s	1.24×10^4 s
SRopt + CVX	0.15×10^4 s	0.38×10^4 s	0.90×10^4 s

A signed rank test on the median CPU times could not establish significant differences between STRopt + PG and STRopt + CVX $p_{\text{test}} = 0.25$. Notably, the 0.25 quantile of the CPU time using PG was significantly smaller that the quantile using CVX. Which was verified by the test of [86] $p_{\text{test}} < 0.01$.

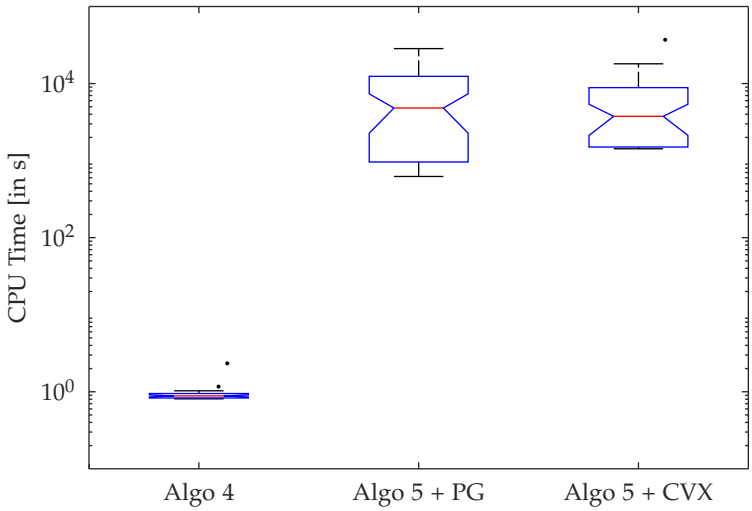

Figure 3.18: Results of the evaluation of the CPU times of the algorithms for sensor control policy computation. The distribution of CPU times is depicted by means box plots. The red line indicates the median CPU time. The blue box denotes the $[0.25, 0.75]$ confidence interval, while whiskers indicate $1.5\times$ the median to quantile distance. SRopt is the approach for computing optimal sensor control policies under restriction of the sensor model. STRopt + PG is the solution approach for SLQGs using the projected gradient method and STRopt + CVX is the solution approach for SLQGs using CVX for the algebraic redundancy check.

Corresponding to the previous analysis of the algorithmic demand of SRopt and STRopt marked differences of the required CPU times have been found. Specifically, the CPU times of STRopt were of almost the same factor of 10^4 higher than those of SRopt. Furthermore, comparison with CVX showed that PG is a competitive approach for solving the redundancy check despite the large number of required iterations. Of course, the CPU times can significantly be reduced by implementing the algorithms in more efficient programming languages. As shown later, SRopt can further be optimized and obtains a CPU time of 0.02s using a fixed size C implementation. However, CVX is not available in C or C++ and importantly does not feature solver optimization for fixed sizes. For these reasons, the evaluation was conducted using the MATLAB implementations. However, based on the analysis of computational demand we expect the ratios of CPU times to be similar in any other programming language. Finally, the analysis of the 0.25 quantile of the CPU times STRopt + PG and STRopt + CVX quantile showed that a quarter of the redundancy checks can be solved in half of the time using PG. We think that this was attributed to the simple step-size rule of 1 used in the experiment. Hence, we expect significant speed-up using more elaborate step-size rules and acceleration techniques which e.g. reviewed in [172].

3.6.3 Discussion

Previously, approaches for computing rational gaze switch policies have been derived which served to implement appropriate glance behavior. This section, presented an evaluation of the methods from the perspective of usage in a real-time distraction warning system. The results showed, the effects of the choice of a quadratic reward of the states related to vehicle control are not in conflict with driver behavior in real traffic. Furthermore, also restricting the sensor model does not produce unrealistic artifacts.

With respect to computational demands, the evaluation shows tremendously increased demand and CPU time of computing rational policies without sensor model restriction (STRopt) compared to solution with sensor model restriction (SRopt). We conclude that restricting the sensor model is the most practicable approach for computing appropriate glance behavior in a distraction warning system. This is because it obtains sufficiently realistic glance behavior under feasible computational demands.

Furthermore, it possesses the important advantage of allowing to compute maximum causal entropy policies which can more accurately take into account realistic driver behavior in definition of appropriate glance behavior.

3.7 Conclusion

This chapter addressed the problem of specifying appropriate glance behavior in secondary task interaction in the driving task of lane keeping. We approached this by means of rational policies in a joint task POMDP. Here, a kinematic model of the driving task was developed that incorporated external parameters which represented the possible variations of the driving situation. Furthermore, explicit models of the driver's sensing of the driving situation and the potential distracting secondary task were presented. Possible approaches for obtaining optimal policies, SRopt (Algo. 4) and STRopt (Algo. 6) as well as rational policies SRMCE (Algo. 7) were discussed thereafter. Given the obtained optimal or the maximum causal entropy policy of the joint task, glance behavior whose Eyes-Off Duration (EOD) did not significantly exceed those produced by the rational policies was defined as appropriate. The exact approaches for computing gaze switch policies were finally evaluated. Here, sufficient realism with respect to driver behavior in real traffic could be verified for all approaches. However, the investigation of the computational demands revealed that only the variants SRopt and SRMCE can be applied in a distraction warning system. We concluded that it is most practical to restrict the sensor model of the joint task POMDP.

As a result of the evaluation of the different variants of the joint task POMDP and the different associated policies we can conclude: Exact and sufficiently fast policy computation is only possible under the assumption that the driver immediately obtains all available information when returning his or her gaze to the road. Although, we also discussed approximate techniques and heuristics, exact computation under clear restrictions has some advantages for defining situationally appropriate glance behavior. This is because, approximate or heuristic methods usually cannot guarantee a bounded loss of performance in terms of reward. This bears the risk that the performance of the computed policy may drop in specific driving situations what conflicts with the goal of situation specific assessment of driver behavior. In contrast, for the optimal glance policy under the simple secondary task model we can establish a uniform bound on loss of primary task performance that depends only on the reward parameters and the time horizon.

Considering, related work on POMDP models of human glance behavior it is important to note that although some work considered more complex models, e.g. [173, 57, 23], in neither of these exact solution techniques could be applied. Therefore, we must expect that extending the approach of this work to driving tasks other than lane keeping will require considering approximate and heuristic methods. As noted in the discussion of inexact methods in this chapter here exhaustive evaluation of performance in the different driving scenarios is necessary.

Finally, this chapter derived the joint task POMDP but left open several important parameters. Specifically, we neither specified the parameters of the introduced reward functions nor did we present concrete values for the quantities involved in the model of the driver's sensor characteristics. Nevertheless, these parameters are of crucial importance as they define the compute appropriate glance behavior. In previous work the values of these parameters were hand-tuned which is not satisfying. Instead, we present sophisticate procedures to obtain these parameters from data of experienced drivers in Cpt. 4 and Cpt. 5.

4 Inferring Driver's Policy and Reward

Computation of appropriate glance behavior has to consider the driver's potentially sub-optimal behavior. Furthermore, a suitable parametrization of the individual components of the reward function is needed. This chapter addresses how both reward parameters and the drivers' policy can automatically be determined in the context of the normative model of glance behavior. We first review frameworks of inverse optimal control that allow to estimate those quantities from behavioral data in Sec. 4.3. Here, methods for inference under optimal policies as well as two variants for inference under the maximum causal entropy policy model are considered. Thereafter, the corresponding algorithmic approaches for the joint task POMDP introduced previously are derived in Sec. 4.4. Sec. 4.5 reports on a first evaluation of both MCE approaches and a baseline technique using simulated data. Thereafter, we introduce a new data set of driver behavior in Sec. 4.6. This was obtained by recording several participants during engagement in a typing task while driving on a public motorway. Finally, Sec. 4.7 studies the performance of the behavior models obtained using inverse optimal control methodology and two different baselines approaches.

This chapter has previously been published in the works [203, 202].

4.1 Introduction

In the previous chapter, a mathematical definition of appropriate glance behavior was developed in form of a normative model of glance behavior. This model features a POMDP model of the joint task of driving and the engagement in a secondary task. Given a reward function that considers both tasks the POMDP model can be solved for an optimal or MCE gaze switch policy. Implementing our mathematical definition in Sec. 3.4 these policies specify appropriate glance behavior which is ultimately used in a distraction warning system (see Sec. 6.3). Previously, the task dynamics and the sensor model of the POMDP model as well as the reward *terms* have been introduced. However, the reward *parameters* are yet unknown and cannot simply be adapted from the literature. This is because our POMDP model differs from the models of other authors. In addition to that, in most previous works the applied reward parameters were given without a motivation or derivation. Nevertheless, those parameters are of crucial importance. They implicitly specify the normative glance behavior and will for example in the context of a distracting warning system explored in Cpt. 6 cause the system to either trigger a warning or suppress a warning. Hence, those parameters must carefully be set to ensure the effectiveness and the usefulness of such a system.

In addition to the reward parameters, the implementation of the definition of appropriate glance behavior requires a model of the driver's policy. When using this definition in a distraction warning system we can only hope to improve the driver's policy for *returning* gaze to the road (see Sec. 6). Consequently, we have to consider the potential sub-optimal policies for vehicle control, secondary task interaction as well as *averting* gaze from the road. For example, the durations of glance off the road that can be tolerated in the model of appropriate glance behavior strongly depend on how well drivers can control the vehicle and how frequent steering errors occur. Furthermore, the viewing time subsequent of gaze aversion influences appropriate glance behavior. This is because in this period the driver can effectively correct his or her lane position which has a strong influence on the overall driving performance. Summarized it is required to include a realistic model of the policies of real drivers into our normative model of glance behavior. If optimal or MCE policies are employed to model the driver's policies these are also defined by reward parameters (see Sec. 3.5.2 and Sec. 3.5.3). That is specifying such policy models also corresponds to finding suitable reward parameters. In this context, especially the MCE policy promises to realistically consider driver behavior as it allows stochastic and sub-optimal behavior. Of course, this requires to additionally estimate the spread of the policy distribution.

In human factors and cognitive science research, as previously reviewed in Sec. 1.1, many works have found rational adaption of human behavior. For example, drivers deliberately chose maximum occlusion times that follow a strong monotonic decrease with the driving speed [207, 67]. Consequently, optimal control has been proposed as a suitable model for manual control performed by well trained and well-motivated humans [19]. This was verified by a close match of experimental data in numerous works [149, 26, 240, 42]. Hence, we could try to obtain suitable reward parameters by minimizing the difference between the optimal policy and behavioral data. This approach is commonly termed as *Inverse Optimal Control* (IOC). Conversely to optimal control where one aims for the optimal policy with respect to a reward, one seeks to find a reward for which the policy underlying the data is optimal in inverse optimal control. However, not all drivers will show optimal behavior. Indeed, this is the reason for distraction related crashes. Nevertheless, IOC can be applicable when data is obtained from experienced and carefully instructed drivers. Furthermore, it turns out that IOC is also possible under the maximum causal entropy framework which allows to infer rewards from potentially sub-optimal behavior. In addition to that IOC in the maximum causal entropy model also estimates the amount of variation in the underlying policy. This quantity can be used to take into account the variations in realistic driver behavior in the definition of appropriate glance behavior.

4.2 Related Work

The problem of estimating policies from behavioral data is well known in the literature. The classic approach to this is to use regression techniques to fit a mapping from states to controls on those pairs present in the data. This approach is referred to as behavioral cloning [16] or *Direct Policy Estimation* (DPE) [200]. A variety of methods are available for this purpose as e.g. discussed in [148]. In the context of estimating driver policies these techniques have for example been used to estimate models of steering behavior [197, 78] and models of glance behavior [95]. However, difficulties in the application of DPE arise for states that are not contained in the collected data. In addition, changes in the control scenario can result in adapted behavior e.g. adapting deliberately chosen occlusion times to vehicle speed [207, 67]. Consequently, when using DPE on a sub-set of all possible scenarios there is a risk that the prediction accuracy of the obtained policies deteriorates in previously unseen scenarios, so-called *transfer scenarios*. Otherwise, estimating policies using DPE on several distinct scenarios can wash out the specific differences in behavior. Nevertheless, due to its frequent use in previous approaches for modeling driver behavior DPE is a suitable baseline for evaluating the predictive performance obtained by inverse optimal control application.

Inverse optimal control provides an alternative to DPE and is a natural way to fit the normative model to data. IOC in linear quadratic regulation problems has already been addressed in Kalman's work [101]. In the last two decades, several general concepts for inverse optimal control and Inverse Reinforcement Learning (IRL)[1] in Markov Decision Processes (MPDs) have been proposed [167, 2, 185, 234, 166, 182, 258, 194] and a survey is provided in [66]. These methods are of a similar architecture and cycle between computing a rational policy given an iterate of the reward parameters and an update of the reward parameters. Of these general approaches [167, 2, 185, 234] address IOC/IRL with respect to optimal policies, [166, 182, 194] address IOC/IRL with respect to the Boltzman policy model. The maximum causal entropy policy model was first proposed in the context of IOC/IRL in [258]. Furthermore, several IOC/IRL frameworks have also been extended to general POMDPs [40]. In these cases, however, the computational burden required for obtaining rational policies is a great issue. This is not the case for LQGs were computational efficient IOC is possible in case of the MCE policy [38].

Simulated driving has been used to benchmark IOC/IRL algorithms [2]. Driver's navigation strategies have been inferred by IOC to predict turns [259] and the range map of an electric vehicle [169]. Furthermore, IOC has been applied in the context of driver assistance for risk aware choice of driving speed [211]. Finally, IOC has also been used to obtain socially compliant or individualized trajectories of autonomous vehicles [72, 117, 195]. However, none of these works considered estimating policies in the context of driver's manual sensori-motor control of the vehicle.

[1] This is reward estimation using a reinforcement learning approach (see Sec. 2.1) to obtain a rational policy

IRL/IOC for estimating policies in manual airplane control was addressed in [176]. In that work the class of optimal control of LQGs in the infinite horizon setting was employed. Finally, [192] is, to the best of our knowledge, the single previous work where IRL/IOC was used to estimate the rewards underlying gaze switching policies. Here, walkway navigation simulated in a virtual reality environment was considered. However, in this case only an approximate solution was obtained using the arbitration heuristic of [220, 191] for computing policies. We have already discussed the disadvantages of using heuristic approaches for definition of appropriate glance behavior. In estimation of rewards using IOC the usage of approximation techniques is also problematic. When scenario-specific approximate optimal control is used in iteration, IOC can fail to estimate reward models that are transferable.

In this chapter, we derive *exact* approaches to implement the IOC frameworks of [234] and [258] for the POMDP class of the normative model of glance behavior. Here, we extend previous work on IOC in LQGs [38]. In this context, we also discuss important issues related to the fact that human sensory measurements are rarely available in the data of real world experiments that have not been noted in previous work. To the best of our knowledge, we conduct the first comparison of the classic approach to MCE IOC and the maximum causal likelihood variant, which has alternatively been used in some work [29, 77]. This is done in both simulation and on driver data obtained in real traffic. Finally, we evaluated the prediction performance of the MCE policy given the estimated rewards in comparison to DPE of a generic regression baseline as well as DPE of the established models of [197] and [95].

4.3 Inverse Optimal Control

Inverse optimal control seeks to reconstruct the reward model $r(x_t, u_t)$ underlying observed rational behavior. Here, the behavioral data D is given by several trajectories $\mathsf{D} = \{(\mathbf{u}_{t=0:T}, \mathbf{x}_{t=0:T})^{i=1:n}\}$ produced by an *unknown* policy $\boldsymbol{\pi}_{0:T}$ under *known* initial state p_0 and process model $\boldsymbol{\mathcal{P}}_{0:T}$.

In the following we will first review the IOC frameworks of [234, 258] before we derive IOC approaches for the POMDP class of the normative model of glance behavior in a second step.

4.3.1 Syed's Game-Theoretic Inverse Optimal Control

For deriving IOC/IRL frameworks first consider the following: For any reward function $r(x_t, u_t)$ and its associated optimal policy $\boldsymbol{\pi}^{\star}_{0:T}$ the expected return of the optimal policy is greater or equal the expected return of any other policy $\boldsymbol{\pi}_{0:T}$

$$\mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}^{\star}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big] \geq \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big]. \tag{4.1}$$

Define the *gap* $g(r, \boldsymbol{\pi}_{0:T})$ as the difference of the expected returns of the optimal policy $\boldsymbol{\pi}^{\star,r}_{0:T}$ for reward function r and the expected returns of the policy $\boldsymbol{\pi}_{0:T}$,

$$g(r, \boldsymbol{\pi}_{0:T}) = \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}^{\star,r}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big] - \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big]. \tag{4.2}$$

The gap is non-negative and zero only if $\boldsymbol{\pi}_{0:T}$ is optimal for the reward $r(x_t, u_t)$. Hence, it can serve as a suitable measure of the fit of the reward function $r(x_t, u_t)$ to the policy $\boldsymbol{\pi}_{0:T}$. Consequently, the rewards underlying an assumedly optimal policy can be inferred by trying to minimize the gap

$$\min_{r\in\mathsf{R}} g(r, \boldsymbol{\pi}_{0:T}) = \min_{r\in\mathsf{R}} \Big(\mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}^{\star,r}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big] - \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t)\Big|\boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big]\Big). \tag{4.3}$$

In this context, the reward function is constrained to be in a set R that does not contain the zero function. This restriction ensures that the minimization problem is not trivially solved because a zero function will make any policy optimal.

If only samples D of the underlying policy $\boldsymbol{\pi}_{0:T}$ are available, we can instead try to minimize the *empirical* gap

$$\min_r g(r, \mathsf{D}) = \min_{r \in \mathsf{R}} \Big(\mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t) \Big| \boldsymbol{\pi}_{0:T}^{\star,r}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \Big] - \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t) \Big| \mathsf{D} \Big] \Big), \tag{4.4}$$

by replacing the unknown expected return of $\boldsymbol{\pi}_{0:T}$ in (4.3) by the empirical expectation over the samples. Note, the transitions underlying the sample data could have been "beneficial" with respect to the reward and resulted in an empirical return greater than the expectation. Therefore, the empirical gap can be negative and is generally only asymptotically non-negative. It can be proven by recursion [231, 181] that the initial value function $V_0^{\star,r} = \mathbb{E}[V_0^{\star,r}(x_0)|p_0(x_0)]$ for reward function r is given by the expected return of the optimal policy

$$V_0^{\star,r} = \mathbb{E}[V_0^{\star,r}(x_0)|p_0(x_0)] = \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t) \Big| \boldsymbol{\pi}_{0:T}^{\star,r}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \Big]. \tag{4.5}$$

Hence, the previous optimization problem (4.4) can alternatively be written as

$$\min_r g(r, \mathsf{D}) = \min_{r \in \mathsf{R}} \Big(V_0^{\star,r} - \mathbb{E}\Big[\sum_{t=0}^{T} r(x_t, u_t) \Big| \mathsf{D} \Big] \Big). \tag{4.6}$$

Linear Parameterization of Reward Function In case of a linear parametrization of the reward function $r(x_t, u_t) = \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t)$, $\boldsymbol{\theta} \in \Theta$ the optimization problem (4.3) results in

$$\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta} \Big(V_0^{\star,\boldsymbol{\theta}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big) \tag{4.7}$$

$$= \min_{\boldsymbol{\theta} \in \Theta} \Big(\mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \boldsymbol{\pi}_{0:T}^{\star,\boldsymbol{\theta}}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \Big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big) \tag{4.8}$$

$$= \min_{\boldsymbol{\theta} \in \Theta} \Big(\max_{\boldsymbol{\pi}_{0:T}} \Big[\mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \big| \boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \big] \Big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big). \tag{4.9}$$

Neglecting the constraint $\boldsymbol{\theta} \in \Theta$, the mini-max problem (4.9) is convex in $\boldsymbol{\theta}$ and its (sub)-gradient is given by

$$\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathsf{D}) = \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \boldsymbol{\pi}_{0:T}^{\star,\boldsymbol{\theta}}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \Big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] = \nabla_{\boldsymbol{\theta}} V_0^{\star,\boldsymbol{\theta}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big], \tag{4.10}$$

by means of standard convex analysis [32, 190]. Here, the gradient $\nabla_{\boldsymbol{\theta}} V_0^{\star,\boldsymbol{\theta}}$ can be computed in recursive fashion as presented in [166]. This is possible, because first it holds

$$\nabla_{\boldsymbol{\theta}} V_t^{\star,\boldsymbol{\theta}}(x_t) = \nabla_{\boldsymbol{\theta}} \max_{u_t} \big(Q_t^{\star,\boldsymbol{\theta}}(x_t, u_t) \big) = \nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(x_t, \pi_t^{\star,\boldsymbol{\theta}}(x_t)). \tag{4.11}$$

Second, the gradients of the state-control function $\nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(x_t, u_t)$ are given by

$$\nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(x_t, u_t) = \nabla_{\boldsymbol{\theta}} \Big(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) + \mathbb{E}\big[V_t^{\star,\boldsymbol{\theta}}(x_{t+1}) \big| \mathcal{P}(x_{t+1}|x_t, u_t) \big] \Big) \tag{4.12}$$

$$= \boldsymbol{\varphi}(x_t, u_t) + \mathbb{E}\big[\nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(x_{t+1}, \pi_{t+1}^{\star,\boldsymbol{\theta}}(x_{t+1})) \big| \mathcal{P}(x_{t+1}|x_t, u_t) \big]. \tag{4.13}$$

Consequently, evaluating (4.11) and (4.12) can be used to compute the gradient $\nabla_{\boldsymbol{\theta}} V_0^{\star,\boldsymbol{\theta}}$, which is used to form the gradient of the gap $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathsf{D})$.

Assuming Θ is convex and admits efficient projection, we can finally solve the minimization problem of (4.7) by cycling between computing the gradient and a projection step. As another instance of proximal gradient descent (compare to Algo. 5) this is guaranteed to converge to the global optimal solution $\boldsymbol{\theta}^\star$ of the minimization problem (4.7) [172]. Note, that this approach is largely the solution technique proposed in [234].

Illustrative Example We illustrate the principle of minimizing the gap with a simplified version of the joint task POMDP in Fig. 4.1.

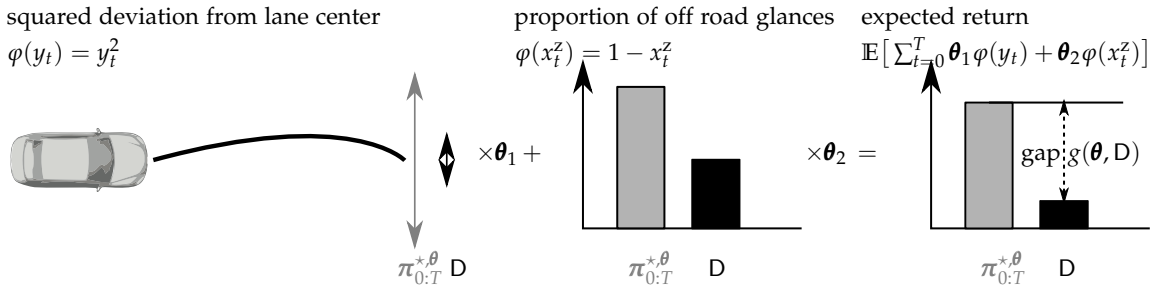

Figure 4.1: Illustration of the principle of minimization of the gap in inverse optimal control. Assume that data of a driver D denoted in black and an optimal policy $\pi_{0:T}^{\star,\boldsymbol{\theta}}$ for candidate parameters $\boldsymbol{\theta}$ denoted in gray are given. Furthermore, assume that the reward parameters are such that long glances off the road are optimal $-1 \ll \boldsymbol{\theta}_1 < 0$, $\boldsymbol{\theta}_2 = 1$. This leads to a high squared deviation from the lane center y_t^2 under the optimal policy. Adding up the expected squared deviation from the lane center scaled by negative factor of small magnitude with the proportion of gaze off the road, the optimal policy obtains high expected return. In contrast, the observed driver showed a significantly smaller proportion of gaze off the road and improved lane position. Under the candidate reward parameters this leads to small empirical return. Consequently, a significant gap $g(\boldsymbol{\theta}, \mathsf{D})$ is present between the return of the optimal policy and the empirical return obtained by the driver. This indicates that the candidate parameters $\boldsymbol{\theta}$ are not the ones that the observed driver behavior is optimal for.

4.3.2 Maximal Causal Entropy Inverse Optimal Control

For the purpose of introducing maximum causal entropy optimal control, we first return to the gradient of the gap $\nabla_{\boldsymbol{\theta}}\, g(\boldsymbol{\theta}, \mathsf{D})$. If it holds true

$$\mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \boldsymbol{\pi}_{0:T}^{\star,\boldsymbol{\theta}}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\Big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D}\Big] = 0, \tag{4.14}$$

for any feasible $\boldsymbol{\theta}$ then the sufficient conditions for a minimizer of the gap are fulfilled due to the convexity of the problem. The role of this so-called *feature-matching* as a sufficient condition in IOC/IRL has first been discovered in [2]. However, this condition is not a *necessary* condition for a minimizer of the gap. This is because the gap $g(\boldsymbol{\theta}, \mathsf{D})$ is not differentiable as it involves the non-differentiable $\max(.)$ function. Hence, convex optimization tells that even the minimizer $\boldsymbol{\theta}^\star$ may not attain zero gradient. For this reason, [2] suggested to mix optimal policies of different reward parameters $\boldsymbol{\theta}$ to obtain feature-matching.

Formal Definition of Maximal Causal Entropy Inverse Optimal Control As a more sophisticated alternative to mixing optimal policies, Maximum Causal Entropy Inverse Optimal Control (MCE-IOC)

was proposed in [258]. Here, a single stochastic policy $\tilde{\boldsymbol{\pi}}_{0:T}$ matching the empirical feature expectation is obtained by solving the maximization problem

$$\max_{\boldsymbol{\pi}_{0:T}} \mathcal{H}(\boldsymbol{\pi}_{0:T}) = -\sum_{t=0}^{T} \mathbb{E}\Big[\int \pi(u_t|x_t) \log \pi(u_t|x_t) \,\mathrm{d}\, u_t \Big| \boldsymbol{\pi}_{0:t-1}, \boldsymbol{\mathcal{P}}_{0:t-1}, p_0 \Big] \tag{4.15}$$

$$\text{s.t.} \quad \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \boldsymbol{\pi}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \Big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] = 0. \tag{4.16}$$

That is, one seeks to find the policy of maximal stochasticity, i.e. entropy, that fulfills the sufficient condition. This is a well-defined problem except for the case that there is not a single policy that can obtain feature matching. That is, if the optimization problem is infeasible.

In Cpt. 2 on the mathematical background of this thesis the maximum causal entropy policy was introduced in a different fashion (see (2.55)). The relation to the definition (4.15) is given in the following. The maximization problem (4.15) can be solved by considering the Lagrangian saddle point problem

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\pi}_{0:T}} \Big(\mathcal{H}(\boldsymbol{\pi}_{0:T}) + \boldsymbol{\theta}^\top \Big[\mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \big| \boldsymbol{\pi}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \big] - \mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \big| \mathsf{D} \big] \Big] \Big) \tag{4.17}$$

$$= \min_{\boldsymbol{\theta}} \Big(\underbrace{\max_{\boldsymbol{\pi}_{0:T}} \Big[\mathcal{H}(\boldsymbol{\pi}_{0:T}) + \mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \big| \boldsymbol{\pi}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \big] \Big]}_{\max_{\boldsymbol{\pi}_{0:T}} \left(\mathcal{H}(\boldsymbol{\pi}_{0:T}) + \mathbb{E}\left[\sum_{t=0}^{T} r(x_t, u_t) \middle| \boldsymbol{\pi}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \right] \right)} - \mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \big| \mathsf{D} \big] \Big) \tag{4.18}$$

with respect to the policy $\boldsymbol{\pi}_{0:T}$ and the Lagrangian multipliers $\boldsymbol{\theta}$ [32, 190]. Here the previous definition (2.55) is found in the inner maximization problem considering that the reward function was parametrized according to $r(x_t, u_t) = \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t)$.

In [257] the following key properties are proven: First, it holds true for the soft value-function $\tilde{V}_0^{\boldsymbol{\theta}}$ for $t = 0$ under the reward function $r(x_t, u_t) = \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t)$

$$\max_{\boldsymbol{\pi}_{0:T}} \Big(\mathcal{H}(\boldsymbol{\pi}_{0:T}) + \mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \big| \boldsymbol{\pi}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \big] \Big) = \tilde{V}_0^{\boldsymbol{\theta}}. \tag{4.19}$$

Second, in case $\tilde{V}_0^{\boldsymbol{\theta}}$ is finite, it is a differentiable function of $\boldsymbol{\theta}$ and its gradient is given as

$$\nabla_{\boldsymbol{\theta}} \tilde{V}_0^{\boldsymbol{\theta}} = \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \tilde{\boldsymbol{\pi}}_{0:T}^{\boldsymbol{\theta}}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \Big], \tag{4.20}$$

where $\tilde{\boldsymbol{\pi}}_{0:T}^{\boldsymbol{\theta}}$ is the maximum causal entropy policy obtainable by the iterations (2.58) and (2.57). Third and finally the minimization wrt. $\boldsymbol{\theta}$ is a convex optimization problem. Hence, any local minimizer is also global minimizer. Consequently, we can obtain suitable reward parameters for the maximum causal entropy policy model by minimizing, what we call the *soft gap* $\tilde{g}(\boldsymbol{\theta}, \mathsf{D})$,

$$\min_{\boldsymbol{\theta}} \tilde{g}(\boldsymbol{\theta}, \mathsf{D}) = \min_{\boldsymbol{\theta}} \Big(\tilde{V}_0^{\boldsymbol{\theta}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big) \tag{4.21}$$

$$= \min_{\boldsymbol{\theta}} \Big(\max_{\boldsymbol{\pi}_{0:T}} \Big[\mathcal{H}(\boldsymbol{\pi}_{0:T}) + \mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \big| \boldsymbol{\pi}, \boldsymbol{\mathcal{P}}_{0:T}, p_0 \big] \Big] - \mathbb{E}\big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \big| \mathsf{D} \big] \Big). \tag{4.22}$$

The maximum causal entropy model has found broad success in application e.g in [259, 211, 169, 117, 195]. This is due to the fact that fitting its reward comes with robust performance guarantees [258]. That is, the MCE policy obtains smallest loss in case the true policy underlying the data is chosen adversarial.

The Maximum Causal Likelihood Variant Considering that the MCE policy is given by

$$\tilde{\pi}_t^{\boldsymbol{\theta}} = \exp\left(\tilde{Q}_t^{\boldsymbol{\theta}}(x_t, u_t) - \tilde{V}_t^{\boldsymbol{\theta}}(x_t)\right),$$

the reward parameters $\boldsymbol{\theta}$ can also be obtained by minimizing the negative log-likelihood of the demonstration data $\mathsf{D} = \{(\mathbf{u}_{t=0:T}, \mathbf{x}_{t=0:T})^{i=1:n}\}$

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T} -\log\left[\exp\left(\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i) - \tilde{V}_t^{\boldsymbol{\theta}}(x_t^i)\right)\right] \tag{4.23}$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\left(\tilde{V}_t^{\boldsymbol{\theta}}(x_t^i) - \tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i)\right). \tag{4.24}$$

This variant was proposed as an alternative to the mini-max problem (4.18) in [257] and was extended in [29, 77]. The minimization problem of (4.23) can be solved by means of the gradient $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ which is given by

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\left(\nabla_{\boldsymbol{\theta}}\tilde{V}_t^{\boldsymbol{\theta}}(x_t^i) - \nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i)\right). \tag{4.25}$$

Similar, as in case of Syed's approach (4.11), (4.12) we can use a recursion technique to obtain the required gradient. First note, that it holds true

$$\nabla_{\boldsymbol{\theta}}\tilde{V}_t^{\boldsymbol{\theta}}(x_t^i) = \nabla_{\boldsymbol{\theta}}\log\int\exp\left(\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t)\right)\mathrm{d}\,u_t = \int\frac{\exp\left(\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t)\right)}{\int\exp\left(\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t')\right)\mathrm{d}\,u_t'}\nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t)\,\mathrm{d}\,u_t \tag{4.26}$$

$$= \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t)\middle|\tilde{\pi}_t^{\boldsymbol{\theta}}(u_t|x_t^i)\right]. \tag{4.27}$$

Furthermore, the gradients $\nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i)$ fulfill the recursive relation

$$\nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i) = \nabla_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^\top\boldsymbol{\varphi}(x_t^i, u_t^i) + \mathbb{E}\left[\tilde{V}_t^{\boldsymbol{\theta}}(x_{t+1})\middle|\mathcal{P}(x_{t+1}|x_t^i, u_t^i)\right]\right) \tag{4.28}$$

$$= \boldsymbol{\varphi}(x_t^i, u_t^i) + \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_{t+1}, u_{t+1})\middle|\tilde{\pi}_t^{\boldsymbol{\theta}}(u_{t+1}|x_{t+1}), \mathcal{P}(x_{t+1}|x_t^i, u_t^i)\right] \tag{4.29}$$

which was first shown in [29]. Hence, we can obtain the gradient $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ by recursively conducting (4.29) and (4.27). The same approach can also be used to compute $\nabla_{\boldsymbol{\theta}}\tilde{V}_0^{\boldsymbol{\theta}} = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\tilde{V}_0^{\boldsymbol{\theta}}(x_0)\middle|p_0(x_0)\right]$ which is required to solve the mini-max problem of the original maximum causal entropy approach to IOC (4.19).

Comparison of Maximum Causal Entropy IOC and Maximum Causal Likelihood IOC When we compare the original MCE approach to the MCL approach, the following interpretations can be made: MCE seeks to minimize the soft gap $\tilde{g}(\boldsymbol{\theta}, \mathsf{D})$ (4.21) which corresponds to finding parameters $\boldsymbol{\theta}$ that result in feature matching (4.14). Hence, this approach focuses on matching the performance of the policy present in the data wrt. to the individual reward features. MCL on the other hand is closer related to direct policy estimation as it seeks to minimize the log-likelihood of the policy given the state-control pairs in the data (4.23). In contrast to DPE the MCE policy model used in MCL results in a significant beneficial inductive bias which improves generalization and transferability.

Although MCE and MCL may estimate different $\boldsymbol{\theta}$ on finite data sets D, notably in [257] is shown that the same $\boldsymbol{\theta}$ is obtained in the infinite sample limit. This is because if D contains infinitely many triples $(x_{t+1}^i, x_t^i, u_t^i)$ the term $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\left(\nabla_{\boldsymbol{\theta}}\tilde{V}_t^{\boldsymbol{\theta}}(x_t^i) - \nabla_{\boldsymbol{\theta}}\tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i)\right)$ is equal to the feature expectation under the policy $\tilde{\boldsymbol{\pi}}_{0:T}^{\boldsymbol{\theta}}$. In practical application, however, one cannot expect that this is the case. Therefore, it is necessary to empirically investigate the differences between MCE and MCL. Before we do so, let us return to the POMDP model of the joint task and derive the IOC algorithms for this problem class.

4.4 Inverse Optimal Control in the Class of the Joint Task POMDP

Previously, we reviewed two inverse optimal control frameworks for MDPs. We introduced objective functions that can be used to infer reward parameters from behavioral data and presented recursions for gradient computation. Similar as in case of the Bellman equation introduced earlier (see Cpt. 2) the gradient recursion is computational intractable in many MDPs and POMDPs. In contrast, the class of POMDPs that are used in the normative model of glance behavior admit exact and efficient computation. This will be shown in the present section.

Syed's inverse optimal control framework and the maximum causal entropy framework have a very similar architecture. In both cases the estimation is posed as an optimization problem (4.6), (4.18) which requires to compare the (soft) value function $\tilde{V}_0^{\boldsymbol{\theta}}, V_0^{\star,\boldsymbol{\theta}}$ to the empirical return of the data D. To obtain the minimizer in both Syed's IOC and MCE IOC gradient based techniques can be used. Here, the gradients can be obtained by evaluating recursion (4.12), (4.11) for the optimal policy or recursion (4.29), (4.27) for the MCE policy given the current parameter iterate $\boldsymbol{\theta}$. Consequently, we derive IOC approaches for the optimal and MCE policy in the POMDP class in a unified view considering its (soft) Bellman equations.

4.4.1 Posing IOC

To obtain algorithms for IOC in the POMDP relevant for the normative model of glance behavior, first both minimization problems (4.6), (4.18) must be posed with regards to a specific class. For this purpose, consider the reward functions that are involved in the POMDP model. These must be rewritten in a linear parameterization with respect to a parameter $\boldsymbol{\theta} := [\mathrm{vec}(\boldsymbol{\Theta}_1); \mathrm{vec}(\boldsymbol{\Theta}_2); \theta_3; \boldsymbol{\theta}_4]$. The reward model used for the task of vehicle control introduced in Sec. 3.3.1 can more generally be written as

$$r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) = -\mathbf{x}_t^{\mathrm{p}\,\top}\mathbf{C}_x\mathbf{x}_t^{\mathrm{p}} - u_t^{\mathrm{p}\,\top}\mathbf{C}_u u_t^{\mathrm{p}} = \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}}\mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}). \tag{4.30}$$

Here the matrices $-\mathbf{C}_x, -\mathbf{C}_u$ are replaced by reward parameters $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$. Furthermore, the reward function of the sensor control (see Sec. 3.3.2) is given by

$$r(u_t^{\mathrm{z}}) = \theta_3 u_t^{\mathrm{z}}, \tag{4.31}$$

and the reward function of the secondary task (see Sec. 3.3.3) is given as

$$r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) = \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}). \tag{4.32}$$

Next, we review the (soft) Bellman equation derived in Cpt. 3 under the linear parametrization. Under a given initial covariance $\boldsymbol{\Sigma}_0^{\mathrm{p}}$ (compare to Sec. 3.5.2) and the optimal policy we obtain

$$Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{Q^\star,\boldsymbol{\theta}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{Q^\star,\boldsymbol{\theta}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{Q^\star,\boldsymbol{\theta},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \tag{4.33}$$

$$V_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{V^\star,\boldsymbol{\theta}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{V^\star,\boldsymbol{\theta}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{V^\star,\boldsymbol{\theta},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}), \tag{4.34}$$

$$\mathbf{M}_t^{Q^\star,\boldsymbol{\theta}} = \begin{cases} [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star,\boldsymbol{\theta}} [\mathbf{A}_t\ \mathbf{B}_t] + \mathrm{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) & \text{if } t < T \\ \mathrm{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) & \text{else} \end{cases} \tag{4.35}$$

$$\mathbf{m}_t^{Q^\star,\boldsymbol{\theta}} = \begin{cases} 2[\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^\star,\boldsymbol{\theta}} \mathbf{a}_t + [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{V^\star,\boldsymbol{\theta}} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \tag{4.36}$$

$$m_t^{Q^{\star,\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \\ \quad + \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})) \\ \quad + \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\right)\right) \\ \quad + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \\ \quad + \mathbb{E}\left[m_{t+1}^{V^{\star,\boldsymbol{\theta}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right] & \text{if } t < T \\ \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:T})) + \theta_3 u_T^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}}) & \text{else} \end{cases}, \tag{4.37}$$

$$\mathbf{M}_t^{V^{\star,\boldsymbol{\theta}}} = \mathbf{M}_{t\ x,x}^{Q^{\star,\boldsymbol{\theta}}} - \mathbf{M}_{t\ x,u}^{Q^{\star,\boldsymbol{\theta}}} [\mathbf{M}_{t\ u,u}^{Q^{\star,\boldsymbol{\theta}}}]^{-1} \mathbf{M}_{t\ u,x}^{Q^{\star,\boldsymbol{\theta}}} \tag{4.38}$$

$$\mathbf{m}_t^{V^{\star,\boldsymbol{\theta}}} = \mathbf{m}_{t\ x}^{Q^{\star,\boldsymbol{\theta}}} - \mathbf{M}_{t\ x,u}^{Q^{\star,\boldsymbol{\theta}}} [\mathbf{M}_{t\ u,u}^{Q^{\star,\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{Q^{\star,\boldsymbol{\theta}}} \tag{4.39}$$

$$\begin{aligned} m_t^{V^{\star,\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = &-\frac{1}{4}[\mathbf{m}_{t\ u}^{Q^{\star,\boldsymbol{\theta}}}]^\top [\mathbf{M}_{t\ u,u}^{Q^{\star,\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{Q^{\star,\boldsymbol{\theta}}} + \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})) + \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x}\right)\right) \\ &+ \max_{u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \Big(\theta_3 u_t^{\mathrm{z}} - \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\right) \\ &+ \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) + \mathbb{E}\left[m_{t+1}^{V^{\star,\boldsymbol{\theta}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]\Big). \end{aligned} \tag{4.40}$$

Whereas, the soft Bellman equations defining the maximum causal entropy policy are given by

$$\tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \tag{4.41}$$

$$\tilde{V}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}), \tag{4.42}$$

In this context the variables $\mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}}, \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}, m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}, \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}}, \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}}$ are given by equations analogous to the equations (4.35)-(4.39) of their optimal policy counterparts $\mathbf{M}_t^{Q^{\star,\boldsymbol{\theta}}}, \mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}}, m_t^{Q^{\star,\boldsymbol{\theta}},1}, \mathbf{M}_t^{V^{\star,\boldsymbol{\theta}}}, \mathbf{m}_t^{V^{\star,\boldsymbol{\theta}}}$. In contrast to that, $m_t^{\tilde{V}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}})$ is given by

$$\begin{aligned} m_t^{\tilde{V}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = &-\frac{1}{4}[\mathbf{m}_{t\ u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^\top [\mathbf{M}_{t\ u,u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{\tilde{Q}^{\boldsymbol{\theta}}} + \frac{1}{2}\log(\det(\pi[\mathbf{M}_{t\ u,u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^{-1})) \\ &+ \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})) + \mathrm{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}\left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x}\right)\right) \\ &+ \mathrm{softmax}_{u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \Big(\theta_3 u_t^{\mathrm{z}} - \mathrm{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\right) \\ &+ \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) + \mathbb{E}\left[m_{t+1}^{\tilde{V}^{\boldsymbol{\theta}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]\Big). \end{aligned} \tag{4.43}$$

Consequently, $\boldsymbol{\theta}$ can be obtained in the problem class of our POMDP model by solving the optimization problem of

$$\begin{aligned} &\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathsf{D}) = \\ &\min_{\boldsymbol{\theta}} \Big([\boldsymbol{\mu}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + \mathbf{m}_0^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + m_t^{V^{\star,\boldsymbol{\theta}},1}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\ &\quad - \mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D}\Big]\Big) \end{aligned} \tag{4.44}$$

to implement Syed's framework for inverse optimal control under the optimal policy model.

For the maximum causal entropy approach to IOC the optimization problem

$$\begin{aligned}
&\min_{\boldsymbol{\theta}} \tilde{g}(\boldsymbol{\theta}, \mathsf{D}) = \\
&\min_{\boldsymbol{\theta}} \Big([\boldsymbol{\mu}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + \mathbf{m}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + m_0^{\tilde{V}^{\boldsymbol{\theta}},1}(x_0^z, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}) + \theta_3 u_t^z + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big)
\end{aligned} \tag{4.45}$$

must be solved. Finally, the MCL variant to obtain the reward parameter in the context of the maximum causal entropy policy, is given by the minimization problem

$$\begin{aligned}
&\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \mathsf{D}) = \\
&\min_{\boldsymbol{\theta}} \mathbb{E}\Big[\sum_{t=0}^{T} \Big([\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta}},1}(\boldsymbol{x^z}_{0:t}, x_t^{\mathrm{i}}) \\
&\qquad - \big[[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}(\boldsymbol{x^z}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^z, u_t^{\mathrm{i}}) \big] \Big) \Big| \mathsf{D} \Big].
\end{aligned} \tag{4.46}$$

4.4.2 The Observed Agent and Its Observer in IOC in POMDPs

Before addressing how gradients of both objectives can be computed let us take a closer look at the terms and variables in the objective function of the IOC approaches (4.44),(4.45) and (4.46).

Apparently, the definition of the (soft) gap involves both the "true" state $\mathbf{x}_t^{\mathrm{p}}$ as well as its expectation $\boldsymbol{\mu}_t^{\mathrm{p}}$ under the belief $b(\mathbf{x}_t^{\mathrm{p}})$ of the agent. This may seem contradictory, because in previous Cpt. 3 we transformed the joint task POMDP into its belief-MDP equivalent to formulate the (soft) Bellman equations. In this process, the unknown "true" state $\mathbf{x}_t^{\mathrm{p}}$ was substituted by the corresponding belief $b(\mathbf{x}_t^{\mathrm{p}})$. Nevertheless, we wish to note that this formulation is exactly the same as in [38].

The reason for the occurrence of both $\mathbf{x}_t^{\mathrm{p}}$ and its expectation $\boldsymbol{\mu}_t^{\mathrm{p}}$ is that IOC assumes a different context than optimal control. When computing rational policies in optimal control we are in the role of the agent that cannot directly access the states but receives sensory measurements of it. In contrast, in inverse optimal control we are in the role of an *external* observer, that observes an agent which acts assumedly rational based on partial information of the states.

In IOC it is commonly assumed that the states and the controls applied by the agent are fully known by the observer [167, 2, 185, 234, 166, 182, 258, 194]. Notably, [107] considered the problem of noisy state and control data in IOC in MDPs. Similar as in the standard approaches to IOC in MDPs, in this thesis the states and the controls of the POMDP model are available in collected behavioral data: This was previously set as an objective for development of the POMDP model in Cpt. 3. In the framework for IOC in POMDPs presented in [40] it was assumed that also the beliefs of the agent or at least its sensory measurements are known by the observer. However, this is definitely not the case in our application: When conducting driving experiments visual measurements made by the driver cannot be recorded. Even worse, the actual sensory measurements made are likely totally different from what we assumed in the crude sensor model used in the joint task POMDP. Consequently, the expected state $\boldsymbol{\mu}_t^{\mathrm{p}}$ according to the belief of the driver must be considered unknown. This aspect will later be especially relevant in the context of inferring sensor models in Cpt. 5.

Fortunately, missing sensory measurements and beliefs in the data do not prevent applying IOC. In [38] MCE IOC in LQGs was applied for modeling mouse cursor movements subject to human delayed sensing of the task states. In this context the issue of missing sensory measurement in application of IOC in POMDPs has not been noted, which is to the best of our knowledge first discussed in this work.

IOC in Syed's framework and in the maximum causal entropy framework is possible for the following reasons: The (soft) value of the belief MDP equivalent of the joint task POMDP (first term of the

IOC objectives (4.44) and (4.45)) is kept. In contrast, the measured empirical return under the true states

$$\mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) + \theta_3 u_t^z + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D}\Big]$$

is used instead of the empirical return in the belief MDP under the unknown beliefs of the observed agent (second term of the IOC objectives (4.44) and (4.45)). This is because the reward function of the belief MDP is $r^b(b(x_t), u_t) = \mathbb{E}[r(x_t, u_t)|b(x_t)]$. Hence, the expected return in the belief MDP equals the expected return under the "true" states x_t

$$\mathbb{E}[\sum_{t=0}^{T} r^b(b(x_t), u_t)|\boldsymbol{\pi}_{0:T}, p^z, \boldsymbol{\mathcal{P}}^b_{0:T}, p^b_0] = \mathbb{E}[\sum_{t=0}^{T} r(x_t, u_t)|\boldsymbol{\pi}_{0:T}, p^z, \boldsymbol{\mathcal{P}}_{0:T}, p_0]. \tag{4.47}$$

For this reason, we use the empirical return under the true states as an unbiased estimator of the empirical return in the belief MDP under the unknown beliefs. Informally, IOC in our POMDP can be described as closing the gap between the return an agent with partial information estimates he can obtain following a rational policy and the empirical return measured by the external observer.

In both (4.44) and (4.45) it is still required to evaluate the (soft) value functions for the initial state $\boldsymbol{\mu}_0^{\mathrm{p}}, x_0^z, x_0^{\mathrm{i}}$. As in the previous case also exact knowledge of $\boldsymbol{\mu}_0^{\mathrm{p}}$ is not needed for inverse optimal control. This is because of the following reason: As a well-known property of the Kalman-Filter it holds $p(\boldsymbol{\mu}_t^{\mathrm{p}}|\mathbf{x}_t^{\mathrm{p}}) = \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}}|\mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p}})$ if the expectation is taken with respect to possible previous sensory measurements $\mathbf{z}_{-\infty:t}$. Therefore, in the role of the external observer we can consider the conditional distribution of $p(\boldsymbol{\mu}_t^{\mathrm{p}}|\mathbf{x}_t^{\mathrm{p}})$ of the unkown $\boldsymbol{\mu}_t^{\mathrm{p}}$ given the known $\mathbf{x}_t^{\mathrm{p}}$. We illustrate the known and unknown quantities of the observed driver and the external observer in Fig. 4.2.

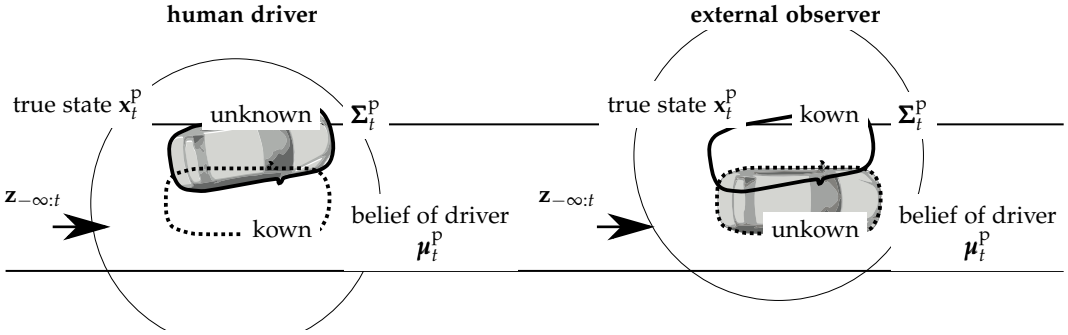

Figure 4.2: Illustration of the known and unknown quantities of the observed driver and the external observer. Arrow denotes the history of sensory measurements $\mathbf{z}_{-\infty:t}$. The slightly rotated car (solid contour) denotes the "true" state $\mathbf{x}_t^{\mathrm{p}}$. The car in the center of the lane denotes (dotted contour) the expected state $\boldsymbol{\mu}_t^{\mathrm{p}}$. In the first case, the pale vehicle denotes the estimated $\boldsymbol{\mu}_t^{\mathrm{p}}$ of the human driver of the true vehicle state $\mathbf{x}_t^{\mathrm{p}}$ indicated by the opaque vehicle. In the second case, the pale vehicle denotes the observer's estimate $\mathbf{x}_t^{\mathrm{p}}$, i.e. the true vehicle state, of the human driver's estimate $\boldsymbol{\mu}_t^{\mathrm{p}}$ indicated by the opaque vehicle. The uncertainty in the corresponding estimation is expressed by the same covariance $\boldsymbol{\Sigma}_t^{\mathrm{p}}$.

Hence, if the expected state $\boldsymbol{\mu}_0^{\mathrm{p}}$ is unknown, minimizing the (soft) gap in Syed's approach to IOC and

in maximum causal entropy IOC can be conducted by considering the expected (soft) gap with respect to the distribution of $\boldsymbol{\mu}_0^{\mathrm{p}}$ given the "true" state $\mathbf{x}_0^{\mathrm{p}}$ and the initial covariance $\boldsymbol{\Sigma}_0^{\mathrm{p}}$

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathsf{D}) \\
&= \min_{\boldsymbol{\theta}} \Big(\mathbb{E}\Big[[\boldsymbol{\mu}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + \mathbf{m}_0^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] \Big| \mathcal{N}(\boldsymbol{\mu}_0^{\mathrm{p}} | \mathbf{x}_0^{\mathrm{p}}, \boldsymbol{\Sigma}_0^{\mathrm{p}}) \Big] + m_0^{V^{\star,\boldsymbol{\theta}},1}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \operatorname{vec}(\boldsymbol{\Theta}_1)^\top \operatorname{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) + \operatorname{vec}(\boldsymbol{\Theta}_2)^\top \operatorname{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big) \\
&= \min_{\boldsymbol{\theta}} \Big([\mathbf{x}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + \operatorname{tr}(\mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} \boldsymbol{\Sigma}_0^{\mathrm{p}}) + \mathbf{m}_0^{V^{\star,\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + m_0^{V^{\star,\boldsymbol{\theta}},1}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \operatorname{vec}(\boldsymbol{\Theta}_1)^\top \operatorname{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) + \operatorname{vec}(\boldsymbol{\Theta}_2)^\top \operatorname{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big)
\end{aligned} \tag{4.48}
$$

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}} \tilde{g}(\boldsymbol{\theta}, \mathsf{D}) \\
&= \min_{\boldsymbol{\theta}} \Big(\mathbb{E}\Big[[\boldsymbol{\mu}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + \mathbf{m}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] \Big| \mathcal{N}(\boldsymbol{\mu}_0^{\mathrm{p}} | \mathbf{x}_0^{\mathrm{p}}, \boldsymbol{\Sigma}_0^{\mathrm{p}}) \Big] + m_0^{\tilde{V}^{\boldsymbol{\theta}},1}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \operatorname{vec}(\boldsymbol{\Theta}_1)^\top \operatorname{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) + \operatorname{vec}(\boldsymbol{\Theta}_2)^\top \operatorname{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big) \\
&= \min_{\boldsymbol{\theta}} \Big([\mathbf{x}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + \operatorname{tr}(\mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_0^{\mathrm{p}}) + \mathbf{m}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + m_0^{\tilde{V}^{\boldsymbol{\theta}},1}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \operatorname{vec}(\boldsymbol{\Theta}_1)^\top \operatorname{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) + \operatorname{vec}(\boldsymbol{\Theta}_2)^\top \operatorname{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big).
\end{aligned} \tag{4.49}
$$

Obviously, the same technique is also applicable for the MCL variant (4.46). Here, the expectations are taken with respect to the true states $\mathbf{x}_t^{\mathrm{p}\,j}$ as well as the covariances $\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{j})$ according to the sensor state sequences $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{j}$ for the states $\mathbf{x}_t^{\mathrm{p}\,j}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{j}$ present in the data D. As a result, we arrive at

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}, \mathsf{D}) \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\Big[\sum_{t=0}^{T} \mathbb{E}\big[[\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) \\
&\qquad - \big([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})) | \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}} | \mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \big] \Big| \mathsf{D} \Big] \\
&= \min_{\boldsymbol{\theta}} \mathbb{E}\Big[\sum_{t=0}^{T} [\mathbf{x}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}] + \operatorname{tr}(\mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) \\
&\qquad - \big([\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \operatorname{tr}(\mathbf{M}_{t\,x,x}^{\tilde{Q}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})) \Big| \mathsf{D} \Big].
\end{aligned} \tag{4.50}
$$

Initial Belief Covariances at Inference Time Previously, we have demonstrated that inverse optimal control is applicable even in absence of the sensory measurements of the observed agent, e.g. the driver. This was possible using the "true" state $\mathbf{x}_0^{\mathrm{p}}$ and the initial covariance $\boldsymbol{\Sigma}_0^{\mathrm{p}}$. While the true state is a measurable quantity contained in the data D, this is not the case for the initial covariance. In principle, $\boldsymbol{\Sigma}_0^{\mathrm{p}}$ depends on the entire sequence of sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{-\infty:0}$ of the observed agent. In practical application to inference of the sensor model underlying gaze switching behavior in driving a heuristic alternative can be applied. We can assume the covariance to quickly converge to a steady state covariance $\hat{\boldsymbol{\Sigma}}^{\mathrm{p}}$ if the driver's gaze is on the road $x_t^z = 0$. A similar assumption was previously made in case of the sensor model restriction Sec. 3.5.2. Given data where the driver's gaze is off the road for $t = 0$, i.e. $x_0^{\mathrm{z}} = 1$, we proceed as follows: We first compute the steady state covariance $\hat{\boldsymbol{\Sigma}}^{\mathrm{p}}$ at the last time step the gaze was on the road, i.e. $t_{\text{gaze aversion}} = \max_{x_t^z=1,\, t<0}(t)$ using the sensor noise

covariance $\mathbf{\Sigma}^{\epsilon^z}(1)$. Thereafter, we conduct the Kalman belief updates for $t = t_{\text{gaze aversion}} : 0$ using the sensor noise covariance $\mathbf{\Sigma}^{\epsilon^z}(0)$ and finally obtain $\mathbf{\Sigma}_0^{\text{p}}$.

4.4.3 Obtaining the Gradients

To obtain a gradient of the (soft) gap, it is effectively required to compute $\nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\text{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\text{i}}, u_t^{\text{p}}, u_t^{\text{z}}, u_t^{\text{i}})$ or $\nabla_{\boldsymbol{\theta}} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\text{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\text{i}}, u_t^{\text{p}}, u_t^{\text{z}}, u_t^{\text{i}})$. Given these quantities, the gradients of the gap, the soft gap and the MCL objective can easily be obtained.

For the purpose of computing the state-control function gradients, the recursions (4.12) and (4.29) are employed. Fortunately, the linear parameterization allows to split the gradient $\nabla_{\boldsymbol{\theta}}$ into the parts $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}$ and $\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4}$. This enables separate treatment of the individual components.

Gradients with Respect to the Rewards of Sensor Model and Secondary Task For the latter part $\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\text{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\text{i}}, u_t^{\text{p}}, u_t^{\text{z}}, u_t^{\text{i}})$ or $\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\text{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\text{i}}, u_t^{\text{p}}, u_t^{\text{z}}, u_t^{\text{i}})$ we can drop the dependency on primary task variables $\boldsymbol{\mu}_t^{\text{p}}, u_t^{\text{p}}$. This is because the optimal policy (3.39) and the MCE policy (3.75) for the sensor control u_t^{z} and the secondary task control u_t^{z} are independent of the primary task variables. Hence, it holds for the state-control function gradients at $t = T$

$$\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_T^{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}, x_T^{\text{i}}, u_T^{\text{z}}, u_T^{\text{i}}) = \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} Q_T^{\star,\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}, x_T^{\text{i}}, u_T^{\text{z}}, u_T^{\text{i}}) = [u_T^{\text{z}}; \boldsymbol{\varphi}(x_T^{\text{i}}, u_T^{\text{i}})]. \tag{4.51}$$

In the remaining time steps t the state-control function gradients are recursively obtained. In this context, the recursion for the optimal policy (4.12) is given as

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\text{i}}, u_t^{\text{z}}, u_t^{\text{i}}) = {} & [u_t^{\text{z}}; \boldsymbol{\varphi}(x_t^{\text{i}}, u_t^{\text{i}})] \\ & + \mathbb{E}\big[\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} Q_{t+1}^{\star,\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t+1}, x_{t+1}^{\text{i}}, u_{t+1}^{\text{z}}, u_{t+1}^{\text{i}})\big| \\ & \qquad \pi_t^{\star}(u_{t+1}^{\text{z}}, u_{t+1}^{\text{i}} | \boldsymbol{x}^{\mathbf{z}}{}_{0:t+1}, x_{t+1}^{\text{i}}), [\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\text{z}} \oplus u_t^{\text{z}}], \mathcal{P}^{\text{i}}(x_{t+1}^{\text{i}} | x_t^{\text{z}}, u_t^{\text{z}}; x_t^{\text{i}}, u_t^{\text{i}})\big]\end{aligned} \tag{4.52}$$

while the recursion for the MCE policy (4.29) can be formulated as

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\text{i}}, u_t^{\text{z}}, u_t^{\text{i}}) = {} & [u_t^{\text{z}}; \boldsymbol{\varphi}(x_t^{\text{i}}, u_t^{\text{i}})] \\ & + \mathbb{E}\big[\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_{t+1}^{\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t+1}, x_{t+1}^{\text{i}}, u_{t+1}^{\text{z}}, u_{t+1}^{\text{i}})\big| \\ & \qquad \tilde{\pi}_t(u_{t+1}^{\text{z}}, u_{t+1}^{\text{i}} | \boldsymbol{x}^{\mathbf{z}}{}_{0:t+1}, x_{t+1}^{\text{i}}), [\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\text{z}} \oplus u_t^{\text{z}}], \mathcal{P}^{\text{i}}(x_{t+1}^{\text{i}} | x_t^{\text{z}}, u_t^{\text{z}}; x_t^{\text{i}}, u_t^{\text{i}})\big].\end{aligned} \tag{4.53}$$

In principle, the recursions can straightly be evaluated as only discrete variables are involved. In the light of the analysis of Sec. 3.4 in the previous chapter, it is clear that this can be infeasible due to the exploding size of the state space of $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$. Fortunately, the recursion remains tractable in the considered scenarios of the algorithms of Cpt. 3 Algo. 4, Algo. 6 and Algo. 7: In the case where the sensor state sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ can be replaced by the EOD d_t the recursion can be evaluated by enumerating all states. If the simple secondary task model is used as in Algo. 6, the optimal gaze policy is given by a single optimal sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}$. Therefore, we can directly obtain the gradient of the initial value function $\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} V_0^{\star,\boldsymbol{\theta}}$ by counting the number of sensor switches $u_t^{\text{z}\,\star} = 1$ and sensor states $x_t^{\text{z}\,\star} = 0$ in the optimal sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}$

$$\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} V_0^{\star,\boldsymbol{\theta}} = \sum_{t=0}^{T} [u_t^{\text{z}\,\star}; 1 - x_t^{\text{z}\,\star}]. \tag{4.54}$$

Gradients with Respect to the Rewards of the Primary Task Unfortunately, in case of $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} Q_t^{\star,\boldsymbol{\theta}}$, $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \tilde{Q}_t^{\boldsymbol{\theta}}$ the dependence on the discrete states cannot be dropped. However, it is possible to split these quantities in several summands similar as in $Q_t^{\star,\boldsymbol{\theta}}, \tilde{Q}_t^{\boldsymbol{\theta}}$. Specifically, it holds

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\big(&\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) + \mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] \\ &+ \mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})\big),\end{aligned} \tag{4.55}$$

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\big(&\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) + \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] \\ &+ \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})\big),\end{aligned} \tag{4.56}$$

where $\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1}, \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \in \mathbb{R}^{(n_x+n_u)^2 \times (n_x+n_u)^2}$, $\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2}, \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2} \in \mathbb{R}^{(n_x+n_u)^2 \times (n_x+n_u)}$ and $\mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}, \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} \in \mathbb{R}^{(n_x+n_u)^2}$. In this context, $\mathbf{P}_{n_x,n_u}^{\mathrm{blk}}(\mathbf{x})$ serves to extract the diagonal blocks corresponding to $\boldsymbol{\mu}_T^{\mathrm{p}}$ and u_T^{p} of the vector $\mathbf{x} \in \mathbb{R}^{(n_x+n_u)^2}$ expanded to a square matrix $\mathbf{X} \in \mathbb{R}^{(n_x+n_u)\times(n_x+n_u)}$. This verified by considering (4.12) and (4.29). Using matrix calculus (see e.g. [175]) we obtain for time step $t = T$

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) &= \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \\ &= \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\Big([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top \operatorname{blk}(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \operatorname{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T})) + \theta_3 u_T^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}})\Big) \quad (4.57) \\ &= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\mathbf{I}^{(n_x+n_u)^2} \operatorname{vec}([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}][\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top) + \mathbf{0}[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \operatorname{vec}\big(\operatorname{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}), \mathbf{0})\big)\Big) \quad (4.58) \\ &= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\mathfrak{M}_T^{Q^{\star,\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}][\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top) + \mathfrak{M}_T^{Q^{\star,\boldsymbol{\theta}},2}[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}})\Big) \quad (4.59) \\ &= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\mathfrak{M}_T^{\tilde{Q}^{\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}][\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top) + \mathfrak{M}_T^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}})\Big). \quad (4.60)\end{aligned}$$

For addressing the remaining steps in Syed's approach to IOC under the optimal policy we define the quantities

$$\mathfrak{F}_t^\star := \begin{bmatrix}\mathbf{I}^{n_x} \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\end{bmatrix}, \quad \mathfrak{T}_t^\star := \begin{bmatrix}\mathbf{A}_t & \mathbf{B}_t \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\mathbf{A}_t & \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\mathbf{B}_t\end{bmatrix}, \quad \mathfrak{t}_t^\star := \begin{bmatrix}\mathbf{a}_t \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\mathbf{a}_t + \mathbf{f}_{t+1}^{\star,\boldsymbol{\theta}}\end{bmatrix}, \tag{4.61}$$

by means of the optimal policy of the primary task $\pi_t^\star(u_t^{\mathrm{p}}|\boldsymbol{\mu}_t^{\mathrm{p}}) = \mathcal{I}(u_t^{\mathrm{p}}|\mathbf{F}_t^{\star,\boldsymbol{\theta}}\boldsymbol{\mu}_t^{\mathrm{p}} + \mathbf{f}_t^{\star,\boldsymbol{\theta}})$. Following some algebraic manipulation which can be found in the appendix (Sec. A.2), the recursion (4.12) is given as

$$\begin{aligned}&\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big((\mathbf{I}^{(n_x+n_u)^2} + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1}\mathfrak{T}_t^\star \otimes \mathfrak{T}_t^\star) \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) \\ &+ \big(\mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},2}\mathfrak{T}_t^\star + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1}(\mathfrak{T}_t^\star \otimes \mathfrak{t}_t^\star + \mathfrak{t}_t^\star \otimes \mathfrak{T}_t^\star)\big)[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] \\ &+ \operatorname{vec}\big(\operatorname{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}), \mathbf{0})\big) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1} \operatorname{vec}\big(\mathfrak{t}_t^\star \mathfrak{t}_t^{\star\top} + \mathfrak{F}_t^\star(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x})\mathfrak{F}_t^{\star\top}\big) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},2}\mathfrak{t}_t^\star \\ &+ \mathbb{E}\big[\mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1} \operatorname{vec}\big(-\mathfrak{F}_t^\star \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\mathfrak{F}_t^{\star\top}\big) \\ &\quad + \mathfrak{m}_{t+1}^{Q^{\star,\boldsymbol{\theta}}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) | \pi_t^\star(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | [\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\big]\Big).\end{aligned} \tag{4.62}$$

In the case of IOC in the maximum causal entropy policy the variables

$$\tilde{\mathfrak{F}}_t := \begin{bmatrix}\mathbf{I}^{n_x} \\ \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\end{bmatrix}, \quad \tilde{\mathfrak{T}}_t := \begin{bmatrix}\mathbf{A}_t & \mathbf{B}_t \\ \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\mathbf{A}_t & \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\mathbf{B}_t\end{bmatrix}, \quad \tilde{\mathfrak{t}}_t := \begin{bmatrix}\mathbf{a}_t \\ \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\mathbf{a}_t + \tilde{\mathbf{f}}_{t+1}^{\boldsymbol{\theta}}\end{bmatrix} \tag{4.63}$$

derived from the MCE policy of the primary task $\tilde{\pi}_t(u_t^{\mathrm{p}}|\boldsymbol{\mu}_t^{\mathrm{p}}) = \mathcal{N}(u_t^{\mathrm{p}}|\tilde{\mathbf{F}}_t^{\boldsymbol{\theta}}\boldsymbol{\mu}_t^{\mathrm{p}} + \tilde{\mathbf{f}}_t^{\boldsymbol{\theta}}, \boldsymbol{\Sigma}^{u_t^{\mathrm{p}},\boldsymbol{\theta}})$ are needed. Using these definitions and some algebraic manipulations (Sec. A.2) the recursion of the MCE policy (4.29) can be conducted by computing

$$
\begin{aligned}
&\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big((\mathbf{I}^{(n_x+n_u)^2} + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\tilde{\mathfrak{T}}_t \otimes \tilde{\mathfrak{T}}_t)\,\mathrm{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top)\\
&+ (\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},2}\tilde{\mathfrak{T}}_t + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}(\tilde{\mathfrak{T}}_t \otimes \tilde{\mathfrak{t}}_t + \tilde{\mathfrak{t}}_t \otimes \tilde{\mathfrak{T}}_t))[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]\\
&+ \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}), \mathbf{0})\right) + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\,\mathrm{vec}\left(\tilde{\mathfrak{t}}_t\tilde{\mathfrak{t}}_t{}^\top + \tilde{\mathfrak{F}}_t\left(\mathbf{A}_t\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x}\right)\tilde{\mathfrak{F}}_t{}^\top + \mathrm{blk}(\mathbf{0}, \boldsymbol{\Sigma}^{u_t^{\mathrm{p}},\boldsymbol{\theta}})\right) + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},2}\tilde{\mathfrak{t}}_t\\
&+ \mathbb{E}\big[\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\,\mathrm{vec}\left(-\tilde{\mathfrak{F}}_t\boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\tilde{\mathfrak{F}}_t{}^\top\right)\\
&\qquad + \mathfrak{m}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}})|\tilde{\pi}_t(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}|[\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})]\Big).
\end{aligned}
\tag{4.64}
$$

Similar as in the previous case, additional assumptions must be made to obtain a computational tractable update. This is because in both $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})$ and $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})$ the element $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ is present. Again, replacing $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ with d_t as in Sec. 3.5.2, Sec. 3.5.3 results in a feasible update that can be performed by means of enumeration. Under restriction of the secondary task model as in Sec. 3.5.2 we can directly obtain $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}Q_0^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_0^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:0}, u_0^{\mathrm{p}}, u_0^{\mathrm{z}})$ by computing

$$
\begin{aligned}
&\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}Q_0^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_0^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:0}, u_0^{\mathrm{p}}, u_0^{\mathrm{z}}) && (4.65)\\
&= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\mathfrak{M}_0^{Q^{\star,\boldsymbol{\theta}},1}\,\mathrm{vec}([\boldsymbol{\mu}_0^{\mathrm{p}}; u_0^{\mathrm{p}}][\boldsymbol{\mu}_0^{\mathrm{p}}; u_0^{\mathrm{p}}]^\top) + \mathfrak{M}_0^{Q^{\star,\boldsymbol{\theta}},2}[\boldsymbol{\mu}_0^{\mathrm{p}}; u_0^{\mathrm{p}}] && (4.66)\\
&\quad + \sum_{t=0}^{T-1}\mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^\star), \mathbf{0})\right) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1}\,\mathrm{vec}\left(\mathfrak{F}_t^\star\left(\mathbf{A}_t\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^\star)\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t+1}{}^\star])\right)\mathfrak{F}_t^{\star\top}\right) && (4.67)\\
&\quad + \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star), \mathbf{0})\right)\Big) && (4.68)
\end{aligned}
$$

using the optimal sensor state sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star$.

We admit, the updates of the (soft) state-control function gradients look quite complicated. Note, however that all the involved operations are more or less standard linear algebra operations such as construction, reshaping, transposition and multiplication of matrices. Hence, the gradient updates can analytically and reasonable quickly be performed if appropriate matrix libraries are used in implementation.

4.4.4 Inverse Optimal Control Algorithms

As a result of the previous derivation, four different approaches for inverse optimal control in the joint task model can be obtained:

1. IOC according to Syed's approach in the optimal policy model under sensor model restriction (SRopt)
2. IOC according to the maximum causal entropy approach under sensor model restriction (SRMCE)
3. IOC according to the maximum causal likelihood approach under sensor model restriction (SRMCL)
4. IOC according to Syed's approach in the optimal policy model under secondary task model restriction (STRopt)

Previously the set of feasible reward parameters Θ was introduced (4.9). This served the purpose to avoid trivial solutions of the IOC problem. In addition to that, the joint task POMDP requires to restrict

$\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. This is because the part of the POMDP that relates to linear quadratic regulation requires a reward $r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) = -\mathbf{x}^{\mathrm{p}}{}_t^{\top}\mathbf{C}_x\mathbf{x}_t^{\mathrm{p}} - u^{\mathrm{p}}{}_t^{\top}\mathbf{C}_u u_t^{\mathrm{p}}$ where both $\mathbf{C}_x$ is positive semidefinite and where $\mathbf{C}_u$ is positive definite. Hence, $\boldsymbol{\Theta}_1 = -\mathbf{C}_x$ and $\boldsymbol{\Theta}_2 = -\mathbf{C}_u$ need to be negative semidefinite and negative definite respectively. Thus, we can for example, think of the feasible set Θ defined as

$$\Theta = \{\boldsymbol{\theta} \colon \boldsymbol{\Theta}_1 \preceq \mathbf{0},\ \boldsymbol{\Theta}_2 \preceq -\varepsilon\mathbf{I}^{n_u},\ \theta_3 \geq 0,\ \boldsymbol{\theta}_4 \geq \mathbf{0},\ \theta_3 + \textstyle\sum \boldsymbol{\theta}_4 = 1\}. \tag{4.69}$$

This is a convex set that allows for efficient projection onto [32, 172]. Consequently, we can globally solve all IOC approaches by a generic projected gradient descent technique [172].

In the following, we present a concrete algorithmic solution approaches for (SRopt), (SRMCE), (SRMCL) and (STRopt). Due to space constraints, we only give the simplest computational procedures. In many cases more efficient algorithms are possible exploiting the ideas used for computing optimal and MCE policies. We will indicate where this is the case and refer to the previously introduced algorithms.

Algorithm 8 Generic Projected Gradient Descent IOC [SolveIOC]

function SolveIOC($(\mathbf{A}, \mathbf{a}, \mathbf{B})_{0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D}$)
Require: Stepsize η, feasible set of reward parameters Θ
  $\boldsymbol{\theta} \leftarrow$ Sample(Θ) ▷ sample a random initial parameter from feasible set
  while not converged $\boldsymbol{\theta}$ **do**
    $[v, \nabla_{\boldsymbol{\theta}}] \leftarrow$ Eval<name>$(\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D})$ ▷ evaluate the different IOC objectives SRopt Algo. 9, SRMCE Algo. 11, SRMCL Algo. 12 or STRopt Algo. 14
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}$
    $\boldsymbol{\theta} \leftarrow$ Project$(\boldsymbol{\theta}, \Theta)$ ▷ project parameters on feasible set
  end while
  return $\boldsymbol{\theta}$
end function

Algorithm 9 Evaluation of SRopt [EvalSRopt]

function EvalSRopt($\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathsf{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathsf{z}}), \mathcal{P}^i, \mathsf{D}$)

Require: steady state covariance for gaze on road $\hat{\boldsymbol{\Sigma}}_0^{\mathsf{p}}$

$(\boldsymbol{\mu}_0^{\mathsf{p}}, d_0, x_0^{\mathsf{i}}) \leftarrow \mathsf{D}$

$(\boldsymbol{\Sigma}_0^{\mathsf{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathsf{p}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z, \boldsymbol{\lambda}}(x_t^{\mathsf{z}}))$ ▷ as described in Sec. 4.4.2

$\mathbf{C}_x \leftarrow -\boldsymbol{\Theta}_1, \quad \mathbf{C}_u \leftarrow -\boldsymbol{\Theta}_2, \quad r(u_t^{\mathsf{z}}) \leftarrow \boldsymbol{\theta}_3(u_t^{\mathsf{z}}), \quad r(x_t^{\mathsf{i}}, u_t^{\mathsf{i}}) \leftarrow \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathsf{i}}, u_t^{\mathsf{i}})$

$\left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_t^\star, \pi_t^\star(u_t^{\mathsf{z}}, u_t^{\mathsf{i}} | d_t, x_t^{\mathsf{i}})\right)_{t=0:T} \leftarrow$ OptJTSR$(\mathbf{C}_x, \mathbf{C}_u, r(u_t^{\mathsf{z}}), r(x_t^{\mathsf{i}}, u_t^{\mathsf{i}}), (\mathbf{A}, \mathbf{a}, \mathbf{B})_{0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathsf{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathsf{z}}), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathsf{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathsf{p}})$ ▷ Algo. 4

$v \leftarrow \mathbb{E}\left[\boldsymbol{\mu}_0^{\mathsf{p}\,\top} \mathbf{M}_0^{V^\star} \boldsymbol{\mu}_0^{\mathsf{p}} + \boldsymbol{\mu}_0^{\mathsf{p}\,\top} \mathbf{m}_0^{V^\star} + m_0^{V^\star,1}(d_0, x_0^{\mathsf{i}}) \middle| \mathcal{N}(\boldsymbol{\mu}_0^{\mathsf{p}} | \mathbf{x}_0^{\mathsf{p}}, \boldsymbol{\Sigma}_0^{\mathsf{p}})\right] - \mathbb{E}\left[\sum_{t=0}^T \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathsf{p}} \mathbf{x}_t^{\mathsf{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathsf{p}} u_t^{\mathsf{p}\,\top}) + \theta_3 u_t^{\mathsf{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathsf{i}}, u_t^{\mathsf{i}}) \middle| \mathsf{D}\right]$ ▷ Algo. 10

$(\mathfrak{M}_t^{Q^{\star\boldsymbol{\theta}},1}, \mathfrak{M}_t^{Q^{\star\boldsymbol{\theta}},2}, \mathfrak{m}_t^{Q^{\star\boldsymbol{\theta}}}, \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} Q_t^{\star\boldsymbol{\theta}}(d_t, x_t^{\mathsf{i}}, u_t^{\mathsf{z}}, u_t^{\mathsf{i}}))_{t=0:T} \leftarrow$ QGJTSROpt$((\mathbf{F}_t^\star, \mathbf{f}_t^\star, \pi_t^\star(u_t^{\mathsf{z}}, u_t^{\mathsf{i}} | d_t, x_t^{\mathsf{i}}, \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathsf{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathsf{z}}), \mathcal{P}^i, x_0^{\mathsf{z}}, \boldsymbol{\Sigma}_0^{\mathsf{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathsf{p}})$

$\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow \mathbb{E}\left[\mathbf{P}^{\mathrm{blk}}_{n_x,n_u}(\mathfrak{M}_0^{Q^{\star\boldsymbol{\theta}},1} \mathrm{vec}([\boldsymbol{\mu}_0^{\mathsf{p}}; u_0^{\mathsf{p}}][\boldsymbol{\mu}_0^{\mathsf{p}}; u_0^{\mathsf{p}}]^\top) + \mathfrak{M}_0^{Q^{\star\boldsymbol{\theta}},2}[\boldsymbol{\mu}_0^{\mathsf{p}}; u_0^{\mathsf{p}}] + \mathfrak{m}_0^{Q^{\star\boldsymbol{\theta}}}(x_0^{\mathsf{z}}, x_0^{\mathsf{i}}, u_0^{\mathsf{z}}, u_0^{\mathsf{i}})) \middle| \mathcal{I}(u_0^{\mathsf{p}} | \mathbf{F}_0^{\star,\boldsymbol{\theta}} \boldsymbol{\mu}_0^{\mathsf{p}} + \mathbf{f}_0^{\star,\boldsymbol{\theta}}), \pi^\star(u_0^{\mathsf{z}}, u_0^{\mathsf{i}} | d_0, x_0^{\mathsf{i}}), \mathcal{N}(\boldsymbol{\mu}_0^{\mathsf{p}} | \mathbf{x}_0^{\mathsf{p}}, \boldsymbol{\Sigma}_0^{\mathsf{p}})\right]$

$\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow^- \mathbb{E}\left[\sum_{t=0}^T \begin{bmatrix} \mathrm{vec}(\mathbf{x}_t^{\mathsf{p}} \mathbf{x}_t^{\mathsf{p}\,\top}) \\ \mathrm{vec}(u_t^{\mathsf{p}} u_t^{\mathsf{p}\,\top}) \end{bmatrix} \middle| \mathsf{D}\right]$

$\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \leftarrow \mathbb{E}[\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} Q_0^{\star\boldsymbol{\theta}}(d_0, x_0^{\mathsf{i}}, u_0^{\mathsf{z}}, u_0^{\mathsf{i}}) | \pi^\star(u_0^{\mathsf{z}}, u_0^{\mathsf{i}} | d_0, x_0^{\mathsf{i}})] - \mathbb{E}\left[\sum_{t=0}^T \begin{bmatrix} u_t^{\mathsf{z}} \\ \boldsymbol{\varphi}(x_t^{\mathsf{i}}, u_t^{\mathsf{i}}) \end{bmatrix} \middle| \mathsf{D}\right]$

$\nabla_{\boldsymbol{\theta}} \leftarrow \begin{bmatrix} \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \\ \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \end{bmatrix}$

return $\left(v, \nabla_{\boldsymbol{\theta}}, \left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_t^\star, \pi_t^\star(u_t^{\mathsf{z}}, u_t^{\mathsf{i}} | d_t, x_t^{\mathsf{i}})\right)_{t=0:T}\right)$

end function

Algorithm 10 Computation of the Gradient of the State-Control Function in SRopt [GQSRopt]

function GQSROPT($(\mathbf{F}_t^\star, \mathbf{f}_t^\star, \pi_t^\star(d_t, x_t^{\mathrm{i}}), \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}}$)

  $\mathfrak{M}_T^{Q^{\star,\boldsymbol{\theta}},1} \leftarrow \mathbf{I}^{(n_x+n_u)^2}$

  $\mathfrak{M}_T^{Q^{\star,\boldsymbol{\theta}},2} \leftarrow \mathbf{0}$

  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}} \mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \leftarrow \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(d_T), \mathbf{0})\right)$

  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}} \nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} Q_T^{\star,\boldsymbol{\theta}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \leftarrow [u_T^{\mathrm{z}}; \boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}})]$

  for $t = T-1:0$ **do**

    $\mathfrak{F}_t^\star \leftarrow \begin{bmatrix} \mathbf{I}^{n_x} \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}} \end{bmatrix}, \quad \mathfrak{T}_t^\star \leftarrow \begin{bmatrix} \mathbf{A}_t & \mathbf{B}_t \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}} \mathbf{A}_t & \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}} \mathbf{B}_t \end{bmatrix}, \quad \mathfrak{t}_t^\star \leftarrow \begin{bmatrix} \mathbf{a}_t \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}} \mathbf{a}_t + \mathbf{f}_{t+1}^{\star,\boldsymbol{\theta}} \end{bmatrix}$

    $\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1} \leftarrow \mathbf{I}^{(n_x+n_u)^2} + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1} \mathfrak{T}_t^\star \otimes \mathfrak{T}_t^\star$

    $\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2} \leftarrow \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1} (\mathfrak{T}_t^\star \otimes \mathfrak{t}_t^\star + \mathfrak{t}_t^\star \otimes \mathfrak{T}_t^\star)$

    for all $d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}$ **do** ▷ In analogy to Algo. 4

      $\mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(d_t), \mathbf{0})\right) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1} \mathrm{vec}\left(\mathfrak{t}_t^\star \mathfrak{t}_t^{\star\top} + \mathfrak{F}_t^\star \left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(d_t) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}(d_{t+1}(d_t, u_t^{\mathrm{z}}))\right) \mathfrak{F}_t^{\star\top}\right) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},2} \mathfrak{t}_t^\star$

      $\mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow^+ \mathbb{E}\left[\mathfrak{m}_{t+1}^{Q^{\star,\boldsymbol{\theta}}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) | \pi_t^\star(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), d_{t+1}(d_t, u_t^{\mathrm{z}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$

      $\nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} Q_t^{\star,\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow [u_t^{\mathrm{z}}; \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}})] + \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} Q_{t+1}^{\star,\boldsymbol{\theta}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) | \pi_t^\star(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), d_{t+1}(d_t, u_t^{\mathrm{z}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$

    end for

  end for

  return $(\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1}, \mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2}, \mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}, \nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} Q_t^{\star,\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}))_{t=0:T}$

end function

Algorithm 11 Evaluation of SRMCE [EvalSRMCE]

function EvalSRMCE($\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D}$)

$(\boldsymbol{\mu}_0^{\mathrm{p}}, d_0, x_0^{\mathrm{i}}) \leftarrow \mathsf{D}$

$(\boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z, \boldsymbol{\lambda}}(x_t^{\mathrm{z}}))$ ▷ as described in Sec. 4.4.2

$\mathbf{C}_x \leftarrow -\boldsymbol{\Theta}_1, \quad \mathbf{C}_u \leftarrow -\boldsymbol{\Theta}_2, \quad r(u_t^{\mathrm{z}}) \leftarrow \boldsymbol{\theta}_3(u_t^{\mathrm{z}}), \quad r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \leftarrow \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}})$

$\left(\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \boldsymbol{\Sigma}^{u_t^{\mathrm{p}}, \boldsymbol{\theta}}, \tilde{\pi}_t(u_t^{\mathrm{z}}, u_t^{\mathrm{i}} | d_t, x^{\mathrm{i}})\right)_{t=0:T} \leftarrow$ SRMCE$(\mathbf{C}_x, \mathbf{C}_u, r(u_t^{\mathrm{z}}), r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}), (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}})$ ▷ Algo. 7

$v \leftarrow \mathbb{E}\left[\boldsymbol{\mu}_0^{\mathrm{p}\top} \mathbf{M}_0^{\tilde{V}} \boldsymbol{\mu}_0^{\mathrm{p}} + \boldsymbol{\mu}_0^{\mathrm{p}\top} \mathbf{m}_0^{\tilde{V}} + m_0^{\tilde{V},1}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \middle| \mathcal{N}(\boldsymbol{\mu}_0^{\mathrm{p}} | \mathbf{x}_0^{\mathrm{p}}, \boldsymbol{\Sigma}_0^{\mathrm{p}})\right] - \mathbb{E}\left[\sum_{t=0}^T \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \middle| \mathsf{D}\right]$ ▷ Algo. 13

$\left(\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1}, \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}, \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}, \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})\right)_{t=0:T} \leftarrow$ GQSRMCE$((\tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \tilde{\pi}_t(u_t^{\mathrm{z}}, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}}), \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}})$

$\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow \mathbb{E}\left[\mathbf{P}_{n_x,n_u}^{\mathrm{blk}} (\mathfrak{M}_0^{\tilde{Q}^{\boldsymbol{\theta}},1} \mathrm{vec}([\boldsymbol{\mu}_0^{\mathrm{p}}; u_0^{\mathrm{p}}][\boldsymbol{\mu}_0^{\mathrm{p}}; u_0^{\mathrm{p}}]^\top) + \mathfrak{M}_0^{\tilde{Q}^{\boldsymbol{\theta}},2} [\boldsymbol{\mu}_0^{\mathrm{p}}; u_0^{\mathrm{p}}] + \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}, u_0^{\mathrm{z}}, u_0^{\mathrm{i}})) \middle| \mathcal{N}(u_0^{\mathrm{p}} | \tilde{\mathbf{F}}_0^{\boldsymbol{\theta}} \boldsymbol{\mu}_0^{\mathrm{p}} + \tilde{\mathbf{f}}_0^{\boldsymbol{\theta}}, \boldsymbol{\Sigma}^{u_0^{\mathrm{p}}, \boldsymbol{\theta}}), \tilde{\pi}(u_0^{\mathrm{z}}, u_0^{\mathrm{i}} | d_0, x_0^{\mathrm{i}}), \mathcal{N}(\boldsymbol{\mu}_0^{\mathrm{p}} | \mathbf{x}_0^{\mathrm{p}}, \boldsymbol{\Sigma}_0^{\mathrm{p}})\right]$

$\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow^- \mathbb{E}\left[\sum_{t=0}^T \begin{bmatrix} \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\top}) \\ \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\top}) \end{bmatrix} \middle| \mathsf{D}\right]$

$\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \leftarrow \mathbb{E}[\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_0^{\boldsymbol{\theta}}(d_0, x_0^{\mathrm{i}}, u_0^{\mathrm{z}}, u_0^{\mathrm{i}}) | \tilde{\pi}(u_0^{\mathrm{z}}, u_0^{\mathrm{i}} | d_0, x_0^{\mathrm{i}})] - \mathbb{E}\left[\sum_{t=0}^T \begin{bmatrix} u_t^{\mathrm{z}} \\ \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \end{bmatrix} \middle| \mathsf{D}\right]$

$\nabla_{\boldsymbol{\theta}} \leftarrow \begin{bmatrix} \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \\ \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \end{bmatrix}$

return $\left(v, \nabla_{\boldsymbol{\theta}}, \left(\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \tilde{\pi}_t(u_t^{\mathrm{z}}, u_t^{\mathrm{i}} | d_t, x^{\mathrm{i}})\right)_{t=0:T}\right)$

end function

Algorithm 12 Evaluation of SRMCL [EvalSRMCL]

function EvalSRMCL($\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z), \mathcal{P}^i, \mathsf{D}$)

  $(\boldsymbol{\mu}_0^{\mathrm{p}}, d_0, x_0^{\mathrm{i}}) \leftarrow \mathsf{D}$

  $(\boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z, \boldsymbol{\lambda}}(x_t^z))$ ▷ as described in Sec. 4.4.2

  $\mathbf{C}_x \leftarrow -\boldsymbol{\Theta}_1, \quad \mathbf{C}_u \leftarrow -\boldsymbol{\Theta}_2, \quad r(u_t^z) \leftarrow \boldsymbol{\theta}_3(u_t^z), \quad r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \leftarrow \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}})$

  $\left(\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \boldsymbol{\Sigma}^{u_t^{\mathrm{p}}, \boldsymbol{\theta}}, \tilde{\pi}_t(u_t^z, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}})\right)_{t=0:T} \leftarrow$ SRMCE$(\mathbf{C}_x, \mathbf{C}_u, r(u_t^z), r(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}), (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}})$ ▷ Algo. 7

  $\left(\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1}, \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}, \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}, \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}})\right)_{t=0:T} \leftarrow$ GQSRMCE$((\tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \tilde{\pi}_t(u_t^z, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}}), \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}})$ ▷ Algo. 13

  $v \leftarrow 0, \quad \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow \mathbf{0}, \quad \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \leftarrow \mathbf{0}$

  for $t = 0 : T$ **do**

    $v \leftarrow^+ \mathbb{E}\left[\boldsymbol{\mu}_t^{\mathrm{p}\,\top} \mathbf{M}_t^{\tilde{V}} \boldsymbol{\mu}_t^{\mathrm{p}} + \boldsymbol{\mu}_t^{\mathrm{p}\,\top} \mathbf{m}_t^{\tilde{V}} + m_t^{\tilde{V},1}(d_t, x_t^{\mathrm{i}}) \middle| \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}} | \mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(d_t)), \mathsf{D}\right]$

    $v \leftarrow^- \mathbb{E}\left[[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{m}_t^{\tilde{Q}} + m_t^{\tilde{Q},1}(d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}) \middle| \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}} | \mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(d_t)), \mathsf{D}\right]$

    $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow^+ \mathbb{E}\Big[\mathbf{P}_{n_x,n_u}^{\mathrm{blk}}(\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) + \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(x_t^z, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}})) \Big| \mathcal{N}(u_t^{\mathrm{p}} | \tilde{\mathbf{F}}_t^{\boldsymbol{\theta}} \boldsymbol{\mu}_t^{\mathrm{p}} + \tilde{\mathbf{f}}_t^{\boldsymbol{\theta}}, \boldsymbol{\Sigma}^{u_t^{\mathrm{p}},\boldsymbol{\theta}}),$
    $\tilde{\pi}(u_t^z, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}}), \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}} | \mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(d_t)), \mathsf{D}\Big]$

    $\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \leftarrow^- \mathbb{E}\Big[\mathbf{P}_{n_x,n_u}^{\mathrm{blk}}(\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) + \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(x_t^z, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}})) \Big| \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}} | \mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p}}(d_t)), \mathsf{D}\Big]$

    $\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \leftarrow^+ \mathbb{E}\left[\mathbb{E}[\nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}) | \tilde{\pi}(u_t^z, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}})] - \tilde{Q}_t^{\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}) \middle| \mathsf{D}\right]$

  end for

  $\nabla_{\boldsymbol{\theta}} \leftarrow \begin{bmatrix} \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \\ \nabla_{\boldsymbol{\theta}_3,\boldsymbol{\theta}_4} \end{bmatrix}$

  return $(v, \nabla_{\boldsymbol{\theta}})$

end function

Algorithm 13 Computation of the Gradient of the State-Control Function in SRMCE [GQSRMCE]

function GQSRMCE($(\tilde{\mathbf{F}}_t, \mathbf{f}_t^\star, \pi_t^\star(d_t, x_t^{\mathrm{i}}), \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^{\mathrm{z}}), \mathcal{P}^i, d_0, \boldsymbol{\Sigma}_0^{\mathrm{p}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}}$)

  $\mathfrak{M}_T^{\tilde{Q}^{\boldsymbol{\theta}},1} \leftarrow \mathbf{I}^{(n_x+n_u)^2}$

  $\mathfrak{M}_T^{\tilde{Q}^{\boldsymbol{\theta}},2} \leftarrow \mathbf{0}$

  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}} \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \leftarrow \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(d_T), \mathbf{0})\right)$

  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}} \nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} \tilde{Q}_T^{\boldsymbol{\theta}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \leftarrow [u_T^{\mathrm{z}}; \boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}})]$

  for $t = T-1 : 0$ **do**

    $\tilde{\mathfrak{F}}_t \leftarrow \begin{bmatrix} \mathbf{I}^{n_x} \\ \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}} \end{bmatrix}, \quad \tilde{\mathfrak{T}}_t \leftarrow \begin{bmatrix} \mathbf{A}_t & \mathbf{B}_t \\ \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}} \mathbf{A}_t & \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}} \mathbf{B}_t \end{bmatrix}, \quad \tilde{\mathfrak{t}}_t \leftarrow \begin{bmatrix} \mathbf{a}_t \\ \tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}} \mathbf{a}_t + \tilde{\mathbf{f}}_{t+1}^{\boldsymbol{\theta}} \end{bmatrix}$

    $\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \leftarrow \mathbf{I}^{(n_x+n_u)^2} + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1} \tilde{\mathfrak{T}}_t \otimes \tilde{\mathfrak{T}}_t$

    $\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2} \leftarrow \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1} (\tilde{\mathfrak{T}}_t \otimes \tilde{\mathfrak{t}}_t + \tilde{\mathfrak{t}}_t \otimes \tilde{\mathfrak{T}}_t)$

    for all $d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}$ **do** ▷ In analogy to Algo. 7

      $\mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(d_t), \mathbf{0})\right) + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1} \mathrm{vec}\left(\tilde{\mathfrak{t}}_t \tilde{\mathfrak{t}}_t{}^\top + \tilde{\mathfrak{F}}_t \left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p}}(d_t) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}(d_{t+1}(d_t, u_t^{\mathrm{z}}))\right) \tilde{\mathfrak{F}}_t{}^\top + \mathrm{blk}(\mathbf{0}, \boldsymbol{\Sigma}^{u_t^{\mathrm{p}}, \boldsymbol{\theta}})\right) + \mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},2} \tilde{\mathfrak{t}}_t$

      $\mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow^+ \mathbb{E}\left[\mathfrak{m}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) | \pi_t^\star(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), d_{t+1}(d_t, u_t^{\mathrm{z}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$

      $\nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow [u_t^{\mathrm{z}}; \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}})] + \mathbb{E}\left[\nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} \tilde{Q}_{t+1}^{\boldsymbol{\theta}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) \middle| \pi_t^\star(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), d_{t+1}(d_t, u_t^{\mathrm{z}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$

    end for

  end for

  return $(\mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},1}, \mathfrak{M}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}, \mathfrak{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}, \nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} \tilde{Q}_t^{\boldsymbol{\theta}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}))_{t=0:T}$

end function

Algorithm 14 Evaluation of STRopt [EvalSTRopt]

function EvalSTRopt($\boldsymbol{\theta}$, $(\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}$, $\boldsymbol{\Sigma}^{\epsilon^x}$, $\mathbf{H}(x_t^z)$, $\boldsymbol{\Sigma}^{\epsilon^z}(x_t^z)$, $\mathcal{P}^i$, D)

$(\boldsymbol{\mu}_0^p, x_0^z) \leftarrow \mathsf{D}$

$\boldsymbol{\Sigma}_0^p \leftarrow \textsc{Initialize}(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z, \boldsymbol{\lambda}}(x_t^z))$ ▷ as described in Sec. 4.4.2

$\mathbf{C}_x \leftarrow -\boldsymbol{\Theta}_1, \quad \mathbf{C}_u \leftarrow -\boldsymbol{\Theta}_2, \quad r(u_t^z) \leftarrow \boldsymbol{\theta}_3(u_t^z), \quad r(x_t^z) \leftarrow \theta_4(1 - x_t^z)$

$\left((\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, \mathbf{F}_t^\star, \mathbf{f}_t^\star)_{t=0:T}, \boldsymbol{x^z}_{0:T}{}^\star, m_0^{V^\star,1}(x_0^z)\right) \leftarrow \textsc{STRopt}(\mathbf{C}_x, \mathbf{C}_u, r(u_t^z), r(x_t^z), (\mathbf{A}, \mathbf{a}, \mathbf{B})_{0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z), x_0^z, \boldsymbol{\Sigma}_0^p)$ ▷ Algo. 6

$v \leftarrow \mathbb{E}\left[\boldsymbol{\mu}_0^{p\top} \mathbf{M}_0^{V^\star} \boldsymbol{\mu}_0^p + \boldsymbol{\mu}_0^{p\top} \mathbf{m}_0^{V^\star} + m_t^{V^\star,1}(x_0^z)) \middle| \mathcal{N}(\boldsymbol{\mu}_0^p | \mathbf{x}_0^p, \boldsymbol{\Sigma}_0^p)\right] - \mathbb{E}\left[\sum_{t=0}^T \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^p \mathbf{x}_t^{p\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^p u_t^{p\top}) + \theta_3 u_t^z + \theta_4(1 - x_t^z) \middle| \mathsf{D}\right]$

$(\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1}, \mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2}, \mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}})_{t=0:T} \leftarrow \textsc{GQSTRopt}((\mathbf{F}_t^\star, \mathbf{f}_t^\star, \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T})$ ▷ Algo. 15

$\nabla_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \leftarrow \mathbb{E}\left[\mathbf{P}_{n_x, n_u}^{\mathrm{blk}}(\mathfrak{M}_0^{Q^{\star,\boldsymbol{\theta}},1} \mathrm{vec}([\boldsymbol{\mu}_0^p; u_0^p][\boldsymbol{\mu}_0^p; u_0^p]^\top) + \mathfrak{M}_0^{Q^{\star,\boldsymbol{\theta}},2}[\boldsymbol{\mu}_0^p; u_0^p]) \middle| \mathcal{I}(u_t^p | \mathbf{F}_t^{\star,\boldsymbol{\theta}} \boldsymbol{\mu}_t^p + \mathbf{f}_t^{\star,\boldsymbol{\theta}}), \mathcal{N}(\boldsymbol{\mu}_0^p | \mathbf{x}_0^p, \boldsymbol{\Sigma}_0^p)\right] + \mathbf{P}_{n_x, n_u}^{\mathrm{blk}}(\mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}})$

$\nabla_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \leftarrow^+ \mathbf{P}_{n_x, n_u}^{\mathrm{blk}}\left(\sum_{t=0}^{T-1} \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^p(\boldsymbol{x^z}_{0:t}{}^\star), \mathbf{0})\right) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1} \mathrm{vec}\left(\mathfrak{F}_t^\star \left(\mathbf{A}_t \boldsymbol{\Sigma}_t^p(\boldsymbol{x^z}_{0:t}{}^\star) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^p([\boldsymbol{x^z}_{0:t+1}{}^\star])\right) \mathfrak{F}_t^{\star\top}\right) + \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^p(\boldsymbol{x^z}_{0:T}{}^\star), \mathbf{0})\right)\right)$

$\nabla_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \leftarrow^- \mathbb{E}\left[\sum_{t=0}^T [\mathrm{vec}(\mathbf{x}_t^p \mathbf{x}_t^{p\top}); \mathrm{vec}(u_t^p u_t^{p\top})] \middle| \mathsf{D}\right]$

$\nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} \leftarrow \sum_{t=0}^T [u_t^z{}^\star; 1 - x_t^z{}^\star] - \mathbb{E}\left[\sum_{t=0}^T [u_t^z; 1 - x_t^z] \middle| \mathsf{D}\right]$

$\nabla_{\boldsymbol{\theta}} \leftarrow \begin{bmatrix} \nabla_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2} \\ \nabla_{\boldsymbol{\theta}_3, \boldsymbol{\theta}_4} \end{bmatrix}$

return $\left(v, \nabla_{\boldsymbol{\theta}}, (\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, \mathbf{F}_t^\star, \mathbf{f}_t^\star)_{t=0:T}, \boldsymbol{x^z}_{0:T}{}^\star\right)$

end function

Algorithm 15 Computation of the Gradient of the State-Control Function in STRopt [GQSTRopt]

function GQSTRopt($(\mathbf{F}_t^\star, \mathbf{f}_t^\star, \mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}$)
  $\mathfrak{M}_T^{Q^{\star,\boldsymbol{\theta}},1} \leftarrow \mathbf{I}^{(n_x+n_u)^2}$
  $\mathfrak{M}_T^{Q^{\star,\boldsymbol{\theta}},2} \leftarrow \mathbf{0}$
  $\mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}} \leftarrow \mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(d_T), \mathbf{0})\right)$
  for $t = T-1:0$ **do**
    $\mathfrak{F}_t^\star \leftarrow \begin{bmatrix} \mathbf{I}^{n_x} \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}} \end{bmatrix}, \quad \mathfrak{T}_t^\star \leftarrow \begin{bmatrix} \mathbf{A}_t & \mathbf{B}_t \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\mathbf{A}_t & \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\mathbf{B}_t \end{bmatrix}, \quad \mathfrak{t}_t^\star \leftarrow \begin{bmatrix} \mathbf{a}_t \\ \mathbf{F}_{t+1}^{\star,\boldsymbol{\theta}}\mathbf{a}_t + \mathbf{f}_{t+1}^{\star,\boldsymbol{\theta}} \end{bmatrix}$
    $\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1} \leftarrow \mathbf{I}^{(n_x+n_u)^2} + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1}\mathfrak{T}_t^\star \otimes \mathfrak{T}_t^\star$
    $\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2} \leftarrow \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1}(\mathfrak{T}_t^\star \otimes \mathfrak{t}_t^\star + \mathfrak{t}_t^\star \otimes \mathfrak{T}_t^\star)$
    $\mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},1}\,\mathrm{vec}\left(\mathfrak{t}_t^\star \mathfrak{t}_t^{\star\top}\right) + \mathfrak{M}_{t+1}^{Q^{\star,\boldsymbol{\theta}},2}\mathfrak{t}_t^\star + \mathfrak{m}_{t+1}^{Q^{\star,\boldsymbol{\theta}}}$
  end for
  return $(\mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},1}, \mathfrak{M}_t^{Q^{\star,\boldsymbol{\theta}},2}, \mathfrak{m}_t^{Q^{\star,\boldsymbol{\theta}}})_{t=0:T}$
end function

4.5 Evaluation on Simulated Data

In the previous section approaches for IOC for the class of POMDPs of the normative model of glance behavior have been proposed. Specifically, we presented methods for inverse optimal control under the assumption of optimal behavior SRopt-IOC Algo. 9, STRopt-IOC Algo. 14 and under the assumption that the agent applies a maximum causal entropy policy SRMCE-IOC Algo. 11, STRMCL-IOC Algo. 12.

SRopt-IOC and STRopt-IOC address different variants of the joint task POMDP, the sensor model restriction and the secondary task model restriction. The viability of the individual policy models and the employed restrictions (of either sensor model or secondary task model) can only assessed by evaluation on real data. For example the accuracy in predicting the observed behavior can be investigated.

SRMCE-IOC and SRMCL-IOC both address the POMDP under sensor model restriction. Whereas, SRMCE-IOC estimates reward parameters by means of minimizing the soft gap (4.49) SRMCL-IOC tries to infer these by means of minimizing the negative log-likelihood of the MCE policy (4.50). As noted before, both approaches estimate the same reward parameters in the limit of infinite data, however the estimation can differ on a *finite* dataset. Both MCE-IOC and MCL-IOC are known in the literature, [29, 77] used MCL-IOC in their works, whereas most authors applied the original variant e.g. [259, 117]. However, to the best of our knowledge no comparison of both variants has been made so far.

4.5.1 Scenario

To obtain a better understanding of both variants we therefore conduct comparison of MCE-IOC and MCL-IOC. To the best of our knowledge this is the first empirical investigation of the difference between those methods. For this purpose, data simulated by a MCE policy for a fixed reward parameter $\boldsymbol{\theta}$ is used. This is because in real behavioral data the model assumptions might not perfectly be fulfilled what can have an influence on the results. In the course of the evaluation we also conduct a comparison to a directly estimated policy (Direct Policy Estimation, DPE) as employed in e.g. [78, 95]. This serves to investigate potential benefits of the IOC methods. We will consider both the evaluation on the same driving scenario as well as the transfer to a previously unseen driving scenario.

Inverse Optimal Control Methods

The evaluation was conducted with the joint task POMDP under sensor model restriction (Sec. 3.5.3). Correspondingly, Algo. 7 was used to compute the maximum causal entropy policies whereas SRMCE-IOC or MCE-MCL were used for reward inference. The parameters of the vehicle model were set to the values estimated from the test vehicle used in the driving experiment I (Sec. 4.6). In the evaluation the reward model

$$r(\mathbf{x}_t^{\mathrm{p}}, u_t^{\mathrm{p}}) = \theta_1 (y_t)^2 + \theta_2 (\dot{y}_t)^2 + \theta_3 (\alpha_t)^2 + \theta_4 (u_t^{\mathrm{p}})^2$$

based on the lane position, the lateral velocity, the steering angle and the steering angle velocity was employed which was previously introduced in Sec. 3.3.1. Furthermore, the sensor model restriction was implemented by assuming that the driver can fully sense all primary task states when his gaze is on the road. Specifically, we set the steady state covariance to $\hat{\boldsymbol{\Sigma}}_0^{\mathrm{p}} = \mathbf{0}$. For the case that the driver's gaze was off the road we employed the sensor noise covariance of $\boldsymbol{\Sigma}^{\epsilon^z}(x_t^z = 0) = [\infty; 0; \infty; 0]$ which models a driver that does not obtain any information of the forward road scenery when averting his or her gaze. As secondary task model the simple task model $x_t^{\mathrm{i}} = x_t^z$, $r(x_t^{\mathrm{i}}) = \theta(1 - x_t^{\mathrm{i}}) = \theta(1 - x_t^z)$ that was introduced in Sec. 3.3.3 was employed. The corresponding optimization problems for SRMCE-IOC and SRMCL-IOC were solved using a standard unconstrained optimization solver. In this context, a barrier $-10^{-4} \sum_{i=1}^{4} \log(-\boldsymbol{\theta}_i)$ was added to the objectives $\tilde{g}(\boldsymbol{\theta}, \mathsf{D})$, $l(\boldsymbol{\theta}, \mathsf{D})$ (4.49), (4.50) to ensure that the LQR sub-MDP of the joint POMDP is well defined. Finally, a relative gradient norm below a tolerance of 10^{-6} was used as a termination criterion.

Direct Policy Estimation Baseline

Given the knowledge, that the policy resulting from both IOC approaches factorizes into $\tilde{\pi}_t(u_t^{\mathrm{p}}|\boldsymbol{\mu}_t^{\mathrm{p}}) = \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}}|\tilde{\mathbf{F}}_t^{\boldsymbol{\theta}}\boldsymbol{\mu}_t^{\mathrm{p}} + \tilde{\mathbf{f}}_t^{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_t^{u_t^{\mathrm{p}}})$ and $\tilde{\pi}_t(u_t^{\mathrm{z}}|d_t)$ (3.75), we used the parametrization

$$\pi^{\mathrm{base}}(u_t^{\mathrm{p}}, u_t^{\mathrm{z}}|\boldsymbol{\mu}_t^{\mathrm{p}}, d_t) \propto \mathcal{N}(u_t^{\mathrm{p}}|\boldsymbol{\Lambda}_1^{\mathrm{base}}\boldsymbol{\mu}_t^{\mathrm{p}} + \boldsymbol{\lambda}_2^{\mathrm{base}}, \boldsymbol{\Sigma}^{\mathrm{base}}) \exp\left(u_t^{\mathrm{z}}(\lambda_3^{\mathrm{base}} d_t + \lambda_4^{\mathrm{base}} x_t^{\mathrm{z}} + \lambda_5^{\mathrm{base}}(1 - x_t^{\mathrm{z}}))\right), \quad (4.70)$$

for comparison. Time-invariant parameters were employed, as first experiments showed that this did not negatively affect performance. In every experiment $\boldsymbol{\Lambda}_1^{\mathrm{base}}, \boldsymbol{\lambda}_2^{\mathrm{base}}, \lambda_3^{\mathrm{base}}, \lambda_4^{\mathrm{base}}, \lambda_5^{\mathrm{base}}$ were inferred by L1- regularized maximum likelihood estimation for generalized linear models [165]. We will denote direct policy estimation using this model as DPE

Metrics

In this evaluation we are interested in comparing the prediction error of policies estimated by DPE, MCE and MCL. That is, the difference in the state distribution resulting from simulation of the inferred policies in the joint task POMDP and ground truth data needs to be quantified. For this purpose, the Kullback-Leibler divergence

$$\mathrm{KL}(p(d_t)||p'(d_t)) = \sum_{d_t=0}^{T} p(d_t)(\log[p(d_t)] - \log[p'(d_t)]), \quad (4.71)$$

was used to compare the distribution $p(d_t)$ of EOD in the ground truth data and the data obtained from the inferred policies. This metric has also been employed in [95].

The distributions of the states related to vehicle control were also compared by means of the Kullback-Leibler divergence. Here, first the state distribution was approximated by a Gaussian $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. Thereafter, the average Kullback-Leibler divergence for Gaussians

$$\mathrm{KL}^{\mathrm{G}}(p(\mathbf{x}_t^{\mathrm{p}})||p'(\mathbf{x}_t^{\mathrm{p}})) = \frac{1}{2T}\sum_{t=0}^{T} \mathrm{tr}[(\boldsymbol{\Sigma}_t')^{-1}\boldsymbol{\Sigma}_t'] + (\boldsymbol{\mu}_t - \boldsymbol{\mu}_t')^{\top}(\boldsymbol{\Sigma}_t')^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_t') \quad (4.72)$$

$$- \dim(\boldsymbol{\mu}_t) + \log[\det(\boldsymbol{\Sigma}_t')] - \log[\det(\boldsymbol{\Sigma}_t)] \quad (4.73)$$

was computed which was used as metric.

Finally, the reward parameters $\boldsymbol{\theta}'$ estimated by MCE-IOC and MCL-IOC were evaluated by the relative average deviation from the true $\boldsymbol{\theta}$,

$$\mathrm{RD}(\boldsymbol{\theta}, \boldsymbol{\theta}') := \frac{1}{n}(\sum_{i=1}^{n} |\boldsymbol{\theta}_i' - \boldsymbol{\theta}_i| / |\boldsymbol{\theta}_i|). \quad (4.74)$$

Protocol

In the evaluation the initial state $\mathbf{x}_0^{\mathrm{p}} = [0\,\mathrm{m}; 0\,\mathrm{m/s}; 0; 0]$, $d_0 = 0$ and a horizon T corresponding to 7s were used. We considered a driving situation, denoted Same, in which data was generated to both estimate parameters of the policy models and to evaluate the policy models. Furthermore, a different driving situation, denoted Trans, was used in which data was generated which was not available during inference and only used for evaluation. This had the purpose of investigating the transferability of the estimated parameters to previously unseen situation. Both driving situations Same and Trans were defined by specific external variables v_t, κ_t:

1. **Same:** Driving speed of $\mathbf{v}_{0:T} = 50\,\mathrm{km/h}$ and
   a moderate curve to the left on the motorway of $\boldsymbol{\kappa}_{0:T} = +1.4 \times 10^{-3}\,\mathrm{m}^{-1}$
2. **Trans:** Driving speed of $\mathbf{v}_{0:T} = 80\,\mathrm{km/h}$ and
   a moderate curve to the right on the motorway of $\boldsymbol{\kappa}_{0:T} = -1.4 \times 10^{-3}\,\mathrm{m}^{-1}$

To obtain training data, first the MCE policies for the reward parameters $\boldsymbol{\theta} = [-0.5\,\mathrm{m}^{-2}; -8\,\mathrm{s}^2\mathrm{m}^{-2}; -11; -200\,\mathrm{s}^2; 0.07; -3.5]$ were computed for both scenarios Same and Trans. Thereafter, these policies were used to perform 3000 roll-outs for both driving situations. For $k = 0, 1, \ldots, 10$, we then selected the first 2^k sequences of instance Same, applied DPE and estimated $\boldsymbol{\theta}$ using SRMCE-IOC and SRMCL-IOC. Here, the original $\boldsymbol{\theta}$ was used as initial guess for optimization in SRMCE-IOC and SRMCL-IOC. Finally, the policies (indirectly) inferred by DPE and the IOC methods were used to sample 1976 new sequences for **both** Same and Trans. We report the difference between the original state distribution and the distribution obtained after policy inference.

4.5.2 Results

The results of the evaluation on simulated data are summarized in the following. We report the medians of the metrics between the sampled behavioral data after estimation of policy parameters and all the original data of Same and Trans. Tab. 4.1 shows the error in predicting the three methods SRMCE-IOC, SRMCL-IOC and DPE. The errors in estimating the reward parameters $\boldsymbol{\theta}$ using the IOC methods SRMCE and SRMCL are summarized in Tab. 4.2 In both tables we indicated the best result, i.e. the least *median* error per condition by underlining.

Tabular 4.1: Simulated data evaluation

Num. Train. Seq.	**Metrics**	**Methods**					
		SRMCE		SRMCL		DPE	
		Same	Trans	Same	Trans	Same	Trans
2^0	KL^G	19.89	19.87	<u>19.39</u>	<u>19.54</u>	19.90	606.7
	KL	<u>0.0901</u>	<u>0.2332</u>	0.1219	0.2684	0.1645	0.4057
2^4	KL^G	19.29	19.42	<u>19.27</u>	<u>19.39</u>	19.35	686.3
	KL	0.0041	0.0057	<u>0.0018</u>	<u>0.0043</u>	0.1653	0.3495
2^8	KL^G	19.28	19.40	<u>19.22</u>	<u>19.39</u>	19.30	718.2
	KL	0.0014	0.0009	<u>0.0008</u>	<u>0.0006</u>	0.1746	0.3560

Tabular 4.2: Deviation from true reward

Methods	**Num. Train. Seq.**					
	2^0	2^2	2^4	2^6	2^8	2^{10}
SRMCE	0.723	0.367	<u>0.266</u>	<u>0.212</u>	0.226	<u>0.198</u>
SRMCL	<u>0.431</u>	<u>0.323</u>	0.279	0.236	<u>0.208</u>	0.208

Additionally, we depict the results of the evaluation in Fig. 4.3, Fig. 4.4 and Fig. 4.5.

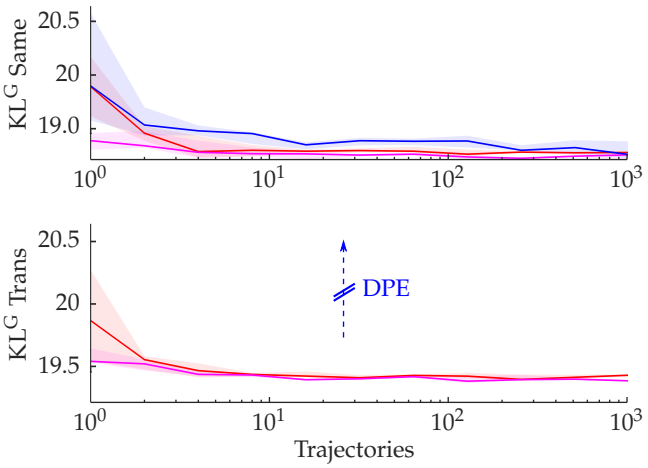

Figure 4.3: KL^G per number of trajectories. The medians are denoted by the line in red for SRMCE-IOC, in magenta for SRMCL-IOC and in blue for DPE. The shaded areas indicated the $[0.25, 0.75]$ interval for SRMCE-IOC, SRMCL-IOC and DPE. The first plot with the continuous lines shows the results for Same while the second plot with the shows the results for Trans.

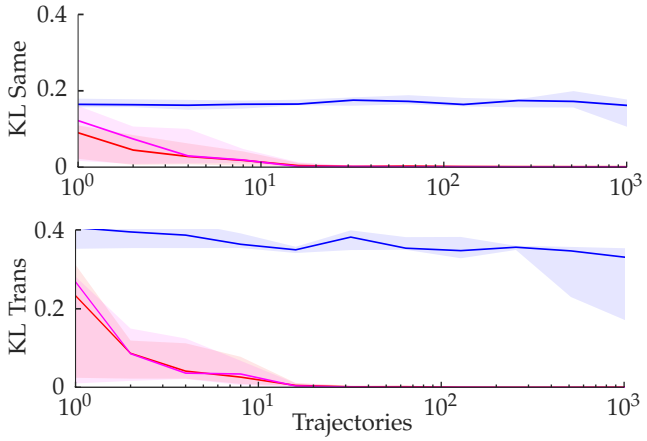

Figure 4.4: KL per number of trajectories (Legend in Fig. 4.3).

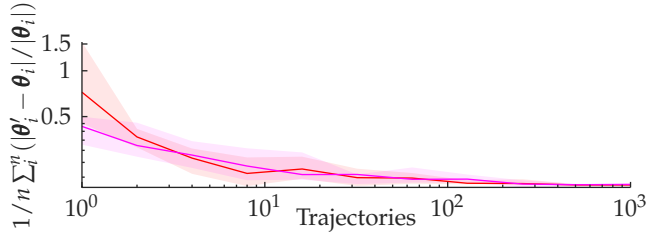

Figure 4.5: RD in *log-scale* per number of trajectories (Legend in Fig. 4.3).

The results for Same show significantly lower prediction error of the primary task states according to KL^G of the *best* IOC method compared to DPE on little data $p_{\text{test}} < 0.01$. The error in KL^G of DPE does not significantly differ from MCE on one trajectory and both have a higher KL^G than MCL $p_{\text{test}} < 0.01$. On a number of 24 to 128 trajectories samples DPE performed significantly worse than IOC $p_{\text{test}} < 0.01$ (larger KL^G). On more training data, no significant differences in performance were present between the individual methods. All methods approached a KL^G of 19.15 which is the average KL^G for sets of 1976 trajectories even under the true parameters.

In evaluation on the transfer scenario Trans all methods showed an increase in prediction error when trained on few data. However, whereas the prediction error in KL^G of DPE exploded, it only slightly increased in the case of the IOC methods. In addition to that, the IOC methods showed a decrease of

the prediction error in terms of KL^G to a lower bound of 19.5 which was attributed to the sampling error as in the case of Same.

In terms of predicting EODs, DPE constantly preformed significantly worse than both IOC methods in Same. The prediction of DPE did not decrease if more data were used for training, whereas the error in predicting EODs decreased in the case of SRMCE-IOC and SRMCL-IOC. Furthermore, no significant differences between the IOC methods could be established. In prediction in the transfer scenario Trans, errors significantly increased for all methods if less than 10 trajectories were used for training. When more trajectories were used for training no significant increase in prediction error in terms of KL of the IOC methods could be established.

4.5.3 Discussion

In the discussion of MCE and MCL in Sec. 4.3.2 we referred to the theorem of [257], that shows equivalence of MCE and MCL in case of infinite training data. The results of the present evaluation are in line with that theorem. SRMCE-IOC and SRMCL-IOC already started to converge for a comparably small amount of 10 trajectories. However, we note that the evaluation was based on simulated data that was obtained under exactly the same joint task model used in the IOC methods. Hence, the data distribution could accurately be reproduced.

For a small amount of training data, SRMCE-IOC and SRMCL-IOC differed in both the estimated reward parameters and the resulting state distributions. This is an important result, as the differences between MCE-IOC and MCL-IOC have not be considered in any previous work. In practical application where usually few trajectories are available similar difference may arise and hence both approaches should be evaluated. The significantly higher prediction error of SRMCE-IOC compared to SRMCL-IOC when trained on few trajectories can possibly be attributed to the usage of the empirical feature expectations. A single trajectory of 7 s contains 175 state control pairs in the model frequency of 25 Hz. Maximizing the likelihood of the policy as in the case of MCL-IOC might more efficiently use information present than first condensing the data into the empirical expectation as in the case of MCE-IOC.

Finally, the evaluation clearly showed the advantage of IOC over DPE especially in terms of sample efficiency and transferability of the inferred quantities: Whereas the recomputed policy for the inferred rewards has low prediction error in the transfer scenario, the directly estimated policy is unable to account for the adapted behavior. This is in line with other work on MCE for policy inference that found a similar advantage of IOC e.g. [2, 260, 118]. Admittedly, it also turned out that the policy for sensor switches under the MCE model is substantially more complex than the used policy model in DPE, hence a more elaborated baseline should be used in further evaluations.

4.6 A Real-Traffic Driving Experiment

We demonstrated the potential advantage of MCE approaches over DPE in the previous section. It was shown that policies estimated by DPE show a significant increase of prediction error on previously unseen driving scenarios. However, the evaluation was done by means of simulated data from exactly the same POMDP model that was also used in the IOC methods. Sec. 3.6.1 showed that the MCE policy of the joint task POMDP has similar fundamental properties as driver behavior in real traffic. Nevertheless, it is unlikely that all model parts perfectly match the relations underlying real-world behavior. Therefore, it is necessary to evaluate the prediction performance using data from real driving.

4.6.1 Protocol

To obtain data, we conducted a driving experiment on a segment of the German motorway A81 which is depicted in Fig. 4.6.
The reason for this choice were speed limits of $80, 90, 100, 110$ km/h in sub-segments which corresponded to the experimental conditions and a low traffic volume (see Fig. 4.7).
We decided against a study in a driving simulator. This is because of possible influences on the participants' behavior by absence of real risk. In addition to that, it is important to evaluate the robustness of prediction on realistic sensor input.

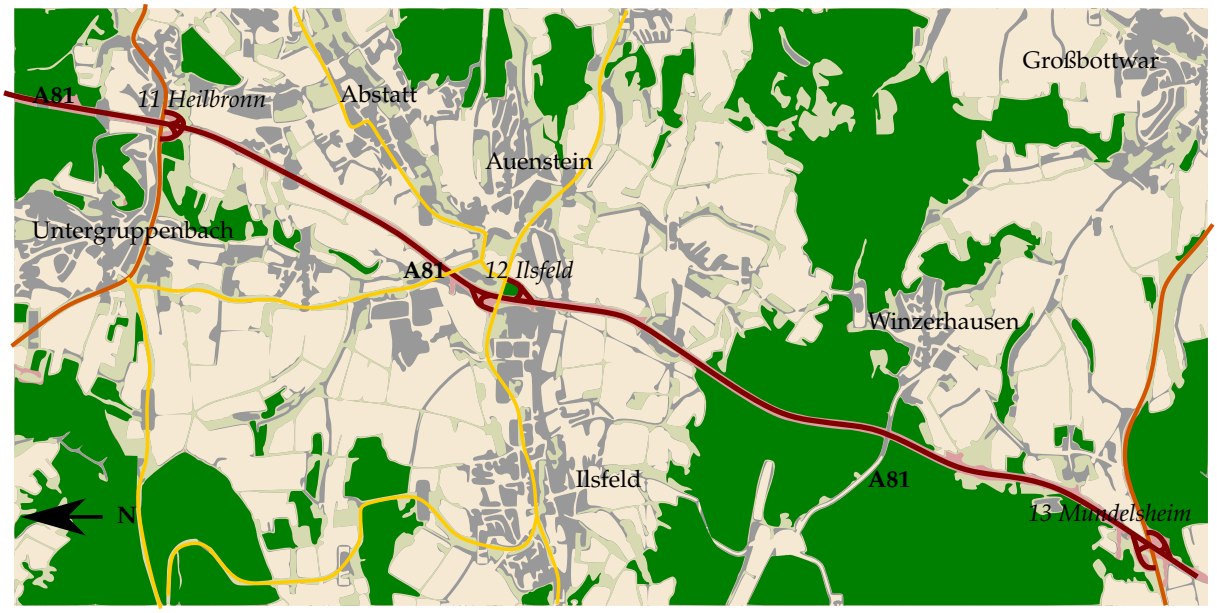

Figure 4.6: Segment of German motorway A81 used for the driving experiment. The recordings were conducted driving between exit *11 Heilbronn* and *13 Mundelsheim*. Obtained from [170], License CC-BY-SA 2.0

Figure 4.7: Impressions from the segment of the motorway A81 used for the driving experiment. The segment features moderate curves and low traffic. As it can be seen on the left picture the higher driving speeds of 100, 110 km/h were driven on the center lane.

For participants we recruited seven drivers (six male, one female) from the Robert Bosch Group. As part of the safety concept only drivers were selected that had previously taken a special in-house driving safety training. Using well-trained and experienced drivers to collect behavioral data is also beneficial for inverse optimal control. Although the MCE policy model accounts for potential sub-optimal behavior it is less suited to model highly erroneous behavior [59]. Hence, using selected drivers supports estimating consistent rewards.

The experiment consisted of four fixed driving speed conditions $\{80, 90, 100, 110\}$ km/h. Vehicle speed was controlled by the vehicle's Adaptive Cruise Control (ACC) to prevent drivers from adjusting their speed as a compensatory action while being engaged in a secondary task as e.g. observed in [12]. Furthermore, a conservative time gap was employed to ensure that the distance to preceding vehicles did have the least possible influence on the drivers' behavior. When the vehicle traveled at the required speed, the measurement periods were started. Such a period was either a reference, where the participants drove fully attentive or involved a visually distracting secondary task. At each speed three secondary tasks and three reference periods per participant were triggered by the investigator. The experimental conditions are summarized in Tab. 4.3.
As a secondary task we used the task of typing random numbers $\{1, 2\}$ that was described earlier

Tabular 4.3: Experimental Conditions of Driving Experiment I

Secondary Task	**Driving Speeds**			
	80 km/h	90 km/h	100 km/h	110 km/h
Reference	3×	3×	3×	3×
With Secondary Task	3×	3×	3×	3×

in Sec. 3.3.3. The random numbers were incrementally displayed in 3 rounds of 10 numbers and the drivers were in total required to type 30 numbers. The numbers were shown on a display at the position of the vehicle's central information display. This required significant aversion of gaze from the road scenery as can be seen in Fig. 4.8.

Figure 4.8: An example of averting gaze aversion to conduct the typing task. (Personal agreement of the depicted participant was obtained.)

The input buttons were implemented by a number pad which was placed next to the gearshift. The arrangement is depicted in Fig. 4.9.

Figure 4.9: Artificial secondary task used in the first driving experiment. Left picture shows the vehicle's interior, where the driver is reading the random numbers from the display and typing by means of a number pad. The right picture shows the presentation of the random numbers.

This artificial task was chosen as it resembles the principle of a variety of real visual-manual tasks performed while driving and possesses several advantages. First, the task state is fully measurable and can easily be modeled, in contrast to the vehicle's infotainment system. Second, the participants needed only little practice to reach maximum execution performance, resulting in no significant learning effects during the experiment.

Typically, drivers have a significant personal interest in the secondary task they are engaging into during vehicle control. Hence, to obtain realistic behavior the participants were instructed to *"perform the secondary task as quickly and correctly as possible while not endangering driving safety"*, as suggested in [6].

4.6.2 Recorded Data and Preprocessing

We used the MPC2 system (Robert Bosch GmbH, Stuttgart, Germany) for tracking the lane boundaries and recorded the position of the lane y_t, the angle between tangent of the lane boundary ϕ_t and the vehicle's longitudinal axis and the curvature κ_t via the vehicle's Controller Area Network (CAN). A SmartEye Pro (SmartEye AB, Gothenburg, Sweden) three-camera infra-red eye-tracking system with active illumination was used to estimate the driver's gaze direction. The cameras were positioned at the left a-column, in front of instrument cluster and on the dashboard above the display (see Fig. 4.10).

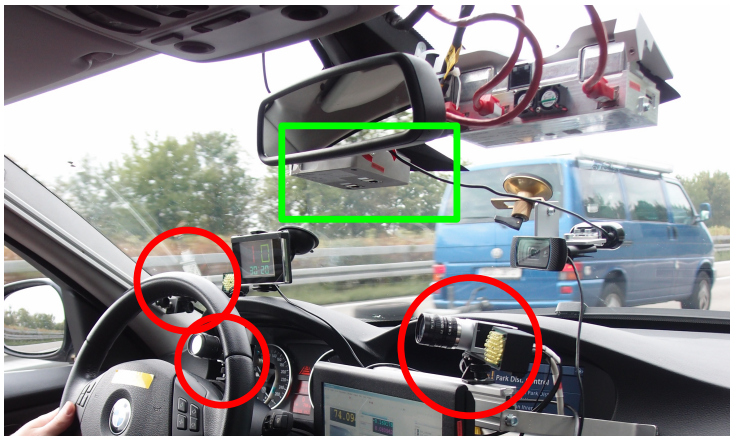

Figure 4.10: Sensors used in experiment I: Red circles denote the cameras of the eye-tracking system. Green rectangle highlights the camera used for tracking the lane boundaries.

Steering wheel position and velocity as well as absolute velocity measured by standard in-vehicle sensors were also recorded via the CAN. Hence, beside the eye-tracking systems we relied solely on signals that are already accessible in today's series-production vehicles.

In order to ensure sufficient quality for numerical evaluation, pre-processing and filtering steps were performed on the collected raw data.

Selection of Valid Trials We automatically excluded lane changes and their preparation phases. Being a different driving maneuver than lane keeping it requires a different driving and gaze policy. Also situations where the ACC controller reduced the vehicle speed by more than 10% were left out due to possible influence on the drivers' behavior. The final data set consisted of 136 valid segments comprising of 53 reference and 83 secondary task periods with an average duration of $\approx$ 50 s.

Sub-sampling and Filtering of Vehicle Signals As the used sensors operate on different frequencies, e.g. the eye-tracking on 60 Hz but the lane-tracking on 25 Hz, we sub-sampled all signals to 25 Hz. Thereafter, the Rauch-Tung-Striebel-smoother [186] was employed for filtering of the partially low resolution signals $y_t, \phi_t, \alpha_t, \dot{\alpha}_t, \kappa_t$ using the kinematic vehicle model introduced in Sec. 3.3.1. The steering angle transmission ratio c was estimated by means of least squares regression while the covariance of the noise in the vehicle model $\mathbf{\Sigma}^{\epsilon^{\mathrm{p}}}$ was estimated by expectation-maximization as suggested for vehicle models by [222].

Pre-processing of Eye-Tracking Data The eye-tracking data was pre-processed by first detecting whether the gaze of the driver was on the road or off the road. This was done by detecting intersections of the eye gaze direction vector with a rectangular region one meter in front of the driver. This region spanned across the forward road scenery similar as in [108] (further details are reported in Cpt. 6). The sensor states x_t^{z} that resulted from this approach were cleansed of jitter by removing all intervals of the same sensor state that had a duration less than 0.1 s. The sensor state of those intervals were set to the preceding sensor state. Finally, all sensor state data was manually checked and remaining errors were removed.

4.6.3 Behavioral Statistics

Before we present the numerical evaluation on the behavioral data, we wish to report on key statistics of the behavior of the participants. These statistics were analyzed to validate that the driving experiment produced realistic driver behavior suitable for evaluation of the inference procedures.

A relevant quantity is the distribution of duration of glances off the road which we will denoted as $\max(d_t)$. The rationale behind this notation is that the duration of a glance off the road is a local maximizer of the eyes-off duration. Furthermore, we also investigated the lane keeping performance as measured by the Root Mean Squared Error (RMSE) of the lane position (deviation from the lane center) and the Standard Deviation (STD) of the lane position. As mentioned before, these metrics of the lane position have previously been used to quantify the effects of distraction [252].

Before we start reporting and analyzing the participants' behavior we wish to comment on the employed methodology. The purpose of the driving experiment was to collect data for evaluating inverse optimal control approaches and baselines. Here, a part of the data is used to estimate the model parameters, e.g. the reward parameters $\boldsymbol{\theta}$ in IOC. This is done *without personalization*. That is, a single global parameter vector is inferred for all participants. Against this background, the present section also analyzes the participants' behavior on global scale without considering identities. For example, difference between the distribution of glance durations of all participants in the experimental conditions are investigated. Consequently, statistical test conducted in this sections will assume that the distributions of those quantities are independent across experimental conditions. Note that quantities of participants' behavior (such as glance durations) in the different experimental conditions are positively correlated, hence treating them as independent can only increase p-values of the tests.

As can be seen in Fig. 4.11, the secondary task used in the experiment resulted in a strong concentration of $\max(d_t)$ around the medians which were at 1.1, 1.2, 1.0, 1.1 s at the speed conditions 80, 90, 100, 110 km/h. Furthermore, mean durations of glances off the road of 1.4, 1.4, 1.1, 1.2 s were present.

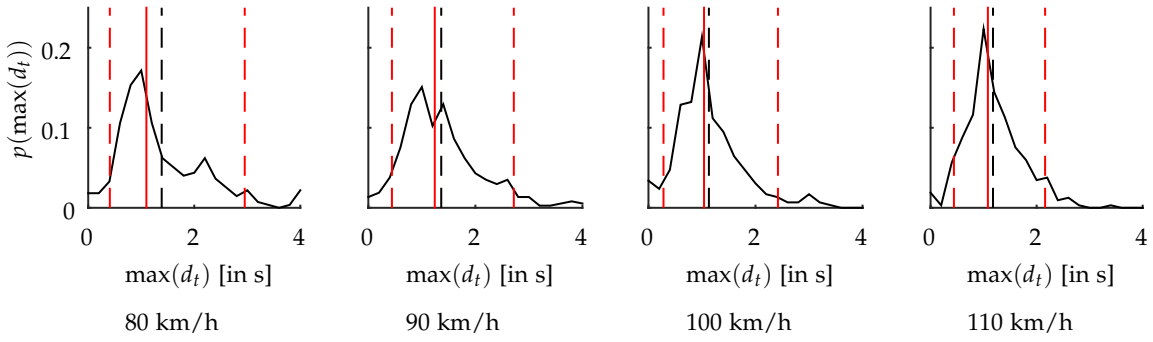

Figure 4.11: Distributions of the durations of glance of the road denoted as $\max(d_t)$ at the different driving speeds. Dashed red lines indicate the $[0.05, 0.95]$ quantiles, while the solid red line indicates the median. The mean maximum glance duration is denoted by a dashed black line.

Considering the quantiles in the speed conditions, which are depicted in Fig. 4.12, the following observations were made. In the quantiles below 0.75 no significant differences among the driving conditions were present. In contrast, the 0.75 quantiles 3 s, 2.7 s, 2.5 s, 2.1 s showed a significant decrease ($p_{\text{test}} < 0.01$) of each of the lower speeds 80 km/h and 90 km/h compared to each of the higher speeds of 100 km/h to 110 km/h according to the quantile test of [86]. At the 0.95 quantile a significant monotonous decrease could be established ($p_{\text{test}} < 0.01$). With respect of the means a significant decrease between the groups 80, 90 km/h and 100, 110 km/h could be established ($p_{\text{test}} < 0.01$). All other differences turned out to be not significant.

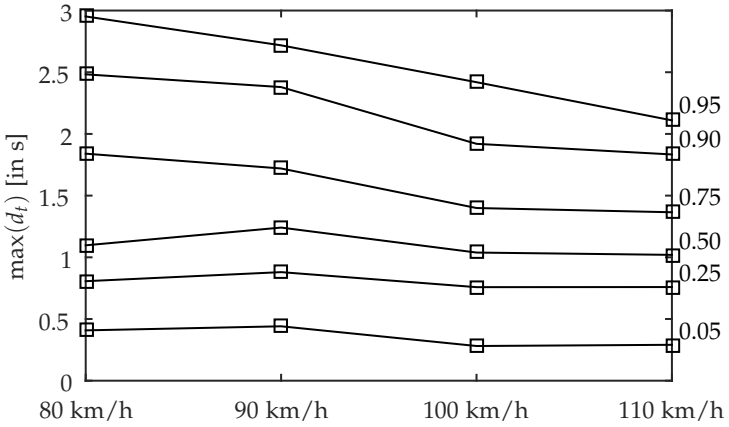

Figure 4.12: Quantiles of the durations of glances off the road $\max(d_t)$ in the driving experiment I. Plot shows the $0.05, 0.50, 0.75, 0.90, 0.95$ quantiles.

Regarding lane keeping performance effects of the secondary task engagement could be established: As can be seen in Fig. 4.13, the standard deviation of the lane position showed a small but significant increase from a median of 0.158 m during driving without a secondary task to a median of 0.165 m during driving in presence of the secondary task according to a Wilcoxon rank sum test ($p_{\text{test}} = 0.01$). The RMSE of the lane position was more sensitive with respect to the presence of the secondary task, which was present in a highly significant increase from 0.239 m to 0.274 m ($p_{\text{test}} \ll 0.01$).

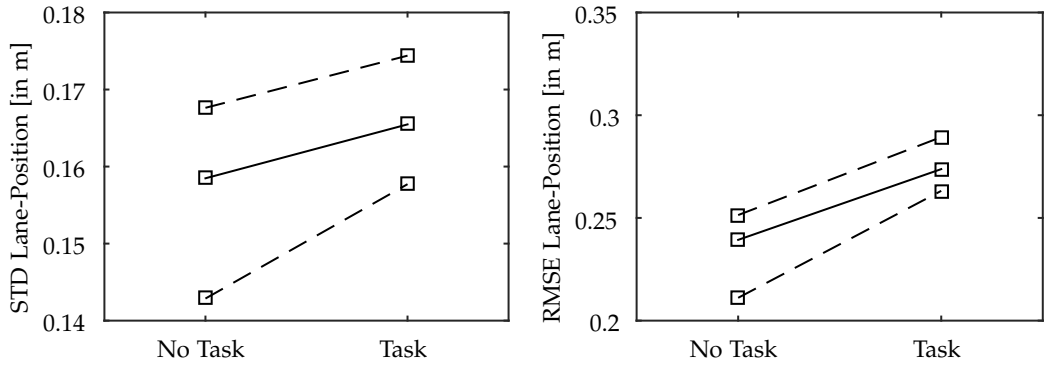

Figure 4.13: Left plot depicts the statistics of the standard deviation of the lane position, right plot shows the statistics of the root mean squared deviation from the lane center. Solid line indicates the median of the statistic, whereas the dashed lines indicate the $[0.05, 0.95]$ confidence interval.

Finally, the time required to complete the secondary task strongly concentrated around the median of 28 s which was close to the 0.05 quantile of 24 s. The distribution of the durations of the secondary task is depicted in Fig. 4.14.

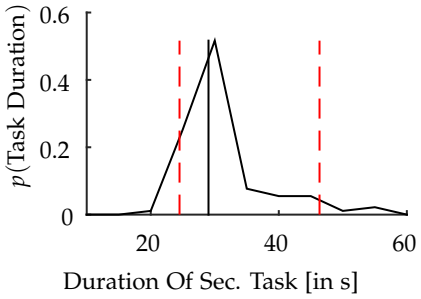

Figure 4.14: Distribution of the time required by the participants to correctly type 30 random digits 1, 2. Dashed red lines indicate the $[0.05, 0.95]$ quantiles, while the solid black line indicates the median.

4.6.4 Discussion

The distribution of the duration of glances off the road is well in line with those obtained in previous real traffic driving experiments. For example, [245] found a similar distribution with means ranging from 0.9 s to 1.4 s depending on the difficulty of the secondary task. In contrast to that work, in our case the percentage of glances off the road whose duration exceeded 2 s was higher then 0.25 even for 110 km/h. However, this may be explained by the fact that the participants were experienced drivers.

[245] showed that mean and median maximum glance duration are sensitive to secondary task difficulty. Hence, we hypothesize that the influence of the chosen secondary task dominates over the influence of the speed conditions with respect to the medians and means. A strong influence of the speed condition only in the higher quantiles is plausible considering previous occlusion experiments [207, 69]. Here, speed dependence was found in the median *maximum* occlusion time the participants tolerated.

Finally, similar effects of secondary task on lane keeping performance have been found in other works. Similar small effects of the presence of a secondary task on the STD were also found in the naturalistic driving experiment of [55]. In [174] the standard deviation of the lane position was 0.1 m for attentive driving and 0.15 m for driving in presence of a secondary task. The differences observed in the absolute values observed in our driving experiment compared to that study may be explained by the following fact. The data of [174] contained a high proportion of driving on rural roads. Compared to driving on the motorway this is associated with smaller lanes and lower speeds. These aspects are both likely to decrease the overall STD of the lane position. Interestingly, in contrast to the STD the RMSE of the lane position was sensitive to the presence of the secondary task. However, we currently lack a sufficient explanation of this effect.

As a result of this discussion we can conclude that the driving experiment was successful in providing realistic data of adaptive driver behavior in engagement in a secondary task. The behavioral statistics in both the distribution of durations of glances off the road as well as the distribution of lane positions were comparable to those obtained in other experiments. In addition to that, the drivers significantly adapted the tail of the duration of long glances to the experimentally imposed different driving speeds.

4.7 Evaluation on Real Traffic Data

Given the behavioral data obtained in the previously described real traffic experiment IOC methods for policy and reward inference can realistically be evaluated. At this point, we would like to refer back to the purpose of this estimation problem. The normative model of glance behavior developed in the previous chapter requires to specify the reward parameters that is accepted by the drivers. Furthermore, the driver's likely future behavior must be considered when defining appropriate glance behavior. Inverse optimal control is a technique to obtain all reward parameters and a realistic policy based on behavioral data. This is done by searching for reward parameters such that the corresponding policy reproduces the observed behavior. In case of the maximum causal entropy model we can obtain

both the normative reward parameters and a realistic sub-optimal policy model under the assumption that the observed drivers show at least rational behavior.

The goal of incorporating a realistic model of driver behavior in computing appropriate glance behavior can be evaluated using the behavioral data. This is possible by investigating how well the policies obtained by IOC predict the observed behavior. By comparison of the prediction error of the inferred policies to the errors of established behavioral models, additionally, the assumptions made in the development of the joint task POMDP can be assessed. It should be noted that evaluating prediction errors does not directly relate to the *acceptance* of the reward parameters when used in a warning system. This is only the case if drivers desire appropriate glance behavior that is close to their own behavior. In the context of the behavior of autonomous cars this hypothesis has recently been questioned [21]. Therefore, later in this thesis a user test is conducted that directly assesses the acceptance by the driver (see Sec. 6.4). Nevertheless, it is a reasonable objective to require that the computed gaze switching policies are consistent with respect to the driver behavior. That is, that the policies consider the influences of the driving situation on driver behavior which has been shown to increase system effectiveness [195]. This aspect can be evaluated by assessing the errors when predicting behavior in a driving scenario unseen during training. If the prediction performance in the unseen scenario is significantly worse than the performance on the seen scenario, then important influences of the driving situation are missing in the model.

4.7.1 Scenario

For these reasons the obtained behavioral data is used to compare the prediction performance using the IOC approaches to the prediction performance obtained by applying DPE. We consider both a general assessment of prediction errors as well as an analysis of the errors in transfer to unseen driving situations.

Inverse Optimal Control Methods

In this evaluation we considered the maximum causal entropy inverse optimal control for the joint task POMDP under sensor model restriction *and* the simple secondary task, i.e. SRMCE-IOC and SRMCL-IOC. Here, the parameters of the dynamics model of the driving task were set to those values previously inferred in the preprocessing of the data (Sec. 4.6.2). The reward features on the primary task states (3.9) were used as in the evaluation on simulated data (Sec. 4.5.1). Furthermore, we assumed that the driver can fully sense all primary task states when his gaze is on the road, i.e. $\hat{\mathbf{\Sigma}}_0^{\mathrm{p}} = \mathbf{0}$. Similar to the evaluation on simulated data (Sec. 4.5.1) a sensor noise covariance of $\mathbf{\Sigma}^{\epsilon^z}(x_t^z = 0) = [\infty; 0; \infty; 0]$ was used. It seemed reasonable to assume that the drivers obtained no information from the road scenery when conducting the secondary task as it required significant gaze aversion which can be seen in Fig. 4.8.

Although previously algorithms for other policy models and other variants of the POMDP model have been derived, those were not considered in the evaluation. This is due to the following reasons: First, using IOC under the optimal policy in the same POMDP model, i.e. SRopt, offers no advantage over the MCE policy model. This is because the MCE policy can closely approximate any optimal policy using high values of the reward parameter (Sec. 2.55). Second, STRopt-IOC could not be evaluated due to infeasible computational demand. For both, computing optimal policies and computing gradients, tractable algorithms STRopt Algo. 6 and STRopt-IOC Algo. 14 were derived. The previous analysis 3.6.2 showed high computational complexity and computational demands of the policy computation according to STRopt. Inverse optimal control requires to compute optimal policies at each iteration of the gradient-based approach for minimizing the gap. In this specific evaluation, additionally several different instances of the POMDP had to be solved at each iteration. This is because the individual periods recorded in the experiment all differed in track topology and driving speed. Hence, applying IOC using STRopt on the data of the driving experiment is not feasible. Finally, we considered only the simple secondary task in the joint task POMDP.

To prevent over-fitting of the reward parameters we regularized their absolute values. This was done by adding terms

$$0.01 \sum_{i=1}^{n} \left| \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}_i(x_t, u_t) \big| \mathrm{D} \Big] \right| |\boldsymbol{\theta}_i| \tag{4.75}$$

to the objectives of the minimization problems involved in SRMCE-IOC (4.45) and SRMCL-IOC (4.46). Similar as in the evaluation on the simulated data Sec. 4.5.1 a barrier function was used to ensure a well-defined POMDP. Finally, a relative gradient norm of $\leq 10^{-6}$ was used as a termination criterion. For convenience, we abbreviate SRMCE-IOC and SRMCL-IOC as MCE and MCL in figures and tables throughout the evaluation.

Direct Policy Estimation Baselines

For comparison of prediction performance of the policy resulting from IOC, a generic DPE baseline (DPE1) and a DPE baseline using established behavior models for human attention allocation and foresighted steering (DPE2) were employed.

Generic Baseline (DPE1) As a first simple baseline the generic policy model previously introduced in Sec. 4.5.1 was used.

Baseline Using Established Beahvioral Models (DPE2) Furthermore, the following baseline was considered: [94, 95] presented a model for gaze-allocation in visual dual-tasking, where the probability of a gaze switch to a task is a logistic function of the uncertainty in its states. In our case uncertainty is only present in the vehicle states - the random number is either known to the drivers if he/she has seen it on the display or unknown otherwise. Therefore, we applied the following variant of the original approach

$$p(u_t^{\mathrm{z}} | x_t^{\mathrm{z}} = 1) = \frac{\exp(\lambda_1^{\mathrm{base}}) + u_t^{\mathrm{z}}}{\exp(\lambda_1^{\mathrm{base}}) + 1} \tag{4.76}$$

$$p(u_t^{\mathrm{z}} | x_t^{\mathrm{z}} = 0, \boldsymbol{\Sigma}_t^{\mathrm{p}}) = \frac{\exp(\lambda_2^{\mathrm{base}} + \mathrm{tr}(\boldsymbol{\Lambda}_3^{\mathrm{base}} \boldsymbol{\Sigma}_t^{\mathrm{p}})) + u_t^{\mathrm{z}}}{\exp(\lambda_2^{\mathrm{base}} + \mathrm{tr}(\boldsymbol{\Lambda}_3^{\mathrm{base}} \boldsymbol{\Sigma}_t^{\mathrm{p}})) + 1}. \tag{4.77}$$

with parameters $\lambda_1^{\mathrm{base}}, \lambda_2^{\mathrm{base}}, \boldsymbol{\Lambda}_3^{\mathrm{base}}$.

The two-point-steering model of [197] was applied to model human foresighted steering, including curve negotiation. Here, it is assumed that the driver's steering policy builds on a visual near-angle β_t^2 and a visual far-angle β_t^2. Given a lane with width w the visual near-angle β_t^1 is defined as the angle between a line from the vehicle center to a point on the road center typically 2 m ahead and the vehicle's longitudinal axis. The far-angle β_t^2 is defined as the angle of minimal magnitude of the angles between the tangents from the vehicle's center to both lane boundaries and the vehicle's longitudinal axis. Both angles are illustrated in Fig. 4.15. using an arc approximation of the track.

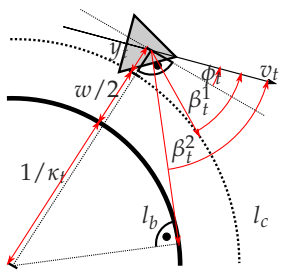

Figure 4.15: Illustration of the variables of the two-point steering model. Here, it is assumed that the track can closely be approximated by means of an arc.

Given collected data, the track topology of the entire recorded periods is available in form of the measured curvature $\boldsymbol{\kappa}_{0:T}$. This can be exploited to precisely compute the near and far angles [78]. However, we found that the track could sufficiently accurately be approximated by an arc using only the current curvature κ_t. In this case both angles of the two-point steering model β_t^1, β_t^2 can be computed in closed form using

$$\beta_t^1 = -\arctan\left(y_t \left(2\,\mathrm{m}\right)^{-1}\right) - \phi_t \tag{4.78}$$

$$\beta_t^2 = \begin{cases} -\arccos\left(\left(1 - \kappa_t \left(w/2 + y_t\right)\right)^{-1}\right) - \phi_t & \kappa_t < 0 \\ +\arccos\left(\left(1 + \kappa_t \left(w/2 - y_t\right)\right)^{-1}\right) - \phi_t & \kappa_t \geq 0 \end{cases}. \tag{4.79}$$

We used the computed near and far angle as well as the steering angle to define the policy

$$\pi_t^{\text{base}}(\dot{\alpha}_t | \beta_t^1, \beta_t^2, \alpha_t) \colon \dot{\alpha}_t = \lambda_4^{\text{base}} \beta_t^1 + \lambda_5^{\text{base}} \beta_t^2 + \lambda_6^{\text{base}} \alpha_t + \epsilon_7^{\text{base}}, \tag{4.80}$$

with parameters $\lambda_4^{\text{base}}, \lambda_5^{\text{base}}, \lambda_6^{\text{base}}$ and a normally-distributed random variable ϵ_6^{base}. This policy differs from the one used in the original work [197] where a proportional-differential (PD) policy was applied. However, this approach turned out to be problematic in the evaluation on real traffic data. The reason is that the curvature measured by the lane tracking often oscillated around 0. This corresponds to an "almost" straight lane which is quite usual on a motorway. However, as a result of the oscillations in κ_t jumps in the far-angle β_t^2 were frequent. These finally led to shaky behavior of the PD controller and loss of control.

The barrier model and the two-point model were linked by replacing y_t, ϕ_t in (4.78), (4.79) with their expectations if the sensor state x_t^z was 0 i.e. when the driver's gaze was off the road. Note, that this approach is very similar to the model applied in [94]. We will denote the resulting policy model as DPE2.

The parameters of the baselines were inferred by restating the models as generalized linear models [165]. Thereafter standard methods for fitting the parameters under $L1$ regularization were used.

Metrics

Similar as in the evaluation in Sec. 4.5.1 the Kullback-Leibler divergence $\mathrm{KL}(p(d_t)||p'(d_t))$ was used to measure the difference between the distributions of EOD in the experimental data and the predicted distribution of EOD.

In contrast, the expected squared error between the true lane position y_t and predicted lane position y_t',

$$\mathrm{SE}(y; \boldsymbol{\pi}_{0:T}) = \mathbb{E}\big[\frac{1}{T}\sum_{t=0}^{T}(y_t' - y_t)^2 | \boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\big] \tag{4.81}$$

was used in the present evaluation. The reason is that the Kullback-Leibler divergence for Gaussians is not defined for single trajectories y_t. As all other states of the kinematic model affect the lane position, the squared error of the lane position is a suitable metric for evaluation of prediction performance.

Protocol

For the numerical evaluation we further subdivided all valid segments of the driving experiment into snippets of $\approx 5\,\mathrm{s}$ (overlap $\approx 2.5\,\mathrm{s}$). This was to account for a realistic prediction horizon in a real-time system. In the evaluation a part of the data was used for (indirectly) estimating policies, whereas a different part was used for evaluation of the prediction performance.

For DPE all training data was merged into a single set on which parameters were inferred. In case of the IOC methods, first for every snippet in the training set the corresponding POMDP model was generated. That is, the values of the first states in the data were used as initial distribution and the driving speed as well as the lane curvature were considered in the POMDP model. Thereafter, the reward parameters were inferred.

The assessment of prediction quality on the test set was done in the following way: We first generated the POMDP models corresponding to each snipped. Then the rewards inferred by the IOC approaches were used to obtain the specific maximum causal entropy policies in the POMDPs. Finally, we used the policies obtained by the DPE and IOC approaches and the POMDP models to sample 100 sequences. Based on these sequences the prediction error was estimated. The evaluation protocol consisted of two distinct evaluations in total:

Overall Prediction Performance We first evaluated the overall prediction performance by splitting the data set into a training set and a test set of equal size randomly and independently of driver, velocity and track-topology. Afterwards the roles of the data sets were swapped. To more precisely estimate the error statistics this 2-fold cross-validation procedure was repeated 10 times.

Transfer Performance To investigate the generalization quality on unseen velocities we conducted a second evaluation. Here we trained on a random selection of half of the data of one single speed condition. We thereafter tested on the remaining half with the same and on all data of other velocities. In this evaluation we performed 5 repetitions.

4.7.2 Results

The results of the evaluation of the overall prediction performance and the transfer performance are summarized in following. In the evaluation the prediction errors followed skewed distributions which can e.g. be seen in Fig. 4.19. Hence, we report on the median errors over the considered snippets.

We first evaluated the overall prediction performance (introduced in Sec. 4.7.1), whose results are depicted in Fig. 4.16 and in Fig. 4.17. As shown in Tab. 4.4 the IOC generally approaches showed a smaller median prediction error in both the SE and the KL metric.

Tabular 4.4: Overall Prediction Performance

Metrics	**Methods**							
	DPE1		DPE2		MCL		MCE	
	Train	Test	Train	Test	Train	Test	Train	Test
SE	0.095	0.096	0.048	0.048	0.021	0.021	<u>0.015</u>	<u>0.015</u>
KL	0.110	0.109	0.100	0.100	<u>0.072</u>	<u>0.073</u>	0.074	0.075

In none of the evaluated methods a significant difference between the prediction errors on the test and the training set could be established ($p_{\text{test}} > 0.01$). Hence, adequate regularization has been employed. With respect to the SE metric the following performance differences were significant ($p_{\text{test}} < 0.01$): DPE1 had a higher median SE than DPE2, DPE2 had a higher median SE than MCL and finally MCL had a higher median SE than MCE. In the KL metric the difference between both baselines were small but still DPE2 had a significantly smaller prediction error than DPE1 ($p_{\text{test}} \approx 0.01$). Furthermore, both IOC methods showed significantly ($p_{\text{test}} \ll 0.01$) lower prediction error than the baselines. However, no significant differences between the IOC method could be established ($p_{\text{test}} > 0.01$).

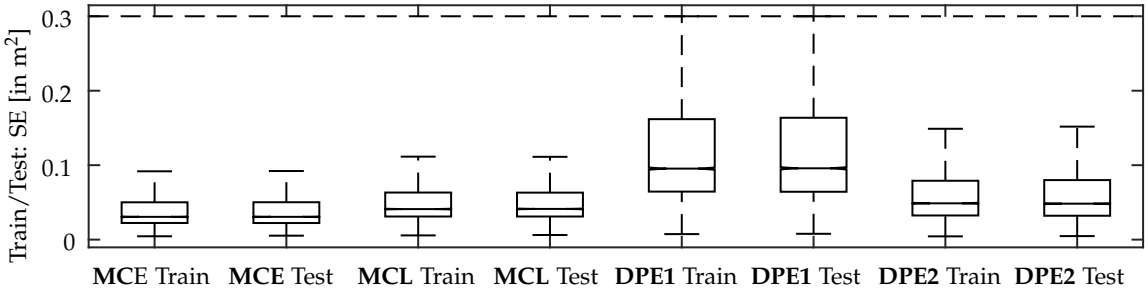

Figure 4.16: Expected squared error in prediction of the lane position. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. *Train* denotes prediction errors on the data set used for training, *Test* denotes prediction errors on the held out data.

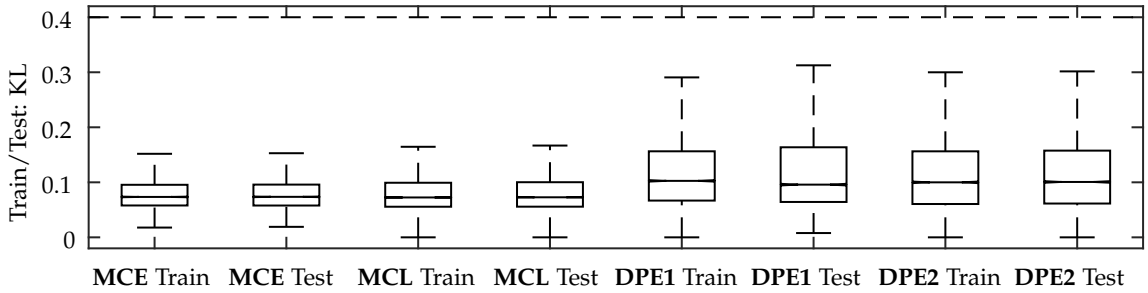

Figure 4.17: Difference of true distribution and predicted distribution of EOD measured by the Kullback-Leibler divergence. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. *Train* denotes prediction errors on the data set used for training, *Test* denotes prediction errors on the held out data.

The results of the evaluation of the transfer performance (Sec. 4.7.1) are shown in Fig. 4.18 and in Fig. 4.19. Tab. 4.5 further summarizes the obtained prediction errors.

Tabular 4.5: Transfer Performance

Metrics	**Methods**							
	DPE1		DPE2		MCL		MCE	
	Same	Trans	Same	Trans	Same	Trans	Same	Trans
SE	0.088	0.097	0.040	0.052	0.021	0.021	<u>0.015</u>	<u>0.015</u>
KL	0.098	0.118	0.112	0.118	<u>0.073</u>	<u>0.073</u>	0.075	0.075

Both IOC method show a significantly $p_{\text{test}} < 0.01$ smaller increase of the prediction error in the transfer than both the DPE baselines. This was verified by conducting signed rank test on the differences of the medians of both metrics SE and KL. Between both IOC methods no significant differences could be established $p_{\text{test}} > 0.01$. In contrast, the transfer performance of the DPE baselines differed: DPE1 showed a small but significant better performance than DPE2 with respect to SE $p_{\text{test}} \approx 0.01$, while the increase of prediction error DPE2 was smaller than the increase of error in DPE with respect to the KL $p_{\text{test}} \approx 0.01$.

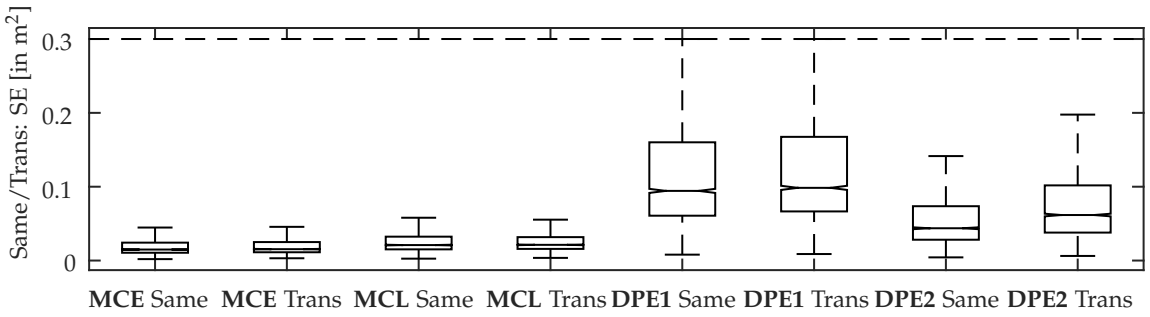

Figure 4.18: Expected squared error in prediction of the lane position. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. *Same* denotes prediction errors on the held out data set of the same speed, *Trans* denotes prediction errors on unseen speeds.

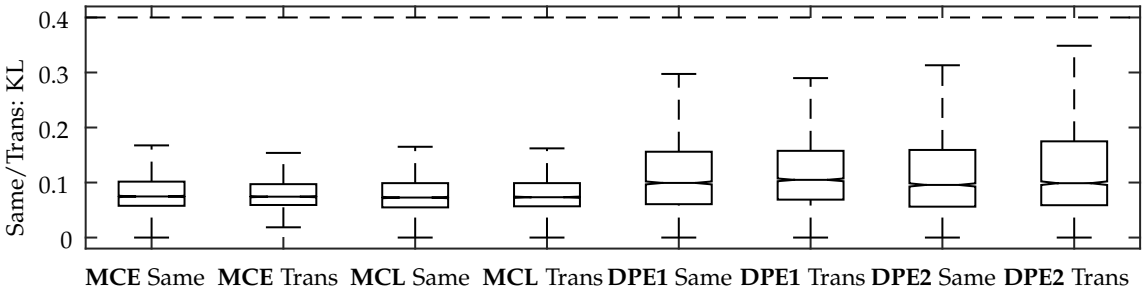

Figure 4.19: Difference of true distribution and the predicted distribution of EOD measured by the Kullback-Leibler divergence. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. *Same* denotes prediction errors on the held out data set of the same speed, *Trans* denotes prediction errors on unseen speeds.

Breaking down the prediction performance in Tab. 4.6 reveals more insights into the difference between the considered DPE2 and SRMCE-IOC. SRMCE-IOC turned out to result in smaller prediction error in both metrics and in all evaluation conditions except for the KL in training on 110 km/h and test on 100 km/h. Furthermore, a Kruskal-Wallis test revealed significant ($p_{\text{test}} < 0.01$) between-condition variations in both methods and both metrics. Consequently, the prediction performance showed large variations in all methods and metrics which can also be seen in Fig. 4.18 and Fig. 4.19. However, in the case of SRMCE-IOC the variation was significantly smaller than in the case of DPE2, which was verified by a Ansari-Bradley test ($p_{\text{test}} < 0.01$).

Tabular 4.6: Breakdown of Transfer Performance

Train Speeds	Metrics	Methods	Test Speeds			
			80 km/h	90 km/h	100 km/h	110 km/h
80 km/h	SE	MCE	0.0154	0.0143	0.0164	0.0154
		DPE2	0.0293	0.0572	0.0725	0.0836
	KL	MCE	0.0851	0.0744	0.0738	0.0701
		DPE2	0.1280	0.1180	0.1025	0.1216
90 km/h	SE	MCE	0.0152	0.0148	0.0159	0.0163
		DPE2	0.0310	0.0365	0.0477	0.0627
	KL	MCE	0.0833	0.0748	0.0730	0.0712
		DPE2	0.1307	0.1146	0.1018	0.1200
100 km/h	SE	MCE	0.0147	0.0160	0.0155	0.0161
		DPE2	0.0311	0.0390	0.0450	0.0584
	KL	MCE	0.0838	0.0763	0.0731	0.0693
		DPE2	0.1374	0.1124	0.1100	0.1173
110 km/h	SE	MCE	0.0157	0.0152	0.0179	0.0177
		DPE2	0.0401	0.0408	0.0479	0.0603
	KL	MCE	0.0825	0.0724	0.0755	0.0721
		DPE2	0.1249	0.1076	0.0901	0.1102

Finally, we present some anecdotal evidence from the evaluation. Fig. 4.21 and Fig. 4.20 depict the trajectory of the lane position y_t and the distribution of the sensor state x_t^z as well as the predictions of DPE and IOC for a snippet of secondary task interaction at 90 km/h. As can be seen in the histogram of the sensor state x_t^z, the driver had his gaze off the road at approximately 89 percent of the time. This was fairly well predicted by the MCE policy, whereas the barrier model included in DPE2 predicted a significantly smaller proportion of gaze off road. When predicting the trajectory of the lane position y_t both DPE2 and SRMCE-IOC made significant errors at the end of the prediction horizon of 5 s. However, in total the prediction of the MCE policy is better than that using the two-point steering model. Notably, the latter approach falsely predicts a significant risk of lane departure.

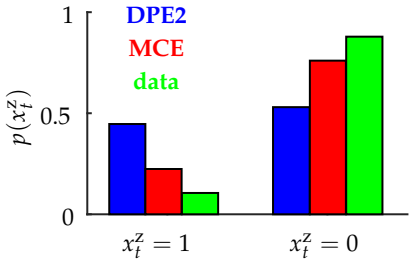

Figure 4.20: Histogram of the distribution of sensor states x_t^z for a snippet of secondary task interaction at 90 km/h. Blue bars indicate the histogram resulting from prediction under **DPE2**, red bars indicate the histogram under prediction under **MCE** and the green bar indicates the actual distribution of sensor states shown by the driver.

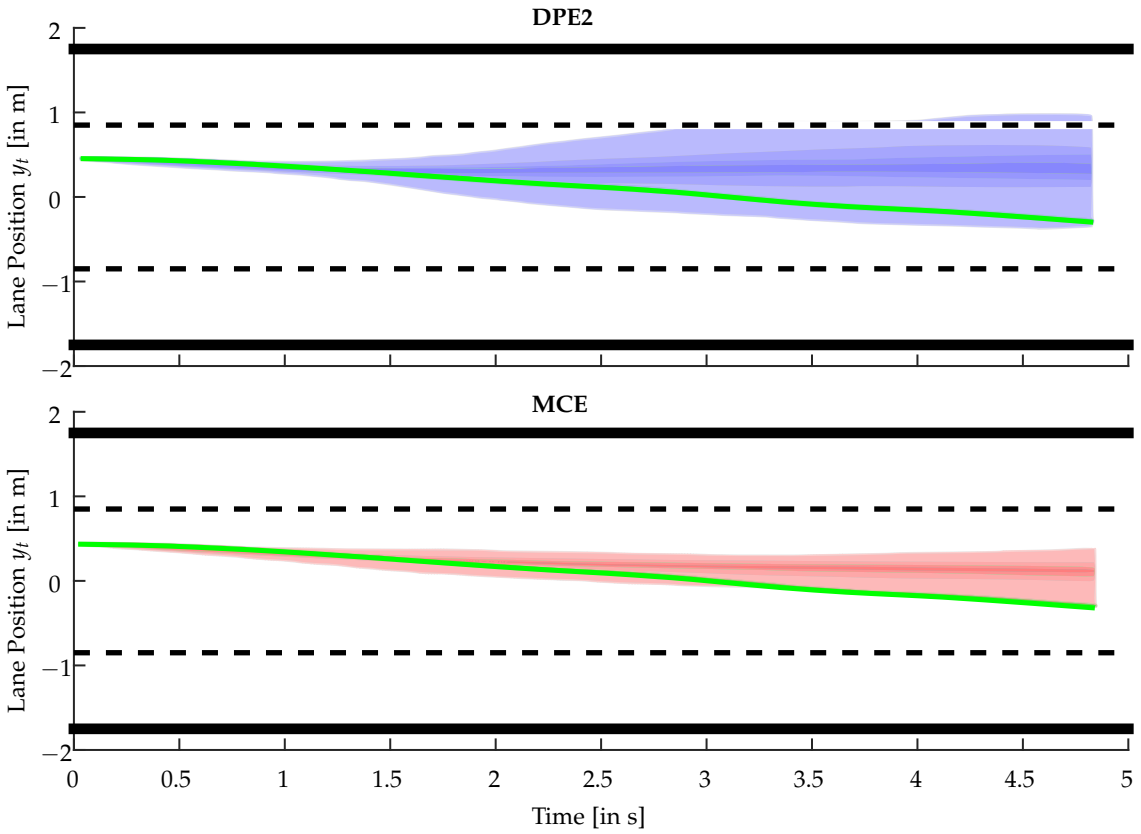

Figure 4.21: Trajectories and trajectory distribution of the lane position y_t for a snippet of secondary task interaction at 90 km/h. The thick black horizontal lines in the plot indicate the lane boundaries which were at approximately ± 1.75 m. The dashed horizontal lines indicate the lane position where the first wheel of the vehicle crosses the lane boundaries. The trajectory driven by the observed driver is denoted by a green line. The first plot shows the trajectory distribution predicted using **DPE2** in blue. The second plot shows the trajectory distribution predicted using **MCE** in red.

4.7.3 Discussion

Comparing the results of the evaluation on simulated data (Sec. 4.5) to the results of the evaluation on real data (Sec. 4.7) the following observations can be made. First, the advantages of IOC over DPE shown on simulated data are reproduced on the real data. However, in the present evaluation prediction errors of policy obtained by DPE did not explode. We think this can largely be attributed to the fact that the real data set is more diverse, which prevents policy overfitting in DPE. A large amount of data (136 of 50s periods obtained in the driving experiment which resulted in more than 1360 snippets of 5s) was used for SRMCE-IOC and SRMCL-IOC. Nevertheless, when evaluating on real traffic data apparently different rewards were estimated. This was visible in significantly different prediction performance. In contrast, the IOC methods inferred similar reward parameters on simulated data under perfect match of the policy and POMDP model. Hence, we hypothesize that effects of model mismatch are present in the behavioral data. MCL-IOC tries to match the state control pairs present in the data by the MCE policy model. In contrast, MCE-IOC tries to directly match the state distribution of the data under both the MCE policy model and the joint task POMDP model. This can explain why MCE resulted in significantly smaller prediction error.

Similar as in many previous work e.g. [2, 260, 211], reward inference by both IOC methods combined with policy recomputation showed excellent transfer performance. This is in strong contrast to directly estimated policy whose prediction error significantly increases when evaluated on unseen driving situations. As reported in the analysis of driver behavior in Sec. 4.6.3, drivers significantly adapted the duration of glances off the road in the driving experiment. This adaptation is taken into account

in reward inference and is closely reproduced by the computed rational policies. DPE infers a static policy instead.

In the evaluation the generic baseline DPE1 resulted in consistently higher prediction error than DPE2 consisting of the barrier and the two-point steering model. This is very plausible as the model parts of DPE2 were explicitly developed for modeling driver behavior. The better prediction performance obtain by application of IOC can be explained by two aspects. First, in the maximum causal entropy policy of the joint task POMDP glance behavior and vehicle control performance as well as the external influences speed and lane curvature are all coupled. Hence, this approach results in a more holistic approach than the baseline where gaze and steering policy are largely independent of each other. This hypothesis is supported by the observed consistently lower prediction error of SRMCE-IOC compared to DPE shown in Tab. 4.6. Second, taking into account behavioral adaptation also improves overall prediction performance. This is because the specific differences in behavior can be predicted which are washed out in DPE because a single policy model is fitted over several different driving scenarios.

4.8 Conclusion

In the present chapter, we addressed inferring the parameters of the reward and the model of the driver's policy. Those quantities are required to specify appropriate glance behavior by means of rational policies in the joint task POMDP. For this purpose we introduced inverse optimal control in form of Syed's game-theoretic IOC and maximum causal entropy IOC. These frameworks were used to derive IOC algorithms for the joint task POMDP under either restriction of the sensor model or the secondary task model as well as optimal and maximum causal entropy policy models. In a first evaluation the original maximum causal entropy approach and the maximum causal likelihood variant were compared. This work was the first to discover significant differences in inferred rewards and predictive performance on small data sets. Furthermore, IOC approaches obtained consistently better prediction performance than directly estimated policies. Thereafter, we introduced a driving experiment conducted to obtain data of realistic and adaptive driver behavior for evaluating the IOC methods. Using this data, an evaluation of two DPE baselines and MCE-IOC as well as MCL-IOC was conducted. Both, overall prediction performance as well as transfer performance were considered. Similar as in the evaluation on simulated data, both IOC methods significantly outperformed the DPE approaches in terms of prediction quality. Notably, the IOC approaches also generalized to previously unseen driving scenarios.

As the evaluations demonstrated, IOC under the maximum causal entropy policy model allows to infer realistic policies of driver behavior. These are beneficial for defining appropriate glance behavior, that takes into account the driver's likely actions for improved effectiveness. Furthermore, we argue that reward inference using IOC on data of *experienced and well-instructed* drivers is also suitable to define normative parameters of the model of glance behavior. This is because in the evaluation the inferred reward parameters allowed to closely reproduce driver adaption to the different speed conditions under a *rational* policy. Hence, the resulting policies in the joint task POMDP take into account the same aspects of the driving situation that shape driver behavior. This is likely also beneficial for good acceptance of the normative model of glance behavior by the drivers. However, we admit that the acceptance of this model can only thoroughly be evaluated by means of subjective judgments of drivers. For this purpose, it is necessary to integrate the model of appropriate glance behavior into a distraction warning system and to conduct a user test. Cpt. 6 introduces a driving experiment which served to evaluate acceptance and effectiveness of the such a warning system.

In Sec. 4.4.2 we made a first a excursion on the difficulties of quantifying the driver's perception. Specifically, we introduced the issues that arise from the fact that the drivers' sensory measurements $\mathbf{z}_t$ cannot be obtained when collecting behavioral data. Fortunately, the inverse optimal control approaches allowed to overcome this problem by means of integrating over the distribution of the expected primary task states $\boldsymbol{\mu}_t^{\mathrm{p}}$ of the driver. Inverse optimal control allows to only infer policy and rewards from behavioral data and we set the sensor model assuming that the driver receives no sensory measurements of the primary task states when averting gaze. This seemed to be a reasonable approach considering the amounts gaze aversions for the required secondary task (see Fig. 4.8). How-

ever, this choice of sensor model parameters may not be appropriate for other situations e.g. other display positions. Ideally, techniques for inference of the parameters of the sensor model are applied. Cpt. 5 shows that this is possible by extending IOC and applying a similar technique of integrating over the distribution of expected primary task states.

5 Inferring Driver's Sensor Characteristics

Human motor behavior is naturally guided by sensing the environment. Therefore, an explicit model of the driver's sensor characteristics has been included into the normative model of glance behavior previously introduced in Cpt. 3. According to its strong influence on the overall model, valid parameters of this model part are important. This chapter presents methodology for inference of sensor models underlying the behavior of experienced drivers. Sec. 5.3 shows how inverse optimal control can generally be extended to inference of the dynamics in an MDP and specifically to estimate the sensor model in a belief-MDP in Sec. 5.4. Sec. 5.5 derives concrete inference procedures for the parameters of the linear Gaussian sensor model of the joint task POMDP. A new driving experiment for evaluating the estimation methods is presented in Sec. 5.6. Finally, Sec. 5.7 investigates the predictive performance when extending inverse optimal control with respect to inference of sensor models.

The inference framework developed in this chapter was strongly motivated by Felix Schmitt's collaboration with Michael Herman, Tobias Gindele, Jörg Wagner and Wolfram Burgard [77]. The contents of the present chapter were largely previously published in [204].

5.1 Introduction

In driving, most of the relevant information is acquired by the driver using vision [214] but also other sensory modalities contribute [164]. Due to the strong decrease of human visual acuity [249] often gaze switching is required when the driver wants to engage in a visually demanding secondary task. In the previous chapter, we demonstrated that the drivers' gaze switching strategies in a specific secondary task can already quite well be predicted using inverse optimal control. This was possible by inferring reward parameters of the concurring tasks of vehicle control and engagement in a secondary task as well as the cost of switching gaze. Here, we assumed that the driver does not obtain any information from the forward road scenery when averting his or her gaze. However, this simple sensor model might be less appropriate for other secondary tasks. In a series of driving experiments [230, 229, 123] researchers investigated the decrements in vehicle control performance when the drivers gazed at various locations in the vehicle's cockpit. For example, lane keeping while the driver's gaze was on the central information display, the mirrors or the speedometer were studied. The results showed significant effects of the amount of gaze aversion (in angular difference to the forward road scenery) on the driving performance. These differences were attributed to varying amount of peripheral vision which contributes to vehicle control by the driver [17]. Consequently, the normative model of glance behavior should account for the specific characteristics of peripheral vision during engagement a secondary task.

The sensor model in the joint task POMDP can implement these differences using different parameter values. However, the linear Gaussian sensor model employed in our work is only a crude approximation of the neuro-biological processes underlying human vision. Hence, we cannot derive suitable parameters from the physiology of the eye. Instead, these parameters must be estimated from experimental data of human behavior.

In estimation a significant challenge must be faced: Neither the sensory measurements made by the driver nor the resulting beliefs are measurable. Several techniques and instruments are available to measure human brain activity such as e.g. electroencephalography, functional magnetic resonance imaging or magnetoencephalography. Still, human neurological information processing is not yet sufficiently well understood to mathematically relate the measured quantities to the sensing of physical states such as the vehicles lane position. Consequently, it is not possible to infer sensor models by means of regression techniques. This is because only the inputs to the sensor model, i.e. the true states, but not the outputs, i.e. sensory measurements made by the observed human subject, are available. Note, that this is in contrast to *technical* sensors whose characteristics can often be estimated by means of regression using a reference senor with significantly higher precision.

Despite this difficulty a variety of practical models of human sensor characteristics has been estimated from behavioral data. Since the seminal works of Weber and Fechner in psycho-physics in the 19th century [60], the following approach has been pursued: Human sensor characteristics can indirectly be estimated from the human reaction they elicit. This can for example be implemented by decreasing the intensity of a stimulus until the human exposed to it cannot notice it anymore. Here, the implicit assumption is that the experimental subjects acts as an optimal detector with respect to his or her capabilities. That is, he or she reports accordingly if any difference is noticeable.

In this chapter, we generalize this established methodology and derive a formal approach for inference of sensor models from their effects on rational policies in partially observable Markov decision processes. This is done by extending inverse optimal control, which only allows for inference of reward parameters, towards estimating sensor model parameters. Importantly, the derived approach does not require any data of the sensory measurements of the observed agent.

5.2 Related Work

The main tool in psycho-physics for estimating sensor models is signal detection theory [60, 233]. Here, so-called psychometric functions are applied that relate stimulus characteristics and human sensor characteristics to the probability of detection by a subject. Fitting such a model to data of stimuli and measured responses allows to infer sensor models [233]. For example, the parameters of models of human sensing of translational acceleration can be estimated [219] and an unpublished application to driver's sensor characteristic has been reported [164]. Fit of signal detection models generally requires carefully controlled experiments where only a single stimulus is presented at a time and in several repetitions. Such an approach is only possible in a laboratory. More important, detection thresholds found in passive conditions can fail to generalize to conditions where the human is actively engaging in a task [243]. Hence, this methodology is problematic if one seeks to characterize drivers' vision during engagement in a secondary task.

Solving the joint task POMDP relied heavily on its transformation into an equivalent belief MDP. This type of MDPs unifies the decision theoretic model of MDPs with Bayesian inference on the states based on both internal models and sensory measurements. A special class of belief MPDs has been popularized under the term Bayesian Decision Theory (BDT) for modeling human perception in cognitive science and neuro science, e.g. [111]. Here, the agent possesses prior knowledge given by internal models of the states and its own sensors which is fused with the sensory measurements. Finally, the agent makes a decision as to minimize a loss function of the belief of the states. Given behavioral data and a Bayesian decision model, model parameters can be inferred by model-inversion. For example, [114] inferred the prior belief of the observed agents. This approach has been formalized in inverse Bayesian decision theory (IBDT) [46, 45]. However, it was noted that IBDT is well defined only under special conditions [3]. Important for the present work, BDT often does not consider temporal dynamics in the decision making problem. Especially, to the best-of-our-knowledge IBDT has only been applied to problems of a single decision step. In contrast, appropriate glance behavior involves several decision steps. For example, the decision to return gaze to the road considers how fast the likely accumulated deviation from the lane center can be corrected under economic steering effort.

Most relevant for this thesis, a few works have considered estimating sensor models within the POMDP class of linear quadratic Gaussian problems. This is because in this class both belief and optimal policies are given in analytical form. Notably, [176] addressed inference of linear quadratic Gaussian sensor models already in the 70s. Here, necessary conditions for identifiability and an estimation procedure were given for the special case of steady-state optimal policies in infinite horizon LQGs. This was possible by exploiting the characterization of the optimal policy derived in [101]. More recently, [70] addressed inference of internal models of agents in general LQGs. The developed approach requires a given model of the policy or the underlying rewards but allows for mismatch between the dynamics and sensor characteristics and the internal models of the agent.

Similar as [176] we consider joint inference of policy and sensor model. However, we relax the assumption of optimal behavior. Instead, our methodology assumes rational behavior according to the maximum causal entropy framework [257]. Motivated by the approach of [77] for inference of dynamics underlying MCE policies in MDPs a general framework estimating sensor models underlying

MCE policies in arbitrary POMDPs is developed. Furthermore, we derive algorithmic approaches to implement the framework in the class of the joint task POMDP. In contrast to [176, 70] who considered LQGs with a fixed sensor model, our work also allows switching of sensor models as present in the model of appropriate glance behavior. Furthermore, we address inference of parts of a POMDP model in continuous states whereas in [77] only procedures for inference of MDP dynamics in small discrete state spaces are given. Finally, the chapter reports on an evaluation of the developed approach on data of a new driving experiment. Here, we study the benefits of inference of sensor models using our approach and variants of [176, 70] in comparison to IOC as well as direct policy estimation. In addition to that, it is also investigated whether using the information available in the driver's glace behavior improves estimation of sensor models.

5.3 Inferring Dynamics From Observed Rational Behavior

In the present section we first review Simultaneous Estimation of Rewards and Dynamics (SERD) [77] for inference of dynamics in Markov decision processes from the behavior of a rational agent acting therein. The ideas used in this context will later be applied to derive a framework for estimating sensor models.

In inverse optimal control we aimed at estimating the reward parameters $\boldsymbol{\theta}$ given behavioral data $\mathsf{D} = \{(\mathbf{u}_{t=0:T}, \mathbf{x}_{t=0:T})^{i=1:n}\}$ produced by an unknown policy $\boldsymbol{\pi}_{0:T}$ under known initial state p_0 and *known* process model $\boldsymbol{\mathcal{P}}_{0:T}$. The parameters $\boldsymbol{\theta}$ could be obtained minimizing the gap

$$\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta} \Big(V_0^{\star,\boldsymbol{\theta}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big),$$

the soft gap

$$\min_{\boldsymbol{\theta}} \tilde{g}(\boldsymbol{\theta}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta} \Big(\tilde{V}_0^{\boldsymbol{\theta}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big),$$

or the negative log-likelihood

$$l(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \big(\tilde{V}_t^{\boldsymbol{\theta}}(x_t^i) - \tilde{Q}_t^{\boldsymbol{\theta}}(x_t^i, u_t^i) \big).$$

Importantly, all the optimization problems could be solved by computing the gradient of the (soft) state-control function $\nabla_{\boldsymbol{\theta}} Q^{\star,\boldsymbol{\theta}}$, $\nabla_{\boldsymbol{\theta}} \tilde{Q}^{\boldsymbol{\theta}}$ using recursions

$$\nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(x_t, u_t) = \boldsymbol{\varphi}(x_t, u_t) + \mathbb{E}\big[\nabla_{\boldsymbol{\theta}} Q_t^{\star,\boldsymbol{\theta}}(x_{t+1}, \pi_{t+1}^{\star,\boldsymbol{\theta}}(x_{t+1})) \big| \mathcal{P}(x_{t+1} | x_t, u_t) \big] \tag{5.1}$$

and

$$\nabla_{\boldsymbol{\theta}} \tilde{Q}_t^{\boldsymbol{\theta}}(x_t, u_t) = \boldsymbol{\varphi}(x_t, u_t) + \mathbb{E}\big[\nabla_{\boldsymbol{\theta}} \tilde{Q}_t^{\boldsymbol{\theta}}(x_{t+1}, u_{t+1}) \big| \tilde{\pi}_t^{\boldsymbol{\theta}}(u_{t+1} | x_{t+1}), \mathcal{P}(x_{t+1} | x_t, u_t) \big].$$

[77] observed that the negative log-likelihood is not only a differential function of the reward parameters $\boldsymbol{\theta}$ but also of parameters $\boldsymbol{\lambda}$ of the dynamics $\boldsymbol{\mathcal{P}}^{\boldsymbol{\lambda}}$. This is because both functions $\tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t)$, $\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t)$ are differentiable functions of $\boldsymbol{\lambda}$. Specifically, it holds for the gradients of the soft value function with respect to the parameters of the dynamics $\boldsymbol{\lambda}$:

$$\nabla_{\boldsymbol{\lambda}} \tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t) = \nabla_{\boldsymbol{\lambda}} \log \int \exp\big(\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) \big) \,\mathrm{d}\, u_t = \int \frac{\exp\big(\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) \big)}{\int \exp\big(\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t') \big) \,\mathrm{d}\, u_t'} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) \,\mathrm{d}\, u_t \tag{5.2}$$

$$= \mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) \big| \tilde{\pi}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_t | x_t) \big]. \tag{5.3}$$

Furthermore, the gradient of the state-control function $\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t)$ is given by

$$\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) = \nabla_{\boldsymbol{\lambda}} \Big(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) + \mathbb{E}\big[\tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}) \big| \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \big] \Big) \tag{5.4}$$

$$= \nabla_{\boldsymbol{\lambda}} \Big(\int \tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}) \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \,\mathrm{d}\, x_{t+1} \Big) \tag{5.5}$$

$$= \int \nabla_{\boldsymbol{\lambda}} \tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}) \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \,\mathrm{d}\, x_{t+1} + \int \tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}) \nabla_{\boldsymbol{\lambda}} \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \,\mathrm{d}\, x_{t+1} \tag{5.6}$$

$$\begin{aligned} &= \mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}, u_{t+1}) \big| \tilde{\pi}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_{t+1}|x_{t+1}), \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \big] \\ &\quad + \int \tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}) \nabla_{\boldsymbol{\lambda}} \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \,\mathrm{d}\, x_{t+1}. \end{aligned} \tag{5.7}$$

The gradients of the *optimal* value function $\nabla_{\boldsymbol{\lambda}} V_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t)$ and state-control function $\nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t)$ fulfill similar relations. Specifically, it holds

$$\nabla_{\boldsymbol{\lambda}} V_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t) = \nabla_{\boldsymbol{\lambda}} \big[\max_{u_t} \big(Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) \big) \big] = \mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) \big| \pi_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(u_t|x_t) \big] \tag{5.8}$$

and

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t, u_t) &= \mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} Q_{t+1}^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}, u_{t+1}) \big| \pi_{t+1}^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(u_{t+1}|x_{t+1}), \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \big] \\ &\quad + \int V_{t+1}^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t+1}) \nabla_{\boldsymbol{\lambda}} \mathcal{P}^{\boldsymbol{\lambda}}(x_{t+1}|x_t, u_t) \,\mathrm{d}\, x_{t+1}. \end{aligned} \tag{5.9}$$

[77] used the previously derived conjecture on the gradients of the soft state-control function to infer parameters $\boldsymbol{\lambda}$ of the MDP dynamics. This was done minimizing the negative log-likelihood of the MCE policy with respect to the reward parameters $\boldsymbol{\theta}$ and the parameters of the dynamics model $\boldsymbol{\lambda}$,

$$\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} l(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \big(\tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t^i) - \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t^i, u_t^i) \big). \tag{5.10}$$

In that work it was shown, that additional inference of the parameters $\boldsymbol{\lambda}$ can significantly improve prediction performance.

As previously discussed in Cpt. 4, gradients of the (soft) state-control functions wrt. to the reward parameters can be used for solving the minimization problems underlying Syed's approach to IOC as well as MCE-IOC and MCL-IOC. Similar, the recursive definition of the gradients of the (soft) state-control function with respect to the parameters of the dynamics can be used to extend all these approaches with respect to estimation of parts of the dynamics. This is possible by minimizing the gap

$$\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda}} \Big(V_0^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big), \tag{5.11}$$

which we will denote OPT-SERD, minimizing the soft gap

$$\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \tilde{g}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda}} \Big(\tilde{V}_0^{\boldsymbol{\theta},\boldsymbol{\lambda}} - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big), \tag{5.12}$$

termed MCE-SERD or minimizing the negative log-likelihood

$$l(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T} \big(\tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t^i) - \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(x_t^i, u_t^i) \big) \tag{5.13}$$

referred to as MCL-SERD with respect to $\boldsymbol{\lambda}$. Considering the analysis of Cpt. 4, we can expect MCE-SERD and MCL-SERD to perform similar under perfect match of model assumptions and infinite amount of data. However, we leave a formal proof open for future research.

5.4 Inferring Sensor Models From Rational Behavior in Belief-MDPs

Previously, we introduced the technique of SERD for inference of model parts of MDPs by extending inverse optimal control. Here, the main idea was to take derivatives of the (soft) state-control functions. It is now shown that a very similar approach can be used for inference of sensor models in POMDPs.

For this purpose, we first return to the reformulation of POMDPs into the equivalent belief-MDPs. In Sec. 2.1.3 the dynamics of the belief MDP $\boldsymbol{\mathcal{P}}^b$ were introduced as

$$\mathcal{P}_t^b(b(x_{t+1})|b(x_t),u_t) = \int_{\mathsf{Z}(b(x_{t+1})|b(x_t),u_t)} \Big(\int p(z_{t+1}|x_{t+1}) \mathcal{P}_{t+1}(x_{t+1}|u_t,x_t) b(x_t) \,\mathrm{d}\, x_{t+1}, x_t \Big) \,\mathrm{d}\, z_{t+1}, x_t,$$

$$\mathsf{Z}(b(x_t)|b(x_{t-1}),u_{t-1}) = \Big\{ \hat{z}_t \in \mathsf{Z} \colon b(x_t) = \frac{\int p(\hat{z}_t|x_t) \mathcal{P}_t(x_t|u_{t-1},x_{t-1}) b(x_{t-1}) \,\mathrm{d}\, x_{t-1}}{\int p(z_t'|x_t) \mathcal{P}_t(x_t|u_{t-1},x_{t-1}) b(x_{t-1}) \,\mathrm{d}\, z_t', x_{t-1}} \Big\},$$

while the reward function of the belief-MDP was

$$r^b(b(x_t),u_t) = \mathbb{E}[r(x_t,u_t)|b(x_t),u_t].$$

Consequently, in the case of a belief-MDP the soft Bellman equations result in the specific form of

$$\tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t)) = \log \int \exp\big(\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t),u_t)\big) \,\mathrm{d}\, u_t \tag{5.14}$$

$$\begin{aligned} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t),u_t) &= r^b(b(x_t),u_t) + \mathbb{E}\big[\tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_{t+1}))\big|\mathcal{P}_t^{b,\boldsymbol{\lambda}}(b(x_{t+1})|b(x_t),u_t)\big] \qquad (5.15)\\ &= r^b(b(x_t),u_t) \\ &\quad + \int \Big[\tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}\Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t) b(x_t) \,\mathrm{d}\, x_t}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t) b(x_t) \,\mathrm{d}\, z_{t+1}', x_t}\Big) \\ &\qquad \cdot p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t) b(x_t)\Big] \,\mathrm{d}\, z_{t+1}, x_{t+1}, x_t. \qquad (5.16) \end{aligned}$$

Given the formulation of the (soft) Bellman equations, analogically to SERD in MPDs the derivatives with respect to the sensor model parameters $\boldsymbol{\lambda}$ can be taken. In case of the soft value function we obtain

$$\nabla_{\boldsymbol{\lambda}} \tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t)) = \mathbb{E}[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t),u_t)|\tilde{\pi}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_t|b(x_t))]. \tag{5.17}$$

The derivative of the soft state-control function $\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t),u_t)$ is

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t),u_t) &= \nabla_{\boldsymbol{\lambda}} \mathbb{E}\big[\tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_{t+1}))\big|\mathcal{P}_t^{b,\boldsymbol{\lambda}}(b(x_{t+1})|b(x_t),u_t)\big] \qquad (5.18)\\ &= \nabla_{\lambda} \int \Big[\tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}\Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t') b(x_t') \,\mathrm{d}\, x_t}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t''}\Big) \\ &\qquad \cdot p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t) b(x_t)\Big] \,\mathrm{d}\, z_{t+1}, x_{t+1}, x_t \qquad (5.19)\\ &= \int \Big(\nabla_{\boldsymbol{\lambda}}\Big[\tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}\Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t') b(x_t') \,\mathrm{d}\, x_t'}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t''}\Big)\Big] \\ &\qquad \cdot p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t) b(x_t) \\ &\qquad + \tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}\Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t') b(x_t') \,\mathrm{d}\, x_t'}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t''}\Big) \\ &\qquad \cdot \nabla_{\boldsymbol{\lambda}} p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t,x_t) b(x_t)\Big) \,\mathrm{d}\, z_{t+1}, x_{t+1}, x_t \qquad (5.20) \end{aligned}$$

$$
\begin{aligned}
&= \int \Big(\nabla_{\boldsymbol{\lambda}} \Big[\tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}} \Big] \Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t') b(x_t') \,\mathrm{d}\, x_t'}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t''} \Big) \\
&\quad \cdot \nabla_{\boldsymbol{\lambda}} \Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t') b(x_t') \,\mathrm{d}\, x_t'}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t''} \Big) \\
&\quad \cdot p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t) b(x_t) \\
&\quad + \tilde{V}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}} \Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t') b(x_t') \,\mathrm{d}\, x_t'}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t} \Big) \\
&\quad \cdot \nabla_{\boldsymbol{\lambda}} p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t) b(x_t) \Big) \,\mathrm{d}\, z_{t+1}, x_{t+1}, x_t
\end{aligned} \tag{5.21}
$$

As can be seen the recursion with respect to the sensor model parameter in belief MDPs is quite similar to the recursion of MCE-SERD in MPD (5.3), (5.7). The only difference is that the derivative is more complicated in belief-MDPs. This is because of the nonlinear association of sensory measurements with resulting beliefs which has to be taken into account by means of the factor

$$
\nabla_{\boldsymbol{\lambda}} \Big(\frac{\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t') b(x_t') \,\mathrm{d}\, x_t'}{\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t''} \Big) \tag{5.22}
$$

$$
\begin{aligned}
&= \frac{\big(\int \nabla_{\boldsymbol{\lambda}} p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t') b(x_t') \,\mathrm{d}\, x_t' \big) \big(\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_{t+1}', x_t'' \big)}{\big(\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t'' \big)^2} \\
&\quad - \frac{\big(\int p^{\boldsymbol{\lambda}}(z_{t+1}|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t') b(x_t') \,\mathrm{d}\, x_t' \big) \big(\int \nabla_{\boldsymbol{\lambda}} p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_{t+1}', x_t'' \big)}{\big(\int p^{\boldsymbol{\lambda}}(z_{t+1}'|x_{t+1}) \mathcal{P}_t(x_{t+1}|u_t, x_t'') b(x_t'') \,\mathrm{d}\, z_t', x_t'' \big)^2}.
\end{aligned} \tag{5.23}
$$

Computation of the recursion (5.21) might look complicated but requires only to integrate over the finite dimensional quantities $z_{t+1}, z_{t+1}', x_{t+1}, x_t, x_t', x_t''$. Furthermore, $\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t), u_t)$ requires to integrate over *potential* sensory measurements z_{t+1}, z_{t+1}' and does not rely on the unknown sensory measurements z_t^i made by the observed agent. Similar to the principle previously applied in IOC (Sec. 4.4.2), this allows us to infer parameters $\boldsymbol{\lambda}$ of sensor models $p^{\boldsymbol{\lambda}}(z_t|x_t)$ by means of solving

$$
\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \tilde{g}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda}} \Big(\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_0) = p_0) - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big), \tag{5.24}
$$

using the gradient

$$
\mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} \tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_0) = p_0, u_0) | \pi^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_0|p_0) \big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big]. \tag{5.25}
$$

Analogously, we can also use the optimal value function $V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}$ for inference of sensor models

$$
\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathsf{D}) = \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\lambda}} \Big(V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_0) = p_0) - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\theta}^\top \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big] \Big), \tag{5.26}
$$

using its gradient

$$
\mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_0) = p_0, u_0) | \pi^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(u_0|p_0) \big] - \mathbb{E}\Big[\sum_{t=0}^{T} \boldsymbol{\varphi}(x_t, u_t) \Big| \mathsf{D} \Big]. \tag{5.27}
$$

We refer to the inference of sensor models using previously introduced approaches as *I See What You See* (ISWYS).

While iteratively conducting equation (5.7) is computationally tractable in small discrete MDPs, the recursion (5.21) is in general not feasible. This is because, it requires back-up of $\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(b(x_t), u_t)$ for every possible belief, which is problematic because $b(x_t)$ is a continuous valued quantity. However, we will show that again the class of the joint task POMDP allows to implement this approach.

5.5 Inferring Sensor Models in The Joint Task POMDP

To implement ISWYS in the joint task POMDP, we first consider the sensor model. As previously introduced in (3.13), in this POMDP the sensory measurements are obtained according to a linear Gaussian sensor model,

$$\mathbf{z}_t \sim \mathcal{N}(\mathbf{H}, \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z)).$$

In this work we assume that unknown parameters are present in the sensor noise covariance $\boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z)$. For example, in the context of application to quantifying drivers' sensor characteristics we are interested in the variance in the noise in sensing the vehicle's lane position $(\sigma_y)^2(x_t^z)$ and orientation in lane $(\sigma_\phi)^2(x_t^z)$, $\boldsymbol{\Sigma}^{\epsilon^z}(x_t^z) = \mathrm{diag}((\sigma_y)^2(x_t^z), 0, (\sigma_\phi)^2(x_t^z), 0)$.

5.5.1 Posing ISWYS in The Joint Task POMDP

Now we return to the (soft) Bellman equations. Here, we assume a linear parametrization of the reward function and an initial covariance $\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}$. We use superscripts to denote which objects depend on the reward parameter $\boldsymbol{\theta}$ and which depend on the parameter of the sensor model $\boldsymbol{\lambda}$. Accordingly, the Bellman equations of the optimal policy are given by

$$Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{Q^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \quad (5.28)$$

$$V_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}), \quad (5.29)$$

$$\mathbf{M}_t^{Q^{\star,\boldsymbol{\theta}}} = \begin{cases} [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} [\mathbf{A}_t\ \mathbf{B}_t] + \mathrm{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) & \text{if } t < T \\ \mathrm{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) & \text{else} \end{cases} \quad (5.30)$$

$$\mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}} = \begin{cases} 2[\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{a}_t + [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{V^{\star,\boldsymbol{\theta}}} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \quad (5.31)$$

$$m_t^{Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \begin{cases} \mathbf{a}_t^\top \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \\ \quad + \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \\ \quad + \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \left(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\ x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\right)\right) \\ \quad + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \\ \quad + \mathbb{E}\left[m_{t+1}^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\ x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right] & \text{if } t < T \\ \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T})) + \theta_3 u_T^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}}) & \text{else} \end{cases}, \quad (5.32)$$

$$\mathbf{M}_t^{V^{\star,\boldsymbol{\theta}}} = \mathbf{M}_{t\ x,x}^{Q^{\star,\boldsymbol{\theta}}} - \mathbf{M}_{t\ x,u}^{Q^{\star,\boldsymbol{\theta}}} [\mathbf{M}_{t\ u,u}^{Q^{\star,\boldsymbol{\theta}}}]^{-1} \mathbf{M}_{t\ u,x}^{Q^{\star,\boldsymbol{\theta}}} \tag{5.33}$$

$$\mathbf{m}_t^{V^{\star,\boldsymbol{\theta}}} = \mathbf{m}_{t\ x}^{Q^{\star,\boldsymbol{\theta}}} - \mathbf{M}_{t\ x,u}^{Q^{\star,\boldsymbol{\theta}}} [\mathbf{M}_{t\ u,u}^{Q^{\star,\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{Q^{\star,\boldsymbol{\theta}}} \tag{5.34}$$

$$\begin{aligned} m_t^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = &-\frac{1}{4} [\mathbf{m}_{t\ u}^{Q^{\star,\boldsymbol{\theta}}}]^\top [\mathbf{M}_{t\ u,u}^{Q^{\star,\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{Q^{\star,\boldsymbol{\theta}}} + \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \\ &+ \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} (\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x})\right) \\ &+ \max_{u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \Big(\theta_3 u_t^{\mathrm{z}} - \mathrm{tr}\left(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\right) \\ &+ \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) + \mathbb{E}\left[m_{t+1}^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \right] \Big). \end{aligned} \tag{5.35}$$

Furthermore, the soft Bellman equations result in

$$\tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \tag{5.36}$$

$$\tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}), \tag{5.37}$$

$$\mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} = \begin{cases} [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{A}_t\ \mathbf{B}_t] + \mathrm{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) & \text{if } t < T \\ \mathrm{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) & \text{else} \end{cases} \tag{5.38}$$

$$\mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} = \begin{cases} 2[\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \mathbf{a}_t + [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} & \text{if } t < T \\ \mathbf{0} & \text{else} \end{cases} \tag{5.39}$$

$$m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \begin{cases} \begin{aligned} &\mathbf{a}_t^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \\ &\quad + \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \\ &\quad + \mathrm{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} (\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}]))\right) \\ &\quad + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \\ &\quad + \mathbb{E}\left[m_{t+1}^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \right] \end{aligned} & \text{if } t < T \\ \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T})) + \theta_3 u_T^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}}) & \text{else} \end{cases}, \tag{5.40}$$

$$\mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} = \mathbf{M}_{t\ x,x}^{\tilde{Q}^{\boldsymbol{\theta}}} - \mathbf{M}_{t\ x,u}^{\tilde{Q}^{\boldsymbol{\theta}}} [\mathbf{M}_{t\ u,u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^{-1} \mathbf{M}_{t\ u,x}^{\tilde{Q}^{\boldsymbol{\theta}}} \tag{5.41}$$

$$\mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} = \mathbf{m}_{t\ x}^{\tilde{Q}^{\boldsymbol{\theta}}} - \mathbf{M}_{t\ x,u}^{\tilde{Q}^{\boldsymbol{\theta}}} [\mathbf{M}_{t\ u,u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{\tilde{Q}^{\boldsymbol{\theta}}} \tag{5.42}$$

$$\begin{aligned} m_t^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) = &-\frac{1}{4} [\mathbf{m}_{t\ u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^\top [\mathbf{M}_{t\ u,u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^{-1} \mathbf{m}_{t\ u}^{\tilde{Q}^{\boldsymbol{\theta}}} + \frac{1}{2} \log(\det(\pi [\mathbf{M}_{t\ u,u}^{\tilde{Q}^{\boldsymbol{\theta}}}]^{-1})) \\ &+ \mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathrm{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} (\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}) \mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x})\right) \\ &+ \mathrm{softmax}_{u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \Big(\theta_3 u_t^{\mathrm{z}} - \mathrm{tr}\left(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\right) \\ &+ \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) + \mathbb{E}\left[m_{t+1}^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \middle| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \right] \Big). \end{aligned} \tag{5.43}$$

In both the optimal and the maximum causal entropy policy, here many parts of the (soft) value and (soft) state-control function do not depend on the parameter of the sensor model $\boldsymbol{\lambda}$. Specifically, only the terms $m_t^{Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}), m_t^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}), m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}), m_t^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}})$ are related to the sensor model. In both cases also the term $\mathrm{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}))$ must be considered. This is in contrast to the general formulation of ISWYS where the immediate reward $r^b(b(x_t), u_t)$ was not relevant for inference of the sensor model (see the gradient (5.18)). Previously, we substituted the *reachable* covariances $\boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})$ by the associated sensor state sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ in the joint task POMDP.

Hence, we need to consider the effects of the sensor model on the covariance associated with a sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$.

Analogously, to the derivations in inverse optimal control in Sec. 4.4.1, we can pose inference of sensor model parameters $\boldsymbol{\lambda}$ and reward parameters $\boldsymbol{\theta}$ as joint minimization of the (soft) gap. That is, we solve

$$\begin{aligned}
&\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} g(\boldsymbol{\theta},\boldsymbol{\lambda},\mathsf{D}) \\
&= \min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \Big(\mathbb{E}\Big[[\boldsymbol{\mu}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + \mathbf{m}_0^{V^{\star,\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] \Big| \mathcal{N}(\boldsymbol{\mu}_0^{\mathrm{p}} | \mathbf{x}_0^{\mathrm{p}}, \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}) \Big] + m_0^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda},1}}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big) \\
&= \min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \Big([\mathbf{x}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + \mathrm{tr}(\mathbf{M}_0^{V^{\star,\boldsymbol{\theta}}} \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}) + \mathbf{m}_0^{V^{\star,\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + m_0^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda},1}}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big)
\end{aligned} \tag{5.44}$$

for estimation under the optimal policy model. Alternatively, inference under the maximum causal entropy policy model is possible considering the minimization problem

$$\begin{aligned}
&\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \tilde{g}(\boldsymbol{\theta},\boldsymbol{\lambda},\mathsf{D}) \\
&= \min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \Big(\mathbb{E}\Big[[\boldsymbol{\mu}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] + \mathbf{m}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_0^{\mathrm{p}}] \Big| \mathcal{N}(\boldsymbol{\mu}_0^{\mathrm{p}} | \mathbf{x}_0^{\mathrm{p}}, \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}) \Big] + m_0^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda},1}}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big) \\
&= \min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \Big([\mathbf{x}_0^{\mathrm{p}}]^\top \mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + \mathrm{tr}(\mathbf{M}_0^{\tilde{V}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}) + \mathbf{m}_0^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_0^{\mathrm{p}}] + m_0^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda},1}}(x_0^{\mathrm{z}}, x_0^{\mathrm{i}}) \\
&\qquad - \mathbb{E}\Big[\sum_{t=0}^{T} \mathrm{vec}(\boldsymbol{\Theta}_1)^\top \mathrm{vec}(\mathbf{x}_t^{\mathrm{p}} \mathbf{x}_t^{\mathrm{p}\,\top}) + \mathrm{vec}(\boldsymbol{\Theta}_2)^\top \mathrm{vec}(u_t^{\mathrm{p}} u_t^{\mathrm{p}\,\top}) + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top \boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \Big| \mathsf{D} \Big] \Big).
\end{aligned} \tag{5.45}$$

Finally, in case of the joint task POMDP we can also derive an inference technique for the parameters $\boldsymbol{\lambda}$ by minimizing the neg. log-likelihood of the maximum entropy policy. This generalizes the maximum causal likelihood variant of inverse optimal control in the joint task POMDP. Here, we first take expectations with respect to the true states $\mathbf{x}_t^{\mathrm{p}\,j}$ as well as the covariances $\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{j})$ according to the sensor state sequences $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{j}$ for the states $\mathbf{x}_t^{\mathrm{p}\,j}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{j}$ present in the data D. Consequently, we arrive at the minimization problem of

$$\begin{aligned}
&\min_{\boldsymbol{\theta},\boldsymbol{\lambda}} l(\boldsymbol{\theta},\boldsymbol{\lambda},\mathsf{D}) \\
&= \min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \mathbb{E}\Big[\sum_{t=0}^{T} \mathbb{E}\big[[\boldsymbol{\mu}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda},1}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) \\
&\qquad - \big([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda},1}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \big) | \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}} | \mathbf{x}_t^{\mathrm{p}}, \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \big] \Big| \mathsf{D} \Big] \\
&= \min_{\boldsymbol{\theta},\boldsymbol{\lambda}} \mathbb{E}\Big[\sum_{t=0}^{T} [\mathbf{x}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}] + \mathrm{tr}(\mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}] + m_t^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}) \\
&\qquad - \big([\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathrm{tr}(\mathbf{M}_{t\,x,x}^{\tilde{Q}^{\boldsymbol{\theta}}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \big) \Big| \mathsf{D} \Big].
\end{aligned} \tag{5.46}$$

5.5.2 Obtaining the Sensor Model Gradients

Following the analysis of the (soft) Bellman equations, computing the derivatives of the (soft) value functions and (soft) state-control functions

$$\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \nabla_{\boldsymbol{\lambda}} m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}), \tag{5.47}$$

$$\nabla_{\boldsymbol{\lambda}} \tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = \nabla_{\boldsymbol{\lambda}} m_t^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}), \tag{5.48}$$

$$\nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \nabla_{\boldsymbol{\lambda}} m_t^{Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}), \tag{5.49}$$

$$\nabla_{\boldsymbol{\lambda}} V_t^{\star,\boldsymbol{\theta}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = \nabla_{\boldsymbol{\lambda}} m_t^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}), \tag{5.50}$$

can be addressed. Here, we drop the dependence on the variables $\boldsymbol{\mu}_t^{\mathrm{p}}, u_t^{\mathrm{p}}$ that do not affect the gradients. In the case of the soft Bellman equations it holds:

$$\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \nabla_{\boldsymbol{\lambda}} m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \tag{5.51}$$

$$= \begin{cases} \nabla_{\boldsymbol{\lambda}} \Big(\operatorname{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})) \\ \quad + \operatorname{tr}\Big(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}\big(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})\mathbf{A}_t^\top - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\big)\Big) \\ \quad + \mathbb{E}\Big[m_{t+1}^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \Big| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\Big]\Big) & \text{if } t < T \\ \nabla_{\boldsymbol{\lambda}} \operatorname{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:T})) & \text{else} \end{cases}, \tag{5.52}$$

$$= \begin{cases} \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}) \operatorname{vec}(\boldsymbol{\Theta}_1) \\ \quad + \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}) \operatorname{vec}(\mathbf{A}_t^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \mathbf{A}_t) \\ \quad - \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}]) \operatorname{vec}(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}) \\ \quad + \mathbb{E}\Big[\nabla_{\boldsymbol{\lambda}} m_{t+1}^{\tilde{V}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \Big| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\Big] & \text{if } t < T \\ \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:T}) \operatorname{vec}(\boldsymbol{\Theta}_1) & \text{else} \end{cases}, \tag{5.53}$$

$$\nabla_{\boldsymbol{\lambda}} \tilde{V}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = \mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} m_t^{\tilde{Q}^{\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \big| \tilde{\pi}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_t^{\mathrm{z}}, u_t^{\mathrm{i}}, \boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}})\big]. \tag{5.54}$$

Furthermore, the gradients of the optimal value function and optimal state-control function are given by

$$\nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \nabla_{\boldsymbol{\lambda}} m_t^{Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \tag{5.55}$$

$$= \begin{cases} \nabla_{\boldsymbol{\lambda}} \Big(\operatorname{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})) \\ \quad + \operatorname{tr}\Big(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}}\big(\mathbf{A}_t \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t})\mathbf{A}_t^\top - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}])\big)\Big) \\ \quad + \mathbb{E}\Big[m_{t+1}^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \Big| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\Big]\Big) & \text{if } t < T \\ \nabla_{\boldsymbol{\lambda}} \operatorname{tr}(\boldsymbol{\Theta}_1 \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:T})) & \text{else} \end{cases}, \tag{5.56}$$

$$= \begin{cases} \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}) \operatorname{vec}(\boldsymbol{\Theta}_1) \\ \quad + \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:t}) \operatorname{vec}(\mathbf{A}_t^\top \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{A}_t) \\ \quad - \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}]) \operatorname{vec}(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}}) \\ \quad + \mathbb{E}\Big[\nabla_{\boldsymbol{\lambda}} m_{t+1}^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}([\boldsymbol{x}^{\mathbf{z}}_{0:t}\, x_t^{\mathrm{z}} \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \Big| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\Big] & \text{if } t < T \\ \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}_{0:T}) \operatorname{vec}(\boldsymbol{\Theta}_1) & \text{else} \end{cases}, \tag{5.57}$$

$$\nabla_{\boldsymbol{\lambda}} V_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}) = \mathbb{E}\big[\nabla_{\boldsymbol{\lambda}} m_t^{Q^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \big| \pi_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(u_t^{\mathrm{z}}, u_t^{\mathrm{i}}, \boldsymbol{x}^{\mathbf{z}}_{0:t}, x_t^{\mathrm{i}})\big]. \tag{5.58}$$

Analyzing the sensor model gradients of the soft Bellman equations and Bellman equations, it turns out that most of the quantities involved are known. These are the matrices $\mathbf{A}_t, \mathbf{\Theta}_1, \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}, \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}}$. In addition to these given quantities, it is required to obtain the gradients of the belief covariance $\nabla_{\boldsymbol{\lambda}} \mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})$ as resulting from the Kalman filter under the sensor states x_t^{z} in sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$. These gradients can recursively be computed by taking the derivatives of the Kalman belief update. For this purpose, we consider the update of the covariance given by

$$
\begin{aligned}
\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} &= \mathbf{A}_t \mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}} \mathbf{A}_t^\top + \mathbf{\Sigma}^{\epsilon^x} \\
\mathbf{K}_{t+1} &= \bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} \mathbf{H}^\top (\mathbf{H} \bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} \mathbf{H}^\top + \mathbf{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}})^+ \\
\mathbf{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} &= (\mathbf{I}^{n_x^{\mathrm{p}}} - \mathbf{K}_{t+1}\mathbf{H}) \bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}},
\end{aligned}
$$

where n_x^{p} denotes the dimension of the primary task states $\mathbf{x}^{\mathrm{p}}$. In the following, we use the notation $\partial_{\mathbf{x}}(\mathbf{f}(\mathbf{x})) = [\nabla_{\mathbf{x}}(\mathbf{f}(\mathbf{x}))]^\top$ and n_z to denote the dimension of the sensory measurements $\mathbf{z}_t$. Applying matrix calculus, we can take the derivative wrt. to the parameters of the sensor model of the individual steps:

$$
\begin{aligned}
\partial_{\boldsymbol{\lambda}} \operatorname{vec}(\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}) &= (\mathbf{A}_t \otimes \mathbf{A}_t) \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}) && (5.59)\\
\mathbf{R}_{t+1}^{\boldsymbol{\lambda}} &:= \mathbf{H} \bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} \mathbf{H}^\top + \mathbf{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}} && (5.60)\\
\partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{R}_{t+1}^{\boldsymbol{\lambda}}) &= (\mathbf{H} \otimes \mathbf{H}) \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}) + \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}) && (5.61)\\
\forall_i \partial_{\boldsymbol{\lambda}_i} \operatorname{vec}(\mathbf{R}_{t+1}^{\boldsymbol{\lambda},+}) &= \big[-(\mathbf{R}_{t+1}^{\boldsymbol{\lambda},+} \otimes \mathbf{R}_{t+1}^{\boldsymbol{\lambda},+}) + (\mathbf{I}^{n_z} - \mathbf{R}_{t+1}^{\boldsymbol{\lambda}} \mathbf{R}_{t+1}^{\boldsymbol{\lambda},+}) \otimes (\mathbf{R}_{t+1}^{\boldsymbol{\lambda},+} \mathbf{R}_{t+1}^{\boldsymbol{\lambda},+}) \\
&\qquad + (\mathbf{R}_{t+1}^{\boldsymbol{\lambda},+} \mathbf{R}_{t+1}^{\boldsymbol{\lambda},+}) \otimes (\mathbf{I}^{n_z} - \mathbf{R}_{t+1}^{\boldsymbol{\lambda},+} \mathbf{R}_{t+1}^{\boldsymbol{\lambda}}) \big] \partial_{\boldsymbol{\lambda}_i} \mathbf{R}_{t+1}^{\boldsymbol{\lambda}} && (5.62)\\
\partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{K}_{t+1}^{\boldsymbol{\lambda}}) &= \big((\mathbf{R}_{t+1}^{\boldsymbol{\lambda},+} \mathbf{H}) \otimes \mathbf{I}^{n_z} \big) \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}) + \big(\mathbf{I}^{n_z} \otimes (\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} \mathbf{H}^\top) \big) \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{R}_{t+1}^{\boldsymbol{\lambda},+}) && (5.63)\\
\partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}) &= \big(\mathbf{I}^{n_z} \otimes (\mathbf{I}^{n_x} - \mathbf{K}_{t+1}\mathbf{H}) \big) \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}) - \big((\bar{\mathbf{\Sigma}}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}} \mathbf{H}^\top) \otimes \mathbf{I}^{n_z} \big) \partial_{\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{K}_{t+1}^{\boldsymbol{\lambda}}). && (5.64)
\end{aligned}
$$

Here, the equations result from standard rules of matrix differentiation [175] and the derivative of the pseudo-inverse that is presented in [224].

Senor Model of Initial Covariance at Inference Time Finally, we must also consider the influence of the sensor model parameters $\boldsymbol{\lambda}$ on the initial covariance of the primary state belief $\mathbf{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}$. This is done in similar fashion as in inverse optimal control Sec. 4.4.2. We first compute the gradient of the steady state covariance $\nabla_{\boldsymbol{\lambda}} \hat{\mathbf{\Sigma}}^{\mathrm{p},\boldsymbol{\lambda}}$. If the driver's gaze is off the road for $t = 0$, we compute the gradient of the steady state covariance at the last time step the gaze was on the road. Thereafter the gradients of the covariance $t_{\text{gaze aversion}} : 0$ are obtained.

Tractable Implementation Combining the gradients wrt. the sensor model parameters $\boldsymbol{\lambda}$ of the (soft) state-control function with the derivative of the Kalman filter allows to implement the approaches of (5.24). However, backing up the gradients of the (soft) state-control functions

$$
\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}), \quad \nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})
$$

is in general computationally infeasible similar as in case of sate-control function itself or the gradients wrt. the reward parameters.

Using the restricted sensor model SR, direct back-up of the (soft) state-control functions is feasible as $\boldsymbol{x}^{\mathbf{z}}{}_{0:t}$ can be replaced by d_t. We refer to the resulting approaches for sensor model inference SR-Opt-ISWYS and SR-MCE-ISWYS.

In the case of the optimal policy in the joint task model with restriction of the secondary task also a computationally feasible approach can be derived. Here, the optimal sensor control policy is a single optimal sequence $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}$. Consequently, computing $\nabla_{\boldsymbol{\lambda}} V_0^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_0^{\mathrm{p}}, x_0^{\mathrm{z}})$ requires only considering

the sensor states of the optimal sensor state sequence. As an alternatively to the intractable recursion (5.55), this conjecture can be used to compute $\nabla_{\boldsymbol{\lambda}} V_0^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\mu}_0^{\mathrm{p}}, x_0^{\mathrm{z}}) = \nabla_{\boldsymbol{\lambda}} m_0^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(x_0^{\mathrm{z}})$ in a forward pass

$$\begin{aligned}\nabla_{\boldsymbol{\lambda}} V_0^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_0^{\mathrm{z}}) = \sum_{t=0}^{T-1} \big(&\nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{\star}) \operatorname{vec}(\boldsymbol{\Theta}_1) \\ &+ \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}{}^{\star}) \operatorname{vec}(\mathbf{A}_t^{\top} \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{A}_t) \\ &- \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t+1}{}^{\star}) \operatorname{vec}(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}})\big) + \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}) \operatorname{vec}(\boldsymbol{\Theta}_1). \end{aligned} \tag{5.65}$$

We will denote this approach as STR-Opt-ISWYS.

5.5.3 Illustrative Example

Before we present prototypical algorithms for sensor model inference, we wish to explain and illustrate the individual parts of the gradient of the optimal value function in Fig. 5.1. Here, the case of the joint task POMDP under restriction of the secondary task model, STR-Opt-ISWYS, is considered.

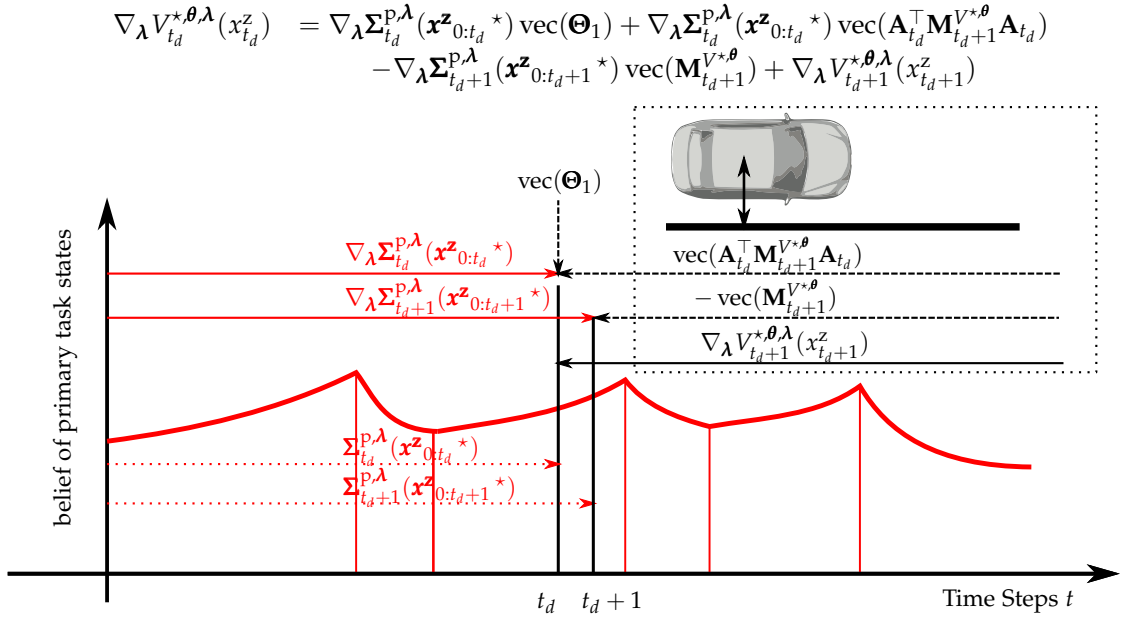

Figure 5.1: Illustrative example of the gradient computation of STR-Opt-ISWYS. Assume that the gradient $\nabla_{\boldsymbol{\lambda}} V_{t_d}^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(x_{t_d}^{\mathrm{z}}{}^{\star})$ for a time step t_d shall be computed. This requires to consider the gradient of the covariance of the belief of the states related to vehicle control of the current time step t_d, $\nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_{t_d}^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{\star})$ as well as the next time step $t_d + 1$, $\nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_{t_d}^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{\star})$. These gradients are computed along the sequence of optimal sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{\star}$. Thereafter, the effects of the covariance $\boldsymbol{\Sigma}_{t_d}^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d}{}^{\star})$ on current and future vehicle control under the optimal steering policy are considered. For example, increased uncertainty in the belief results in increased deviation from the lane center. This is done by multiplying the gradients of the covariance with the immediate reward of the primary task states $\boldsymbol{\Theta}_1$ as well as with the matrices $\mathbf{A}_{t_d}^{\top} \mathbf{M}_{t_d+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{A}_{t_d}$, $-\mathbf{M}_{t_d+1}^{V^{\star,\boldsymbol{\theta}}}$. The latter is the part of the value function related to the quadratic reward function of the primary task states. Finally, the accumulated gradients of the future time steps $t > t_d$, $\nabla_{\boldsymbol{\lambda}} m_{t_d+1}^{V^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t_d+1}{}^{\star})$ are added. Summarized the gradient of the optimal value function with respect to sensor model parameters quantifies the changes in vehicle control performance as resulting from changes in the belief of the primary task states under the current optimal sequence of sensor states $\boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^{\star}$.

5.5.4 ISWYS Algorithms

To infer sensor models, the gap $g(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathsf{D})$ or the soft gap $\tilde{g}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathsf{D})$ must be minimized with respect to the reward parameters $\boldsymbol{\theta}$ and the sensor model parameters $\boldsymbol{\lambda}$. Similar to inverse optimal control (Sec. 4.4.4) the reward parameters $\boldsymbol{\theta} := [\mathrm{vec}(\boldsymbol{\Theta}_1); \mathrm{vec}(\boldsymbol{\Theta}_2); \theta_3; \boldsymbol{\theta}_4]$ are constrained to the feasible set Θ. Furthermore, $\boldsymbol{\lambda}$ is required to result in a well-defined positive semi-definite sensor noise covariance $\boldsymbol{\Sigma}^{\epsilon^z, \boldsymbol{\lambda}}$. For example, if the entire sensor model covariance shall be inferred, i.e. $\boldsymbol{\lambda} = \mathrm{vec}(\boldsymbol{\Sigma}^{\epsilon^z, \boldsymbol{\lambda}})$, $\boldsymbol{\lambda}$ must result in a positive semi-definite matrix. For this purpose, we introduce the set of feasible sensor model parameters Λ.

While the gap and the soft gap are convex functions of the reward parameter $\boldsymbol{\theta}$, both functions are not necessarily convex in the parameter of the sensor model $\boldsymbol{\lambda}$. Consequently, projected gradient descent can not only fail to converge to a *global* optimum but may not even converge to a *local* optimal solution. To obtain a local optimal solution instead other techniques as, e.g. as presented in [112], must be applied.

In the following we will outline prototypical algorithms to infer sensor models using STR-Opt-ISWYS and using SR-Opt-ISWYS as well as using SR-MCE-ISWYS. Similar as in the case of the inverse optimal control algorithms we only give the simplest approach. Note, especially computing the quantities related to the belief covariance and its gradients can be computed more efficient using the techniques introduced in the corresponding policy computation Algo. 4 and Algo. 7.

Algorithm 16 Generic ISWYS solver [SolveISWYS]

function SolveISWYS($(\mathbf{A}, \mathbf{a}, \mathbf{B})_{0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z}(x_t^z), \mathcal{P}_t^i, x_0^z, \mathsf{D}$)
Require: feasible set of reward parameters Θ, feasible set of sensor parameters Λ
  $\boldsymbol{\theta} \leftarrow$ Sample(Θ)
  $\boldsymbol{\lambda} \leftarrow$ Sample(Λ)
  while not converged $\boldsymbol{\theta}, \boldsymbol{\lambda}$ **do**
    $[v, \nabla_{\boldsymbol{\theta},\boldsymbol{\lambda}}] \leftarrow$ Eval<name>$(\boldsymbol{\theta}, \boldsymbol{\lambda}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \mathcal{P}^i, \mathsf{D})$ ▷ evaluate the different ISWYS objectives STR-Opt-ISWYS, SR-Opt-ISWYS, SR-MCE-ISWYS, SR-MCL-ISWYS
    $(\boldsymbol{\theta}, \boldsymbol{\lambda}) \leftarrow$ PerformStep$(\boldsymbol{\theta}, \boldsymbol{\lambda}, v, \nabla_{\boldsymbol{\theta},\boldsymbol{\lambda}}, \Theta, \Lambda)$ ▷ e.g. using the techniques of [112]
  end while
  return $(\boldsymbol{\theta}, \boldsymbol{\lambda})$
end function

Algorithm 17 Evaluation of SR-Opt-ISWYS [EvalSR-Opt-ISWYS]

function EvalSR-Opt-ISWYS($\boldsymbol{\theta}, \boldsymbol{\lambda}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D}$)
  $(\boldsymbol{\mu}_0^{\mathrm{p}}, d_0, x_0^{\mathrm{i}}) \leftarrow \mathsf{D}$
  $(\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}}\hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^{\mathrm{z}}))$ ▷ as described in Sec. 5.5.2
  $\left(v, \nabla_{\boldsymbol{\theta}}, (\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, m_t^{Q^\star}, m_t^{Q^\star,1}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, m_t^{V^\star}, m_t^{V^\star,1}, \mathbf{F}_t^\star, \mathbf{f}_t^\star, \pi_t^\star(u_t^{\mathrm{z}}, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}}))_{t=0:T}\right) \leftarrow$ EvalSROpt$(\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D})$
  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} Q_T^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \leftarrow \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(d_T)\,\mathrm{vec}(\boldsymbol{\Theta}_1)$
  for $t = T-1:0$ **do**
    $\forall_{d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t)\,\mathrm{vec}(\boldsymbol{\Theta}_1) + \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t)\,\mathrm{vec}(\mathbf{A}_t^\top \mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}} \mathbf{A}_t) - \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}(d_{t+1}(d_t, u_t^{\mathrm{z}}))\,\mathrm{vec}(\mathbf{M}_{t+1}^{V^{\star,\boldsymbol{\theta}}})$
    $\forall_{d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} Q_t^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow^+ \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} Q_{t+1}^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) \middle| \pi^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$
  end for
  $\nabla_{\boldsymbol{\lambda}} \leftarrow \mathbb{E}[\nabla_{\boldsymbol{\lambda}} Q_0^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(d_0, x_0^{\mathrm{i}}, u_0^{\mathrm{z}}, u_0^{\mathrm{i}}) | \pi^{\star,\boldsymbol{\theta},\boldsymbol{\lambda}}(u_0^{\mathrm{z}}, u_0^{\mathrm{i}} | d_0, x_0^{\mathrm{i}})]$
  return $(v, \nabla_{\boldsymbol{\theta}}, \nabla_{\boldsymbol{\lambda}})$
end function

Algorithm 18 Evaluation of SR-MCE-ISWYS [EvalSR-MCE-ISWYS]

function EvalSR-MCE-ISWYS($\boldsymbol{\theta}, \boldsymbol{\lambda}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \mathcal{P}^i, \mathsf{D}$)
  $(\boldsymbol{\mu}_0^{\mathrm{p}}, d_0, x_0^{\mathrm{i}}) \leftarrow \mathsf{D}$
  $(\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}}\hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z))$ ▷ as described in Sec. 5.5.2
  $\left(v, \nabla_{\boldsymbol{\theta}}, \left(\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \tilde{\pi}_t(u_t^z, u_t^{\mathrm{i}} | d_t, x^{\mathrm{i}})\right)_{t=0:T}\right) \leftarrow$ EvalSRMCE$(\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z), \mathcal{P}^i, \mathsf{D})$
  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^z, u_T^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_T^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_T, x_T^{\mathrm{i}}, u_T^z, u_T^{\mathrm{i}}) \leftarrow \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(d_T) \operatorname{vec}(\boldsymbol{\Theta}_1)$
  for $t = T-1 : 0$ **do**
    $\forall_{d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}) \leftarrow \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t) \operatorname{vec}(\boldsymbol{\Theta}_1) + \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t) \operatorname{vec}(\mathbf{A}_t^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \mathbf{A}_t) - \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}(d_{t+1}(d_t, u_t^z)) \operatorname{vec}(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}})$
    $\forall_{d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^z, u_t^{\mathrm{i}}) \leftarrow^{+} \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^z, u_{t+1}^{\mathrm{i}}) \middle| \tilde{\pi}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_{t+1}^z, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^z, u_t^z; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$
  end for
  $\nabla_{\boldsymbol{\lambda}} \leftarrow \mathbb{E}[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_0^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_0, x_0^{\mathrm{i}}, u_0^z, u_0^{\mathrm{i}}) | \tilde{\pi}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_0^z, u_0^{\mathrm{i}} | d_0, x_0^{\mathrm{i}})]$
  return $(v, \nabla_{\boldsymbol{\theta}}, \nabla_{\boldsymbol{\lambda}})$
end function

Algorithm 19 Evaluation of SR-MCL-ISWYS [EvalSR-MCL-ISWYS]

function EvalSR-MCL-ISWYS($\boldsymbol{\theta}, \boldsymbol{\lambda}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D}$)
  $(\boldsymbol{\mu}_0^{\mathrm{p}}, d_0, x_0^{\mathrm{i}}) \leftarrow \mathsf{D}$
  $(\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}}\hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^{\mathrm{z}}))$ ▷ as described in Sec. 5.5.2
  $\left(v, \nabla_{\boldsymbol{\theta}}, (\mathbf{M}_t^{\tilde{Q}}, \mathbf{m}_t^{\tilde{Q}}, m_t^{\tilde{Q},1}, \mathbf{M}_t^{\tilde{V}}, \mathbf{m}_t^{\tilde{V}}, m_t^{\tilde{V},1}, \tilde{\mathbf{F}}_t, \tilde{\mathbf{f}}_t, \tilde{\pi}_t(u_t^{\mathrm{z}}, u_t^{\mathrm{i}} | d_t, x^{\mathrm{i}}))_{t=0:T}\right) \leftarrow$ EvalSRMCL$(\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^{\mathrm{z}}), \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^{\mathrm{z}}), \mathcal{P}^i, \mathsf{D})$
  $\forall_{d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_T^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \leftarrow \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(d_T) \operatorname{vec}(\boldsymbol{\Theta}_1)$
  $\nabla_{\boldsymbol{\lambda}} \leftarrow \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_T^{\mathrm{p},\boldsymbol{\lambda}}(d_T) \operatorname{vec}(\mathbf{M}_T^{\tilde{V}} - \mathbf{M}_{T\,x,x}^{\tilde{Q}}) + \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_T^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \middle| \tilde{\pi}(u_T^{\mathrm{z}}, u_T^{\mathrm{i}} | d_T, x_T^{\mathrm{i}})\right] - \nabla_{\boldsymbol{\lambda}} \tilde{Q}_T^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_T, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}}) \middle| \mathsf{D}\right]$
  for $t = T : 0$ **do**
    $\forall_{d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t) \operatorname{vec}(\boldsymbol{\Theta}_1) + \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t) \operatorname{vec}(\mathbf{A}_t^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \mathbf{A}_t) - \nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_{t+1}^{\mathrm{p},\boldsymbol{\lambda}}(d_{t+1}(d_t, u_t^{\mathrm{z}})) \operatorname{vec}(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}})$
    $\forall_{d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}} \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \leftarrow^+ \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_{t+1}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_{t+1}, x_{t+1}^{\mathrm{i}}, u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}}) \middle| \tilde{\pi}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_{t+1}^{\mathrm{z}}, u_{t+1}^{\mathrm{i}} | d_{t+1}, x_{t+1}^{\mathrm{i}}), \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}} | x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\right]$
    $\nabla_{\boldsymbol{\lambda}} \leftarrow^+ \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}}\boldsymbol{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(d_t) \operatorname{vec}(\mathbf{M}_t^{\tilde{V}} - \mathbf{M}_{t\,x,x}^{\tilde{Q}}) + \mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \middle| \tilde{\pi}(u_t^{\mathrm{z}}, u_t^{\mathrm{i}} | d_t, x_t^{\mathrm{i}})\right] - \nabla_{\boldsymbol{\lambda}} \tilde{Q}_t^{\boldsymbol{\theta},\boldsymbol{\lambda}}(d_t, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \middle| \mathsf{D}\right]$
  end for
  return $(v, \nabla_{\boldsymbol{\theta}}, \nabla_{\boldsymbol{\lambda}})$
end function

Algorithm 20 Evaluation of STR-Opt-ISWYS [EvalSTR-Opt-ISWYS]

function EvalSTR-Opt-ISWYS($\boldsymbol{\theta}, \boldsymbol{\lambda}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \mathcal{P}^i, \mathsf{D}$)
  $(\boldsymbol{\mu}_0^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:0}) \leftarrow \mathsf{D}$
  $(\boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}, \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}) \leftarrow$ Initialize$(\mathsf{D}, \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z))$ ▷ as described in Sec. 5.5.2
  $\nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \leftarrow \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}$
  $\boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \leftarrow \boldsymbol{\Sigma}_0^{\mathrm{p},\boldsymbol{\lambda}}$
  $\left(v, \nabla_{\boldsymbol{\theta}}, \left(\mathbf{M}_t^{Q^\star}, \mathbf{m}_t^{Q^\star}, \mathbf{M}_t^{V^\star}, \mathbf{m}_t^{V^\star}, \mathbf{F}_t^\star, \mathbf{f}_t^\star\right)_{t=0:T}, \boldsymbol{x}^{\mathbf{z}}{}_{0:T}{}^\star\right) \leftarrow$ EvalSTROpt$(\boldsymbol{\theta}, (\mathbf{A}_t, \mathbf{a}_t, \mathbf{B}_t)_{t=0:T}, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}(x_t^z), \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z), \mathcal{P}^i, \mathsf{D})$
  $\nabla_{\boldsymbol{\lambda}} = 0$
  for $t = 0 : T-1$ **do**
    $\nabla_{\boldsymbol{\lambda}} \leftarrow^+ \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \operatorname{vec}(\boldsymbol{\Theta}_1) + \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{A}_t^\top \mathbf{M}_{t+1}^{V^\star,\boldsymbol{\theta}} \mathbf{A}_t)$
    $\nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \leftarrow$ KalmanDerivative$(\nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}}, \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}}, \mathbf{A}_t, \boldsymbol{\Sigma}^{\epsilon^x}, \mathbf{H}, \boldsymbol{\Sigma}^{\epsilon^z,\boldsymbol{\lambda}}(x_t^z{}^\star))$ ▷ according to Sec. 5.5.2
    $\nabla_{\boldsymbol{\lambda}} \leftarrow^- \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \operatorname{vec}(\mathbf{M}_{t+1}^{V^\star,\boldsymbol{\theta}})$
  end for
  $\nabla_{\boldsymbol{\lambda}} \leftarrow^+ \nabla_{\boldsymbol{\lambda}} \boldsymbol{\Sigma}^{\mathrm{p},\boldsymbol{\lambda}} \operatorname{vec}(\boldsymbol{\Theta}_1)$
  return $(v, \nabla_{\boldsymbol{\theta}}, \nabla_{\boldsymbol{\lambda}})$
end function

5.6 A Real-Traffic Driving Experiment

Previously, we introduced algorithms to infer sensor models underlying driver behavior in the context of the joint task POMDP. In application to real world driver behavior, here several assumptions are made. First, we assume that the joint task POMDP, specifically the used linear Gaussian sensor model allows to model the real relations sufficiently well. Second, we impose the assumption, that drivers act at least rational with respect to their true sensor characteristics. Note, that this is also the prerequisite in signal detection experiments, although in this context less complex decision making of the human subject is required. Hence, it is necessary to evaluate the derived approaches for inference of sensor models on real data.

[230] investigated the lane keeping performance of driver that monitored a small display placed at different positions in the vehicle's cockpit. The positions employed were directly above the steering wheel, at the position of the vehicle's speedometer and at the position of the vehicles radio. Here, the results showed significant differences in the lane keeping performance for the driver gazing at the individual display positions. The authors hypothesized that these differences could likely be attributed to varying quality of sensing the vehicle's states by means of peripheral vision. Consequently, such an experiment can provide suitable data for evaluating the approaches for inference of sensor models.

5.6.1 Protocol

We conducted our variant of the experiment of [230] on a segment of the German motorway A81 which is depicted in Fig. 5.2. Here we increased the length of the segment of driving experiment I (Sec. 4.6) and collected the behavioral data between motorway exit *10 Weinsberg* and *13 Mundelsheim*.

Figure 5.2: Segment of German motorway A81 used for the driving experiment. The recordings were obtained driving between exit *10 Weinsberg* and *13 Mundelsheim*. Obtained from [170], License CC-BY-SA 2.0

The participants were 17 drivers (16 male, 1 female) recruited from the Robert Bosch Group. Similar to the previous experiment, only drivers were selected that had previously taken an in-house driving safety training.

The experiment consisted of *three* fixed driving speed conditions $\{80, 90, 110\}$ km/h. We used the vehicle's Adaptive Cruise Control (ACC) for speed control and to ensure a conservative time gap to the next vehicle. When the vehicle traveled at the required speed, either a reference period or a secondary task period were triggered by the investigator. At each speed three samples of baseline driving without a secondary task and three samples of driving with each of three variants of the secondary task were

taken. These experimental conditions are summarized in Tab. 5.1.

Tabular 5.1: Experimental Conditions of Driving Experiment II

Display Positions	**Driving Speeds**		
	80 km/h	90 km/h	110 km/h
None	3×	3×	3×
H	3×	3×	3×
C	3×	3×	3×
N	3×	3×	3×

Here, we used the task of typing random numbers $\{1, 2\}$ introduced in Sec. 3.3.3 and applied in the driving experiment of Sec. 4.6. In total 30 random numbers were displayed one at a time using the small screen depicted in Fig. 5.3.

Figure 5.3: The display used for the secondary task in the driving experiment.

Following [230] we investigated three similar display positions which are shown in Fig. 5.4. Specifically, the display was first put right above the steering wheel at the vehicle's wind screen. Second, the display was set above the vehicle's r.p.m. counter in the kombi-instrument. Finally and third, the display was put to approximately the same position that was used in driving experiment I (Sec. 4.6).

Figure 5.4: The positions of the display used in the experiment. **H** is a display position that requires an amount of gaze aversion similar to a Head-Up display, **C** is the position of the vehicle's r.p.m. counter and **N** corresponds to the vehicle's built in display for navigation.

Reading the displayed numbers required varying amounts of gaze aversion. This is exemplary illustrated at one participant in Fig. 5.5. As can be seen gazing at the display at position H corresponding to a head-up display required only a very small amount of aversion of gaze from the road scenery. In contrast reading the numbers at display position N was often only possible by turning the head.

Figure 5.5: Gaze aversions required to conduct the typing task presented at the different display positions. From left to right, the pictures show the driver gazing at the road, gazing at the display at position H, gazing at the display at position C and gazing at the display at position N. Personal agreement of the depicted participant was obtained.

Similar as in the previous driving experiment (Sec. 4.6), the participants were instructed to *"perform the secondary task as quickly and correctly as possible while not endangering driving safety"*.

5.6.2 Recorded Data and Preprocessing

The MPC2 camera system (Robert Bosch GmbH, Stuttgart, Germany) was used to obtain the position of the lane y_t, the vehicles orientation ϕ_t and the curvature κ_t as in the previous driving experiment. Furthermore, the driver's gaze was tracked by means of a SmartEye Pro system (SmartEye AB, Gothenburg, Sweden). We recorded steering wheel position and velocity as well as absolute velocity from the vehicles CAN-Bus. Of the recorded data, lane changes and their preparation phases were excluded. Furthermore, situations where the ACC controller reduced the vehicle speed by more than 10% were discarded. The final data set consisted of 585 valid segments comprising of 141 reference and 444 secondary task periods with an average duration of 25.3 s. Finally, the same preprocessing steps as in the first driving experiment (Sec. 4.6.2) were conducted.

5.6.3 Behavioral Statistics

We first investigate the important behavioral statistics found in the data of the driving experiment. Here, the same methodology of analysis as in the first driving experiment (see Sec. 4.6.3) is employed. In this context, we will report on the distribution of the durations of glances off the road $\max(d_t)$ as well as the statistics of the lane position y_t.

The statistics of the duration of glances off the road are reported in Tab. 5.2.

Tabular 5.2: Statistics of Glance Behavior

Display Pos.	**Statistics**	**Driving Speeds**		
		80 km/h	90 km/h	110 km/h
H	[0.05, 0.50, 0.95] Quantiles	0.28, 0.84, 3.40 s	0.32, 0.88, 4.10 s	0.30, 0.64, 2.64 s
	Mean	1.37 s	1.41 s	0.98 s
C	[0.05, 0.50, 0.95] Quantiles	0.40, 1.00, 3.61 s	0.36, 0.92, 3.68 s	0.33, 0.92, 2.24 s
	Mean	1.36 s	1.30 s	1.03 s
N	[0.05, 0.50, 0.95] Quantiles	0.56, 1.12, 2.56 s	0.56, 1.04, 2.42 s	0.48, 0.96, 2.00 s
	Mean	1.32 s	1.21 s	1.05 s

With respect to the mean duration of glances off the road for all displays a significant decrease from 90 km/h to 110 km/h could be established $p_{\text{test}} < 0.01$. Furthermore, for the display position N also the decrease from 80 km/h to 90 km/h was significant $p_{\text{test}} < 0.01$. All other differences were not significant $p_{\text{test}} > 0.01$. Among the $0.05, 0.25, 0.5, 0.75, 0.95$ quantiles only the 0.75 for 90 km/h to 110 km/h showed a significant decrease for every display position $p_{\text{test}} < 0.01$ according to the quantile test of [86]. All other differences were not significant $p_{\text{test}} > 0.01$. The distributions of the duration of glances off the road are depicted in Fig. 5.6.

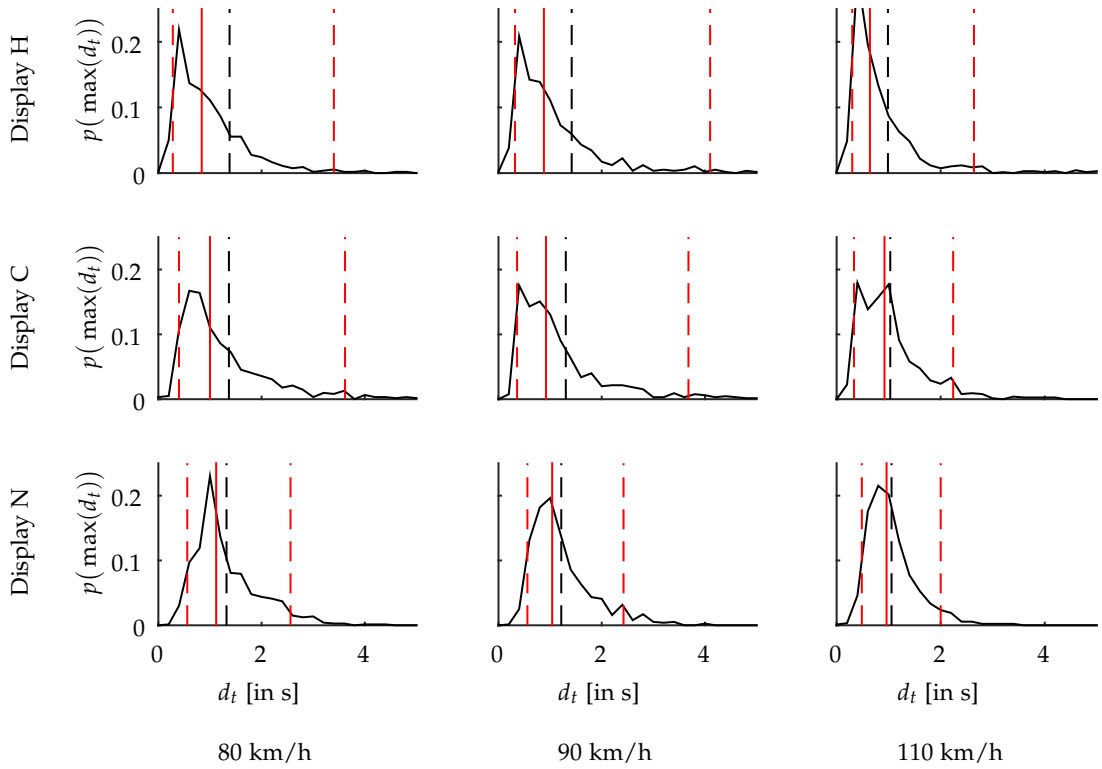

Figure 5.6: Distributions of glances off the road $\max(d_t)$ at the different driving speeds and display positions. Dashed red lines indicate the $[0.05, 0.95]$ quantiles, while the solid red lines indicate the median. The mean duration of glances off the road is denoted by a dashed black line.

Comparing the distributions of glances for the individual display over all speed conditions the following observations, summarized in Tab. 5.3, were made:

Tabular 5.3: Quantiles of the Durations of Glances Off the Road

Display Position	**Quantiles**				
	0.05	0.25	0.50	0.75	0.95
H	0.32s	0.44s	0.80s	1.28s	3.45s
C	0.36s	0.60s	0.92s	1.44s	3.03s
N	0.52s	0.80s	1.04s	1.44s	2.36s

The $0.05, 0.25, 0.50, 0.75$ quantiles showed an increase from display H to C and to N. However, the highest quantile of 0.95 showed a decrease from H to C and to N. All differences in the quantiles were significant $p_{\text{test}} < 0.01$ expect for the difference of the durations of glances at C and N and the 0.75 quantile. With respect to mean durations of glances no significant differences could be established.

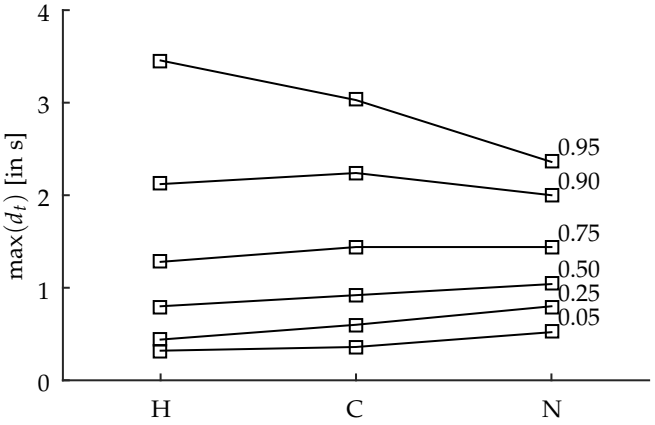

Figure 5.7: Quantiles of the durations of glances off the road $\max(d_t)$ in the driving experiment II. Plot shows the $0.05, 0.25, 0.50, 0.75, 0.90, 0.95$ quantiles.

With respect to lane keeping performance the individual display positions resulted in the following statistics: Driving without a secondary task present resulted in a median STD of lane position of 0.159 m. The median STDs for display position H was 0.160 m, for display position H was 0.166 m and for display position H was 0.166 m. Here, differences turned out be not statistically significant $p_{\text{test}} > 0.01$ according to sum-rank test. With respect to the RMSE of the lane position a median of 0.223 m was obtained for attentive driving, a median of 0.266 m was obtained for display position H, a median of 0.304 m was obtained for display position C and a median of 0.282 m was obtained for display position N. A sum-rank test revealed significantly higher median RMSE for all display positions compared to attentive driving $p_{\text{test}} < 0.01$. All other differences were not significant $p_{\text{test}} > 0.01$ although the differences in RMSE of display position H to both display position C and display position N were close to significance $p_{\text{test}} = 0.08$.

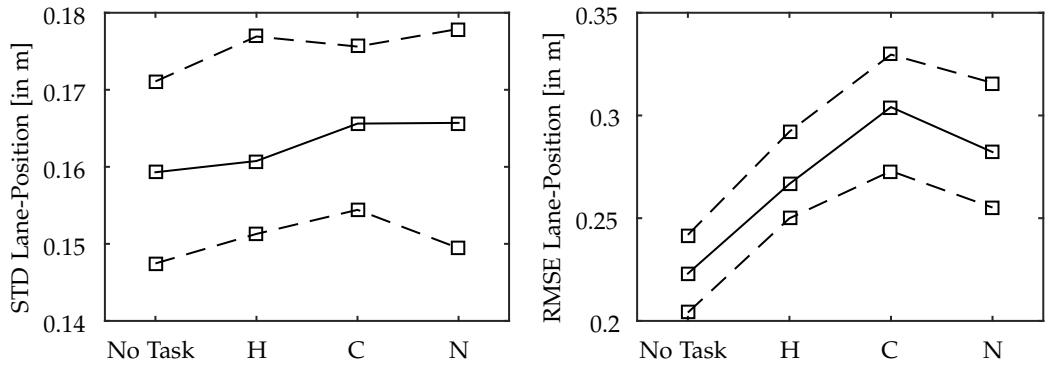

Figure 5.8: Left plot depicts the statistics of the standard deviation of the lane position, right plot shows the statistics of the root mean squared deviation from the lane center. Solid line indicates the median of the statistic, whereas the dashed lines indicate the $[0.05, 0.95]$ confidence interval.

5.6.4 Discussion

Comparing the behavioral statistics of experiment II (Sec. 5.6.3) to those of experiment I (Sec. 4.6.3), glance statistics and the statistics of the lane position are largely similar. In experiment II the distribution of the durations of glances off the road and the median STD as well as the median RMSE of the lane position for display position N almost coincided with those of the secondary task in experiment I. As can be seen by comparing Fig. 4.9 and Fig. 5.4 the display positions in these cases were almost the same. The RMSE of the lane position was similar sensitive to the presence of secondary task in

experiment I and for display position N. In contrast, in the STD of the lane position only effects were present in experiment I.

Reading the generated numbers from the display position H, C and N requires gaze aversion of increasing angles, as can be seen in Fig. 5.5. Hence, under the assumption of rational drivers we expected decreased durations of glances at display position H to glances at display position N. However, such a decrease could only be established for the 0.95 quantile of the durations. In addition to this, the lower quantiles showed an increase from the display position that required least gaze aversion to the display position that required largest gaze aversion. We conclude that drivers did not only adapt to the experimentally manipulated amount of peripheral vision. The significantly shorter glances to the display that required small gaze aversion indicate that the cost of switching gaze seems to be strongly related to the required angular distance. This may be the result of increased muscular effort and longer saccadic suppression of vision. This hypothesis is also supported by the statistics of lane keeping: Corresponding to the increased lower quantiles of the durations of glances off the road, slightly increased RMSE of the lane position for display positions C and N compared to display position H were found.

Comparing the protocol of experiment II (Sec. 5.6.1) to the protocol used in [230] to investigate the role peripheral vision in driving the differences are established: First, in our experiment participants were free to chose their glance behavior, whereas the participants of [230] were forced to avert their gaze from the road. Correspondingly, glances to the road were more frequent in our experiment. In the experimental protocol used in [230] a significantly decreased proportion of the drivers were not able to keep the vehicle in lane while gazing at display positions similar to C and N. In contrast, in our experiment the drivers' lane keeping performance was significantly degraded but the differences between the different display positions were less pronounced. Furthermore, in [230] the angular amount of gaze aversion correlated with more frequent glances at the road whereas in our experiment only the highest quantile of the duration off glances off the road showed a decrease.

We conclude the analysis of behavioral statistics establishing that the experiment succeeded in inducing driver adaption to experimental manipulation of the required amount of glance aversion. No significant variation of the lane keeping performance among the display positions was found. This could be attributed to a decreased duration of long glances off the road for the display positions that required stronger gaze aversion. However, we also noted that the different display position apparently corresponded to different costs of gaze switching which was visible in the distribution of short glances off the road. Consequently, the obtained data can be used to evaluate the developed approaches for inference of sensor models. Nevertheless, to estimate final sensor model parameters the experimental protocol must be improved with respect to mitigation of the effects of different costs of gaze switching.

5.7 Evaluation On Real Traffic Data

In the present section we will report on an evaluation of the developed technique for inference of sensor models under the maximum causal entropy policy model. Similar as in the evaluation of inverse optimal control approaches, here we investigate the prediction performance of policies. Assuming that driver's act rationally with respect to their sensor characteristics prediction performance can be related to quality of estimation of the sensor model. Note, that similar premises are imposed in signal detection experiments.

5.7.1 Scenario

To evaluate our approach for inferring sensor models from gaze switching behavior and lane keeping performances we compared to alternative techniques. Here, we adapted the approaches [176, 70] to the application context of this work. In the resulting methods, sensor models are inferred from their influence on lane keeping performance *alone*. Furthermore, it is investigated, whether estimating sensor models can provide any benefits over the best-guess sensor model used in the evaluation of Sec. 4.5 in previous chapter. As driver behavior in the experiment II partially differed from that of experiment I, also the generic directly estimated baseline policy model (Sec. 4.5.1) is revisited.

Methods for Inference of Sensor Models

In the following we will now describe the specific methods for estimating sensor models which we evaluated on behavioral data obtained in driving experiment II (Sec. 5.6.3).

ISWYS in The Joint Task POMDP In this evaluation we considered our approach for inference of policy and sensor models for the joint task POMDP under sensor model restriction *and* the simple secondary task and under the maximum causal entropy policy model, i.e. SRMCE-ISWYS Algo. 18. Unfortunately, the enormous computational demand required to obtain optimal sensor sequences for the joint task model without the restriction of the sensor model prevented an evaluation of STROpt-ISWYS Algo. 20.

The parameters of the model of the driving task were set to those values previously inferred in the preprocessing of the data (Sec. 4.6.2). The reward function on the primary task states (3.9) was used as in the previous evaluations. We further used task model $x_t^{\mathrm{i}} = x_t^{\mathrm{z}}$, $r(x_t^{\mathrm{i}}) = \theta(1 - x_t^{\mathrm{i}}) = \theta(1 - x_t^{\mathrm{z}})$ of Sec. 3.3.3.

The sensor model parameters inferred in the evaluation were the noise variance in sensing the vehicle's position in lane $(\sigma_y)^2(x_t^{\mathrm{z}} = 1)$ and orientation in lane $(\sigma_\phi)^2(x_t^{\mathrm{z}} = 1)$ of the sensor noise covariance $\mathbf{\Sigma}^{\epsilon^{\mathrm{z}}}(x_t^{\mathrm{z}} = 1) = \mathrm{diag}((\sigma_y)^2(x_t^{\mathrm{z}} = 1), 0, (\sigma_\phi)^2(x_t^{\mathrm{z}} = 1), 0)$. Here, separated noise covariances were used for the different display positions. The parameters of the sensor model for gazing at the road, i.e. $x_t^{\mathrm{z}} = 0$ turned out to be not identifiable: In first tests estimating these parameters resulted in highly unstable optimization of the soft gap (5.45). Therefore we omitted to estimate these quantities and set the sensor noise covariance to $(\sigma_y)^2(0) = 0.64, (\sigma_\phi)^2(0) = 0.01, (\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty$. For long glances at the road this choice of parameters resulted in a steady-state belief of the lane position y with a 0.96 confidence interval of reasonable 0.3 m. Note, that the same values have been used to depict the belief in the lane position when we introduced the joint task POMDP with restriction of the sensor model in Sec. 3.5.2. Reward parameters $\boldsymbol{\theta}$, including the parameter of the reward function on gaze switches, were shared while sensor model parameters were inferred for each of the individual display positions.

To ensure well defined parameters $\boldsymbol{\theta}, \boldsymbol{\lambda}$ the following constraints were imposed: The parameters of the quadratic function of the primary task states and controls were required to be smaller than -10^{-4} whereas the variance of the sensor noise on the lane position and the orientation were required to be greater than 10^{-6}. The final optimization problem comprising of (5.45) plus the additional imposed constraint was solved using an interior point method for non-linear, non-convex optimization.

Phatak et. Al.s' Method Adapted to Joint Task POMDP [176] proposed an approach for inference of sensor models and rewards in infinite time-invariant LQGs under the optimal policy model. Here, a policy was directly estimated under the constraints introduced in [101] that guarantee optimality of the estimated policy for some reward and some sensor model.

The joint task POMDP model is time variant and the sensor model varies according to the driver's glance behavior (Sec. 3.3.2). Hence, the original approach of [176] is not applicable. Furthermore, a rational maximum causal entropy policy allows for more flexibility than a optimal policy. Still, the idea of fitting a rational primary task policy $\tilde{\pi}^{\boldsymbol{\theta},\boldsymbol{\lambda}}(u_t^{\mathrm{p}}|\boldsymbol{\mu}_t^{\mathrm{p}})$ to the observed controls given the distribution $\mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}}|\mathbf{x}_t^{\mathrm{p}}, \mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}))$ of the unknown driver's belief on the primary task states in the data D can be applied. Specifically, we can infer the reward parameter of the primary task $\mathbf{\Theta}_1, \mathbf{\Theta}_2$, and the sensor model parameters $\boldsymbol{\lambda}$ by minimizing the expected negative log-likelihood of the maximum causal entropy *primary task policy*:

$$\min_{\mathbf{\Theta}_1,\mathbf{\Theta}_2,\boldsymbol{\lambda}} \mathbb{E}\Big[\sum_{t=0}^{T} \mathbb{E}\big[\log\big(\tilde{\pi}_t^{\boldsymbol{\theta}}(u_t^{\mathrm{p}}|\boldsymbol{\mu}_t^{\mathrm{p}})\big) | \mathcal{N}(\boldsymbol{\mu}_t^{\mathrm{p}}|\mathbf{x}_t^{\mathrm{p}}, \mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}))\big] \Big| \mathsf{D} \Big] = \tag{5.66}$$

$$\begin{aligned} \min_{\mathbf{\Theta}_1,\mathbf{\Theta}_2,\boldsymbol{\lambda}} \mathbb{E}\Big[\sum_{t=0}^{T} & [\mathbf{x}_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} [\mathbf{x}_t^{\mathrm{p}}] + \mathrm{tr}(\mathbf{M}_t^{\tilde{V}^{\boldsymbol{\theta}}} \mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathbf{m}_t^{\tilde{V}^{\boldsymbol{\theta}}}[\mathbf{x}_t^{\mathrm{p}}] \\ & - \big([\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}}[\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathrm{tr}(\mathbf{M}_{t\,x,x}^{\tilde{Q}^{\boldsymbol{\theta}}} \mathbf{\Sigma}_t^{\mathrm{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}[\mathbf{x}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]\big) \Big| \mathsf{D} \Big]. \end{aligned} \tag{5.67}$$

This optimization problem is a simplified variant of the SR-MCL-ISWYS. In SR-MCL-ISWYS the likelihood of the policy for all controls including the sensor control and secondary task controls is maximized (Sec. 5.46) whereas our version of Phatak et al.s' only maximizes the likelihood of the primary task policy. Consequently, Algo. 19 can be adapted for solution. To finally also obtain a policy for the sensor control u_t^z and the secondary task control u_t^{i} this approach was followed by SR-MCL in the evaluation. Here, the sensor model was fixed to the previously estimated sensor model parameters $\boldsymbol{\lambda}$ while all reward parameters were re-estimated.

For numerical optimization we used the same additional constraints and the same optimizer as in ISWYS.

Golub et. Al.s' Method Adapted to Joint Task POMDP Inference of internal models in LQGs given a linear Gaussian, e.g. maximum causal entropy policy for the primary task, was addressed in [70]. In that work an Expectation-Maximization (EM) approach was proposed which iterates between inference of the agent's belief unknown to the observer and inference of the agent's internal models. Here, both internal models of the sensor and the dynamics were addressed.

In our application the internal dynamics model is assumed to equal the true dynamics model and we only seek to estimate the sensor model. Consequently, Golub et. al.s' method corresponds to estimating sensor model parameters $\boldsymbol{\lambda}$ by minimizing the expected negative log-likelihood of a *fixed* primary task policy $\pi_t(u_t^{\text{p}}|\boldsymbol{\mu}_t^{\text{p}})$:

$$\min_{\boldsymbol{\lambda}} \mathbb{E}\Big[\sum_{t=0}^{T} \mathbb{E}\big[\log\left(\pi_t(u_t^{\text{p}}|\boldsymbol{\mu}_t^{\text{p}})\right)|\mathcal{N}(\boldsymbol{\mu}_t^{\text{p}}|\mathbf{x}_t^{\text{p}},\boldsymbol{\Sigma}_t^{\text{p},\boldsymbol{\lambda}}(\boldsymbol{x}^{\boldsymbol{z}}_{0:t}))\big]\Big|\mathsf{D}\Big]. \tag{5.68}$$

Similar as in case of Phatak et al.'s approach, the minimization problem can be solved by adapting Algo. 19. Note, that the original EM-approach used in [70] is just another technique for solving the same optimization problem. In the evaluation, we used the maximum causal entropy primary task policy $\tilde{\pi}_t^{\boldsymbol{\theta}}(u_t^{\text{p}}|\boldsymbol{\mu}_t^{\text{p}})$ for random parameters sampled from the range observed in the previous evaluation of inverse optimal control (Sec. 4.7).

Estimation was completed by fixing the previously inferred sensor model parameters $\boldsymbol{\lambda}$ followed by inverse optimal control according using SR-MCL. Here, the same additional constraints and the same optimizer as in ISWYS were employed.

Inverse Optimal Control

In addition to previous methods for inference of sensor models and policies, we also considered inverse optimal control. Here, we employed the maximum causal entropy approach under the same model assumptions as in ISWYS, i.e. SR-MCE Algo. 11. Furthermore, the simple secondary task model was used.

To evaluate the advantage of inference of sensor models, a "best guess" sensor model was applied in inverse optimal control: We set the sensor noise covariance for gaze on the road to $(\sigma_y)^2(0) = 0.64, (\sigma_\phi)^2(0) = 0.01, (\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty$ as in ISWYS and used a sensor noise covariance of $(\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty, (\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty$ for the driver gazing at the display. Note, that this sensor model differs from the one of the previous evaluation (4.7) where we assumed that the driver can perfectly sense the primary task states. In a pre-test, we empirically found slightly decreased prediction error using sensory noise for sensor state gaze at the road.

For numerical solution of the corresponding optimization problem (Sec. 4.4.4) the same additional constraints and the same optimizer as in ISWYS were employed.

Baseline

Finally, we also evaluated a baseline policy model obtained by direct policy estimation. For this purpose, the policy model

$$\pi^{\text{base}}(u_t^{\text{p}}, u_t^{\text{z}}|\boldsymbol{\mu}_t^{\text{p}}, d_t) \propto \mathcal{N}(u_t^{\text{p}}|\boldsymbol{\Lambda}_1^{\text{base}}\boldsymbol{\mu}_t^{\text{p}} + \boldsymbol{\lambda}_2^{\text{base}}, \boldsymbol{\Sigma}^{\text{base}}) \exp\left(u_t^{\text{z}}(\lambda_3^{\text{base}} d_t + \lambda_4^{\text{base}} x_t^{\text{z}} + \lambda_5^{\text{base}}(1 - x_t^{\text{z}}))\right), \tag{5.69}$$

was used. Inference of $\boldsymbol{\Lambda}_1^{\text{base}}, \boldsymbol{\lambda}_2^{\text{base}}, \lambda_3^{\text{base}}, \lambda_4^{\text{base}}, \lambda_5^{\text{base}}$ was conducted by L1-regularized maximum likelihood estimation for generalized linear models [165]. We term the baseline model as DPE1 in the evaluation.

5.7.2 Metrics

Following the evaluation in Sec. 4.7 the Kullback-Leibler divergence $\mathrm{KL}(p(d_t)||p'(d_t))$ was used to measure the difference between the distributions of eyes-off durations (EOD) in the experimental data $p(d_t)$ and the predicted distribution of EOD $p'(d_t)$.

Furthermore, the expected squared error between the true lane position y_t^i and predicted lane position y_t,

$$\mathrm{SE}(\mathbf{y}_{0:T}^i; p(\mathbf{y}_{0:T})) = \mathbb{E}\big[\frac{1}{T}\sum_{t=0}^{T}(y_t - y_t^i)^2 | \boldsymbol{\pi}_{0:T}, \boldsymbol{\mathcal{P}}_{0:T}, p_0\big] \tag{5.70}$$

was used in the present evaluation. As an alternative to the expected squared error in prediction of the lane position, we additionally considered the likelihood of the primary state trajectory present in the data

$$\mathrm{NLL}(\mathbf{x}_{0:T}^{\mathrm{p}\ i}; p(\mathbf{x}_{0:T}^{\mathrm{p}})) := -\frac{1}{T}\sum_{t=0}^{T}\log p(\mathbf{x}_t^{\mathrm{p}\ i} | \boldsymbol{\pi}, p^z, \boldsymbol{\mathcal{P}}, p_0), \quad \mathbf{x}_t^{\mathrm{p}\ i} \in \mathsf{D}. \tag{5.71}$$

In this context, a Gaussian approximation of the predictive primary state distribution was employed.

5.7.3 Protocol

For the numerical evaluation all valid segments of the driving experiment were split into snippets of 5 s length. One half of the data set was used to infer the latent parameters (training set) and the other half was used for evaluation (test set). The split was conducted randomly and independently of driver, speed or display position.

To estimate the parameters of DPE1 all training data was merged into a single set. In case of inverse optimal control and the methods for sensor model inference, first for every snippet in the training set the corresponding POMDP model was generated as previously described in Sec. 4.7.1. Here, in both types of approaches the sensor model for the sensor state gaze on road was fixed. In case of IOC also the sensor model for the sensor state gaze at display was fixed, whereas its noise covariance was estimated in the case of the other methods.

The evaluation on the test set was done in the following way: We first generated the POMDP models corresponding to each snippet. In case of the approaches that estimated sensor models, the sensor model corresponding to the position of the display in the snippet was used. For evaluation of inverse optimal control the "best guess" sensor model $(\sigma_y)^2(0) = 0.64, (\sigma_\phi)^2(0) = 0.01, (\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty$, $(\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty, (\sigma_y)^2(1) = \infty, (\sigma_\phi)^2(1) = \infty$ was employed. Thereafter, the maximum causal entropy policies given the inferred rewards were computed. Finally, the obtained policies were used to sample 100 state/control sequences in the POMDP associated with the snippet of the experimental data. Metrics were estimated based on the 100 samples.

5.7.4 Results

As a very first result, we would like to mention once again that *joint* inference of a sensor model for the sensor state $x_t^z = 1$, i.e. eyes on the road, *and* the driver's policy turned out to be impossible. For all applicable implementations found in Mathworks MATLAB optimization toolbox [238], the optimization problems of ISWYS and our variant of Phatak et al.'s approach were found to be unbounded. This was the case also when restricting the training set to a specific speed and/or to a specific display position. Interestingly, this was not the case for the used variant of Golub et. al.'s approach. However, in this case no significantly improved performance could be established when inferring both sensor noise covariances.

We summarize the results of the evaluation in Tab. 5.4. Similar as in previous evaluation, the distribution of prediction errors was found to be strongly skew shaped. Hence, we report on the medians of the individual metrics.

Tabular 5.4: Prediction Performance

Metrics	**Methods**									
	DPE1		MCE		Golub		Phatak		ISWYS	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
SE	0.081	0.080	0.020	0.020	0.020	0.021	0.020	0.020	0.020	0.020
NLL	−5.99	−6.03	−8.59	−8.62	−8.51	−8.61	−8.68	−8.55	−8.56	−8.67
KL	0.60	0.59	0.55	0.56	0.28	0.28	0.27	0.27	0.25	0.25

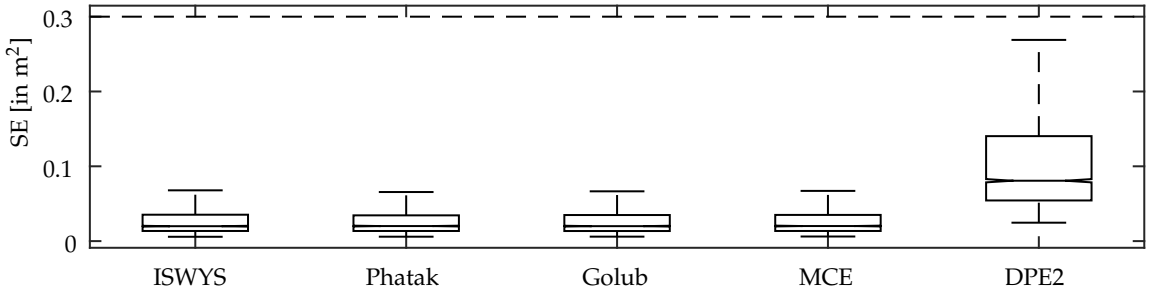

Figure 5.9: Expected squared error in prediction of the lane position. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. Plot depicts the prediction errors on withheld data.

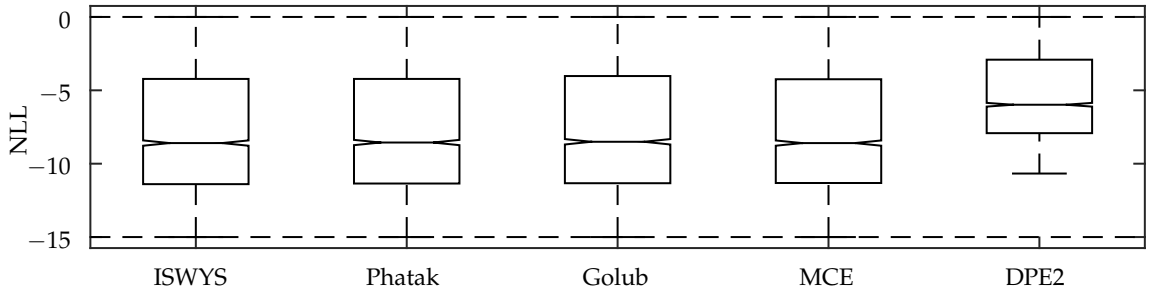

Figure 5.10: Likelihood of the sequences of primary task states under the predictive distribution. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. Plot depicts the prediction errors on withheld data.

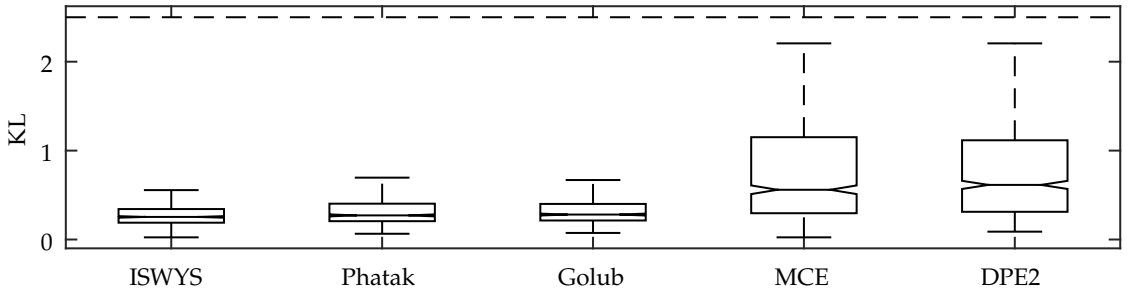

Figure 5.11: Difference of true distribution and predicted distribution of EOD measured by the Kullback-Leibler divergence. Box indicates the $[0.25, 0.75]$ interval, while the notch depicts the median. Whiskers indicate $1.5\times$ the median to quantile distance. Plot depicts the prediction errors on withheld data.

In the numerical experiment, the following observations were made: With respect to predicting the primary task states all other approaches outperformed DPE1 in both the metrics SE and NLL according to pairwise signed-rank tests $p_{\text{test}} < 0.01$. However, no significant differences between the approaches for inference of sensor model and inverse optimal control using the best guess sensor model could be established $p_{\text{test}} > 0.01$.

Considering prediction of glance behavior, the baseline model had significantly higher prediction error than the other approaches $p_{\text{test}} < 0.01$ as measured by the KL. However, the difference between the baseline and MCE was rather small. In contrast to the other metrics here inverse optimal control and the approaches for inference of sensor models showed significant variation in prediction error: First, using the best guess sensor model resulted in significantly higher prediction error compared to inference of individual sensor models $p_{\text{test}} < 0.01$. Second, while our variants of Phatak et al.'s and Golubs et al.'s approaches showed no significant difference, prediction error was slightly but significantly smaller for estimation of sensor models using ISWYS $p_{\text{test}} < 0.01$.

The approaches for inference of sensor models were used to estimate the variance of the sensor noise with respect to the vehicle's position in lane $\sigma^2(y_t)$ and the vehicle's orientation in lane $\sigma^2(\phi_t)$. The parameters obtained in the 5 different splits are shown in Fig. 5.12. In this context, differences in the

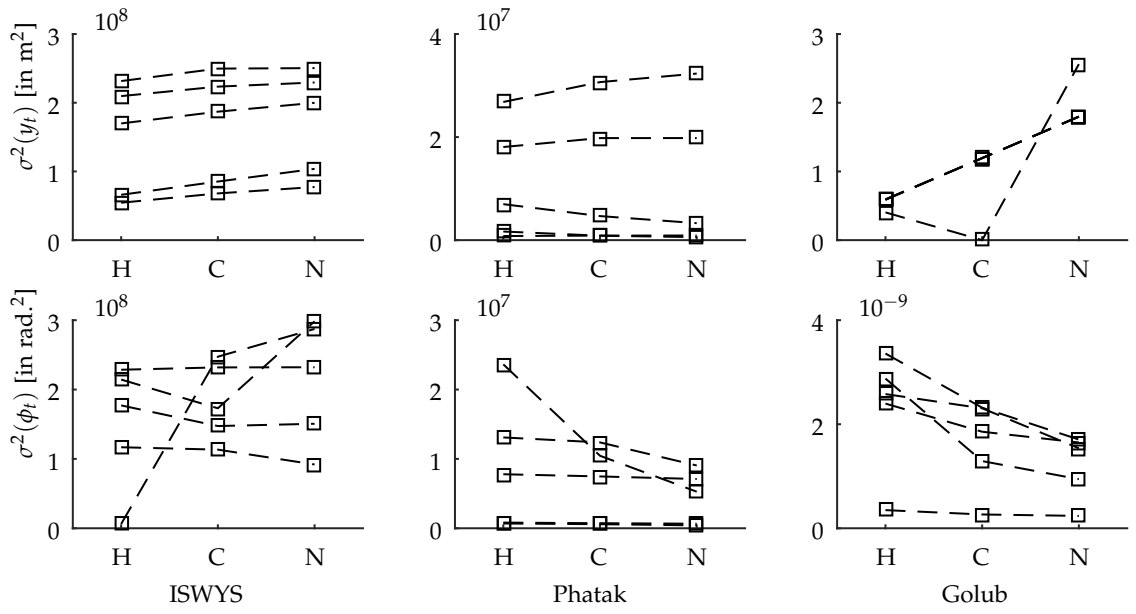

Figure 5.12: Inferred sensor noise parameters in experiment II. Upper plots depict the variance of the noise in sensing the vehicle's position in lane. Lower plots depict the variance in sensing the vehicle's orientation in lane.

values for the individual display positions H, C, N were not significant in neither method according to signed-rank test at a niveau of $p_{\text{test}} = 0.01$. However, in the case of ISWYS a strong tendency towards increasing $\sigma^2(y_t)$ could be established (H to C $p_{\text{test}} = 0.03$, C to N $p_{\text{test}} = 0.03$) and also a tendency (H to C $p_{\text{test}} = 0.06$, C to N $p_{\text{test}} = 0.03$) could be found in our variant of Golub et al.'s approach.

Finally, we wish to provide some anecdotal evidence from the evaluation. In Fig. 5.13 and Fig. 5.14 sample sequences obtained from the reward parameters estimated by IOC and obtained from the reward and sensor model parameters inferred by ISWYS are shown. Here, we depict the "true" lane position of the vehicle in the lane as well as the driver's belief of the lane position at driving speed 80 km/h for display position H.

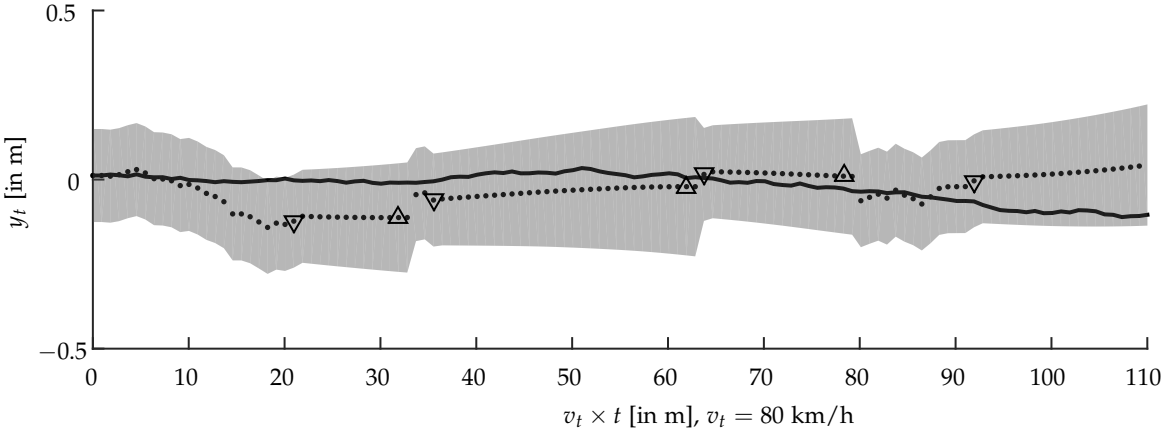

Figure 5.13: A sample sequence resulting from inference of reward under the best guess sensor model using IOC. Thick line $(-)$ denotes the "true" lane position y_t, dotted line $(\cdots)$ the expected lane position $\mathbb{E}[y_t|b(\mathbf{x}^{\mathrm{p}})]$ under the driver's belief $b(\mathbf{x}^{\mathrm{p}})$, shaded area the 96% confidence interval of the belief $b(\mathbf{x}^{\mathrm{p}})$. Triangles ∇ indicate gaze switches $u_t^z = 1$ from the road to display H, triangles Δ indicate gaze switches $u_t^z = 1$ from the display H to the road.

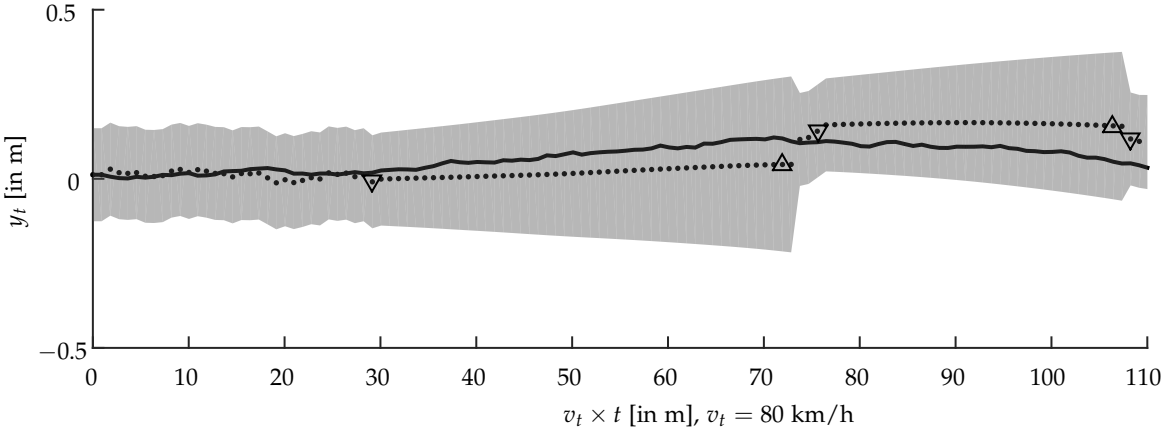

Figure 5.14: A sample sequence resulting from inference of reward and sensor model parameters using ISWYS. Thick line $(-)$ denotes the "true" lane position y_t, dotted line $(\cdots)$ the expected lane position $\mathbb{E}[y_t|b(\mathbf{x}^{\mathrm{p}})]$ under the driver's belief $b(\mathbf{x}^{\mathrm{p}})$, shaded area the 96% confidence interval of the belief $b(\mathbf{x}^{\mathrm{p}})$. Triangles ∇ indicate gaze switches $u_t^z = 1$ from the road to display at H, triangles Δ indicate gaze switches $u_t^z = 1$ from the display at H to the road.

As can be seen in Fig. 5.13 the best guess sensor model typically resulted in shorter glances off the road than those found in the data for display position H. In contrast, inference of the underlying sensor model lead to more realistic longer glances off the road. Notably, despite the high noise variance $\sigma^2(y_t)$ of 10^8 inferred by ISWYS the confidence interval of the driver's belief has a relatively small maximum width of 0.5 m.

5.8 Discussion

In the numerical evaluation we faced issues with unbounded optimization problems when trying to infer the reward parameters and the sensor model parameters for both sensor states "gaze on the road" and "gaze off the road". In this context we employed the maximum causal entropy policy model which results in a stochastic sub-optimal policy. Therefore, our empirical observations are in line with the theoretical analysis of over-parametrization of models of sensori-motor behavior in [176, 3]. [176] showed that the noise in sensing and the noise in execution of controls can generally not be separated if both the belief and the policy are time-invariant. Specifically, if no data of the sensory measurements

obtained by the human agent is available, deviation from the optimal linear-affine policy $u_t^{\mathrm{p}} = \pi_t^\star(\mathbf{x}^{\mathrm{p}})$ in LQGs can be explained by a stochastic policy $\pi_t^1(u_t^{\mathrm{p}}|\mathbf{x}^{\mathrm{p}})$: $u_t^{\mathrm{p}} = \pi_t^\star(\mathbf{x}^{\mathrm{p}}) + \boldsymbol{\epsilon}^1$ (for example the MCE policy model), the optimal policy acting on noisy estimates of the state $\pi_t^2(u_t^{\mathrm{p}}|\mathbf{x}^{\mathrm{p}}) = \pi_t^\star(u_t^{\mathrm{p}}|\boldsymbol{\mu}^{\mathrm{p}}) = \pi_t^\star(u_t^{\mathrm{p}}|\mathbf{x}^{\mathrm{p}} + \boldsymbol{\epsilon}^2)$ or a combination of both.

In the present application the linear-affine policy for the steering angle velocity is time-variant and dependent on the external variables v_t, κ_t which vary among the snippets of behavioral data used in estimation. However, we used a time-invariant belief with covariance $\hat{\boldsymbol{\Sigma}}_0^{\mathrm{p},\boldsymbol{\lambda}}$ for the sensor state "gaze on the road" in the applied estimation variant SRMCE-ISWYS (see evaluation scenario Sec. 5.7.1 and algorithm definition Algo. 18). Consequently, for this sensor state sensory and policy noise is probably not separable. In our variant of Golub et al.s' approach this issue is avoided, as the policy and sensor model are not estimated simultaneously. However, this came at the cost of higher KL compared to ISWYS in the numerical evaluation.

If the driver averts his or her gaze uncertainty in the belief $b(\mathbf{x}_t^{\mathrm{p}})$ monotonously increases which can be seen in Fig. 5.14. Therefore, the contribution of perceptual uncertainty to sub-optimal choice of controls changes over time. Likely, this variation allowed to separate the contribution of the sensorial noise from the contribution of noise in the execution of actions. This could be the reason why joint estimation of a sensor model for the sensor state "gaze off the road" and a stochastic policy was possible in the evaluation.

The results of the numerical show a clear benefit of inferring sensor models for the quality of predicting glance behavior. As shown in Fig. 5.12 different sensor model parameters for the individual display positions were estimated. Hence, improved prediction of glance behavior compared to IOC with a fixed sensor model is due to the fact that more precise individual sensor models for the different display positions were inferred. Furthermore, it turned out that using the additional information contained in the driver's gaze switching behavior as in SRMCE-ISWYS can further reduce prediction error.

Human sensorial capabilities in peripheral vision strongly decrease with increasing eccentricity [36, 61, 226]. In the experiment the characteristics of peripheral vision were experimentally manipulated by requiring the driver to gaze at displays at different positions. The positions were chosen that the forward road scenery had increasing eccentricity with respect to the driver's fovea when his or her gaze was on the display. Considering the statistics of drivers' behavior in the driving experiment Sec. 5.6.3, the individual display position did not result in significant variation in the lane position. Furthermore, the effects on glance behavior were ambiguous: The duration of short glances increased with the amount of gaze aversion, whereas the duration of long glances decreased with the amount of gaze aversion. All considered inference approaches estimate sensor models from their effects on behavior. Consequently, the absence of a significant correlation between sensor noise magnitude in the inferred sensor models and the amount of gaze aversion may be due to the weak effects of the amount of gaze aversion on driver behavior in the conducted driving experiment.

The methods for inference of sensor models improved prediction of glance behavior. However, predicting lane position was not improved compared to maximum causal entropy inverse optimal control using the best guess sensor model. In the driving experiment II presence of the secondary task came with significantly increased deviation from the lane center (Sec. 5.6.3). However, no significant differences between the individual display positions could be established. Hence a single sensor model may already obtain small prediction error in all different display positions.

The sensor noise covariance inferred by ISWYS had values of order 10^8 even for the experimental condition where the driver gazed at the display H right above the steering wheel. These values appear to be very large. Unfortunately, to the best of our knowledge there is currently no published work that could be used for comparison. As shown in Fig. 5.14 these sensor models did result in rather small perceptual uncertainty. Comparably large sensorial noise but small uncertainty in the resulting belief indicates that assuming full knowledge of the kinematic model of the driving situation by the driver might not be valid. In the present work this assumption roots in using the belief MDP (see Sec. 2.1.3) to model driver vehicle control and gaze switching policy. Instead drivers may utilize simple approximate internal models of the dynamics as hypothesized in other contexts of sensori-motor behavior [18, 70].

5.9 Conclusion

This chapter presented a general framework for estimating sensor models underlying rational behavior in partially observable decision processes. Our approach was motivated by the idea of quantifying the characteristics of human sensing from the reactions they elicit from psycho-physics. The framework proposed in this chapter was developed by extending inverse optimal control. A concrete implementation of the framework was derived for the class of POMDPs of the previously introduced normative model of glance behavior in driving in presence of a secondary task. Here, computational tractable exact solution approaches SROpt-ISWYS, SRMCE-ISWYS, SRMCL-ISWYS and STROpt-ISWYS could be obtained. For the purpose of evaluation, a new driving experiment on lane keeping in presence of a secondary task was introduced. In this study, we experimentally manipulated the amount of gaze aversion from the road scenery required for secondary task engagement. Here, effects of this manipulation on the drivers' lane keeping performance and glance behavior could be established. The obtained behavioral data was used to compare to inverse optimal control with a best guess senor model and directly estimated baseline models. The errors in prediction of the recorded behavior showed significant benefits of inference of sensor models using the new methodology. However, problems of over-parametrization were present in the application of our framework and improvement over IOC could not be established with respect to every considered metric.

The methodology developed in this chapter allows to obtain the parameters of characterizing the driver's sensing of vehicle states. These are of crucial importance for computing appropriate glance behavior by means the normative model introduced in previous Cpt.3. Furthermore, inferring the individual parameters of driver's sensor characteristics of common secondary tasks could result in a more sensitive distraction warning system.

The effects of manipulating the drivers' sensor characteristics were rather small in the conducted driving experiment. Hence, in future work subjects should more strongly be motivated towards long glances off the road. This could be done by applying the forced peripheral driving paradigm [230, 229] on a closed test track. Here, we would expect strong effects of the amount of gaze aversion on the driver's behavior, especially on the lane keeping performance. Obtaining the optimal gaze switch policies in the joint task POMDP without restricting the sensor model has problematic computational demand. As a consequence, we were not able to evaluate the implementation of our framework for inference of sensor models STROpt-ISWYS in this variant of the joint task POMDP. In future work therefore more efficient solution techniques should be developed. Additionally, criteria for a-priory detection of over-parametrization as [3] are relevant. This would allow to find an appropriate parametrization and a good experimental design for data collection. Finally, another important research direction is investigating and quantifying the drivers' internal models of the driving situation and to validate the inferred sensor parameters with respect to psycho-physiological findings.

6 Distraction Mitigation by Computation of Appropriate Glance Behavior and Its Evaluation

Current state-of-the-art visual distraction warning systems assess attention based on statistics of the glance behavior alone. Typically, these systems warn if the time passed since the driver had his or her gaze on the road for the last time exceeds a certain predefined threshold. However, drivers show highly adaptive and rational glance behavior. Hence, the state-of-the-art systems are neither optimal in terms of improving driving performance nor in terms of user acceptance.

This chapter develops and evaluates a warning system based on the previously established model of situational appropriate glance behavior. We first present the architecture and the implementation of the warning system including preprocessing of sensor signals, policy computation and warning generation in Sec. 6.3. Sec. 6.4 introduces a comparative evaluation of both the developed warning system and an implementation of a state-of-the-art warning system in a user test. Importantly, both systems used the same presentation of warning and only differed in the triggering mechanism. Finally, we analyze both objective measures of driving performance as subjective ratings of the warning systems by the participants in Sec. 6.5.

This chapter will be published in [201].

6.1 Introduction

Previously, a normative model of Appropriate Glance Behavior (ABG) in driving was developed in Cpt. 3. This model builds on rational gaze-switch policy that takes into account the current driving situation, the secondary task the driver is engaging in and the corresponding characteristics of the driver's sensing of the forward road scenery. In real traffic experienced drivers show adaptive glance behavior. For example, drivers reduce the duration of long glances off the road with increasing driving speed, which is long known in human factors research [207, 69] and was also visible in our driving experiments introduced in Sec. 4.6 and Sec. 5.6. Using the techniques for estimating model parameters established in Cpt. 4 and Cpt. 5, the normative model could accurately predict this adaptive behavior. Hence, we hypothesize that a system that warns drivers if they deviate from the glance strategies underlying the model of appropriate glance behavior is well received. In addition to that, also improvement of driving performance can be expected as an established physical model of the driving task and an empirical model of the driver's sensing and manual control characteristics is taken into account in normative model of glance behavior.

In several previous works distraction warning systems have been proposed. Here, typically warnings are triggered based on the driver's behavior alone. In the simplest approach, the driver is warned if the time passed since his or her gaze was on the forward road scenery exceeds a certain threshold. From theoretical perspective such an approach is definitely inferior to a warning system based on computing situationally appropriate glance behavior as it neglects the influence of the driving situation. However, it is not clear if the theoretical benefits of AGB are also present in practical application: First, driver may find the warnings produced by AGB harder to understand compared to a simpler system which can result in additional distraction [13]. This can be the case for example if some model assumptions made in the normative model of glance behavior are invalid. Here, system behavior can be perceived as inconsistent by the users. Second, drivers have different preferences with respect to the sensitivity of the warning system which can result in significant variation in the individual judgments. Finally, in real driving the performance of the warning system is affected by sensor errors, for example loss of eye-tracking. For these reasons it is necessary to comparatively evaluate both an implementation of a state-of-the-art warning system and the AGB system in a realistic setting.

In this chapter we investigate the benefits of employing AGB to adapt warnings to the driving speed. This done by comparing against a classical Eyes-On-Road detection (EOR) warning system, similar to [108]. In this approach a warning is triggered if the time passed since the driver had his or her gaze

on the road for the last time exceeds a fixed predefined threshold. For both considered warning systems a similar architecture was used that finally triggered the same audio-visual warning interface. Importantly, both warning systems were calibrated on data of a preliminary driving experiment to ensure equal overall sensitivity of EOR and AGB. The evaluation was conducted in form of a user test. In this study subjective ratings of the number of warnings, timing of warnings and usefulness of the warning systems as well as objective measures of driving performance were collect and analyzed.

6.2 Related Work

As reviewed in the introduction a variety of approaches for assessment of attention have been proposed which do not take into account the current driving situation [52, 105, 58, 250, 126, 135]. Of these the methods presented in [108, 52, 105, 126] are solely based on the driver's glance behavior. In the simplest case [108], first a rectangular Region Of Interest (ROI) is defined. The ROI is chosen to cover the forward road scenery. Based on whether the driver's gaze intersects the ROI the gaze is classified into the classes "gaze on road" or "gaze off road". Finally, the proportion of the class eyes-off-road is used for distraction assessments [108]. The other works mainly differ in the shape of the ROI and the statistics computed from the in-ROI classification. Although, more complex approaches allowed to more precisely detect engagement in visually demanding secondary tasks [126], they did not show significant advantages for predicting crash risk compared to the Eyes-Off Duration (EOD) [141]. [62] assessed glance behavior combined with driving speed. Here, a distraction warning index was obtained by dividing EOD by the squared driving speed. EOD divided by the absolute value of the driving speed was used for driver attention assessment in [119]. Furthermore, similar as in the workload manager of [52] secondary tasks were blocked in manually coded locations of high driving demands such as intersections.

A driver distraction warning system was evaluated in a driving simulator in [52]. Here, the warning system significantly altered the drivers' glance behavior but no significant improvement of driving performance could be established. The approach of [105] was used in an extended field study [5], where drivers used the warning system for an average mileage of several thousand kilometers. However, in that study no significant effects but a tendency towards decreased durations of glances off the road was observed. [126] evaluated a driver distraction warning system considering several different secondary tasks in driving simulation. Here, lane keeping performance was significantly improved for one task and significantly worsened for a single other task. Considering all tasks no significant improvement could be established. Furthermore, the distraction warning systems resulted in a higher concentration of the drivers' gaze off the road while no significant effects on glance duration were present. In the evaluation of their distraction warning system [119] could not establish a reduction of the duration of glances off the road but subjective ratings of usefulness were high. In that work no objective measures of driving performance were analyzed.

As an alternative to a standalone distraction warning system, an estimated driver's attention state has been used as an additional feature in the situation analysis of classic ADAS systems. [179] implemented a lane-keeping support system were interventions by means of steering torque are triggered earlier when the driver was classified as being distracted. Furthermore, [187, 236] proposed systems that adapt forward collision warnings with respect to the driver's attention state. A combined approach for detecting driver distraction and corresponding decision-making in lane keeping assistance as well headway keeping assistance was given in [122]. This was implemented as a hidden-mode (corresponding to the driver's distraction state) POMDP. Notably, in all these works the adapted driving assistance systems were neither evaluated wrt. effectiveness nor acceptance by the users. In contrast, [28] showed that early lane keeping assistance for driver engaging into a visually demanding secondary task reduces lane deviation while being similar well accepted as the standard late lane keeping assistance. However, in that work the early warning mode was triggered manually.

Similar as in [105, 5], in this work a standalone distraction warning system is considered. However, our system assesses driver attention from both the glance behavior and the situational context. In contrast to [62, 119], our approach is based on rational policies under an established physical model of the driving task and empirical models of driver's perception and manual control characteristics. As shown in Sec. 3.5.2, this leverages a uniform bound on the loss of vehicle control performance

under the computed glance behavior. In addition to that, we also present a thorough evaluation of the potential benefits of the new warning system in a real driving user test.

6.3 Warning System Design

In this section we present the architectures of the distraction warning system based on a threshold on the eyes-off duration termed Eyes-On-Road (EOR) and by computing appropriate glance behavior (AGB). Here, we restrict the distraction warning system to the driving task of lane keeping similar as in the previous chapters of this thesis.

The warning systems received as an input the estimates of the driver's gaze and head orientation from an eye-tracking system and vehicle states obtained from the CAN-BUS. First, the eye-tracking data was pre-processed returning a classification into "gaze on road" and "gaze off road" as well as an estimate of the duration of the current class. These quantities were then used by the driver attention assessment. Whereas EOR only processed the signals derived from eye-tracking. AGB identified the parameters of a kinematic model of the driving task and generated the POMDP by predicting the future speed profile. Thereafter the maximum causal entropy policy was computed which defined a situation specific threshold on the EOD. Finally, both systems returned warning triggers. The overall architecture including the individual sampling times Δ_t is shown in Fig. 6.1.

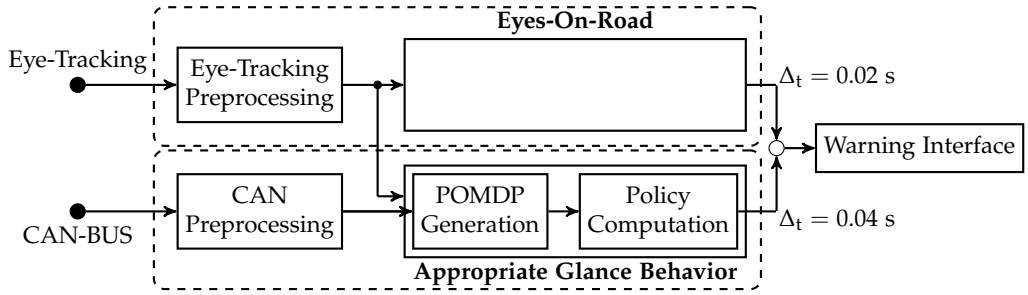

Figure 6.1: Overview of the warning systems. Eye-Tracking preprocessing and EOR run with a sample time of $\Delta_t = 0.02$ s, whereas the CAN preprocessing and AGB run with a sample time of $\Delta_t = 0.04$ s.

In the following we will describe the individual components of the warning system. In this context, the index t' with apostrophe denotes the current time step in application of the warning system. The index t without apostrophe denotes time steps of the POMDP model, ranging from $(t = 0) \equiv t'$ to $(t = T) \equiv t' + T$.

6.3.1 Test Vehicle

The warning system was implemented in a BMW 520d Touring F11 (Bayerische Motoren Werke Aktiengesellschaft, München). For this purpose we integrated the MPC2 system (Robert Bosch GmbH, Stuttgart, Germany) for lane tracking (see Fig. 6.2). Similar as in previous driving experiments a SmartEye Pro (SmartEye AB, Gothenburg, Sweden) infra-red eye-tracking system with active illumination was used to estimate the driver's gaze direction and head orientation. To obtain a larger operation range a four-camera system was employed in the user test. Here, the individual cameras were positioned at the left A-column, in front of instrument cluster, to the right of the central information display and at the right a-column as can been seen in the right picture of Fig. 6.2. Two computers in the back of the vehicle were used for the eye-tracking software (SmartEye AB, Gothenburg, Sweden) and the CANape software (Vector Informatics, Stuttgart, Germany) for measuring CAN-Data, synchronizing data streams as well as interfacing to the implemented algorithms.

Algorithms were first implemented as SIMULINK models (The MathWorks Inc., Natick, United States). Thereafter, we used automatic code generation to obtain highly optimized C code which was compiled for CANape target.

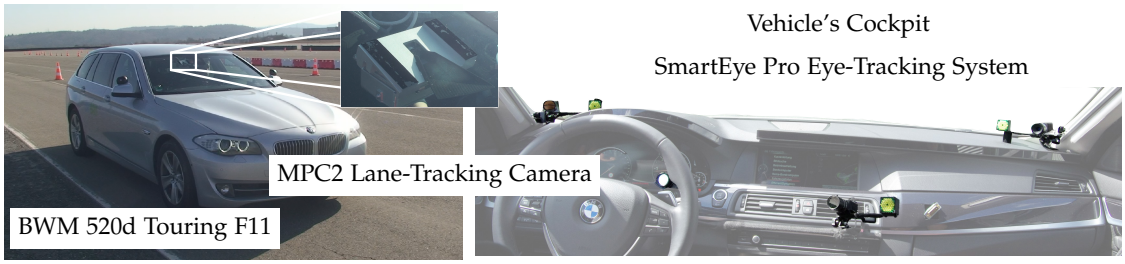

Figure 6.2: Test vehicle in user test. Left picture shows the vehicle and the MPC2 camera, right picture shows the individual cameras of the eye-tracking system in the vehicle's cockpit.

6.3.2 Processing of Eye-Tracking Data

This section describes the processing of the eye-tracking data for distraction warning. Here, first classification of the driver's gaze is considered. Thereafter, the estimation duration of the classified state is explained.

Eye-Tracking Data

The employed eye-tracking system returns an estimate of the driver's eye position. In addition to that it supplies an estimate of the gaze direction by means of its heading which is the angular deviation from a "null" direction along the horizontal axis and its pitch which is the angular deviation from "null" in the vertical axis is supplied. The manufacturer reports that the system uses both the so-called dark and white pupil effect (which are infra-red analogs of the well-known red light effect in photography using a flash) for eye detection and gaze estimation. We refer to the survey article of [75] for more details on the state-of-the-art in eye-tracking. In addition to quantities related to the driver's gaze the system also estimates the driver's head pose using facial landmarks. Most importantly for this work an estimate of the driver's nose-pointing direction is given in form of heading and pitch. In the present work details on head-pose estimation are omitted but can be found for example in the survey article of [237].

Eyes-On-Road Classification

The first processing step with respect to the eye-tracking data was to classify it into the sensor states x_t^z of the joint task POMDP. We followed the approach of [108] and employed a rectangular region of interest for classifying the driver's gaze. The region was placed roughly above the vehicles steering similar as depicted in Fig. 6.3. The region's vertical and horizontal extend was optimized with respect to classification precision using approximately 3 hours of manually annotated gaze data of drivers engaging into a visually demanding secondary task at the central information display. The final region with center $c^{g,x}$, $c^{g,y}$ and width $w^{g,x}$, $w^{g,y}$ had an extend of approximately 0.3 m in the vertical and horizontal axis.

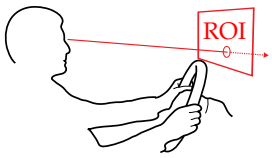

Figure 6.3: Illustration of a typical rectangular region of interest (ROI) as used in this work.

Remote eye-tracking in real driving is a challenging task. Several authors reported imprecise gaze-estimation or even entire loss of tracking when using eye-tracking systems in real driving, for example [5]. As a consequence, classifying gaze into sensor states x_t^z using an ROI approach is typically subject to significant noisy. Fig. 6.7 (2) shows an example of the eye-tracking data and the noisy classification. In that case decreased quality of eye-tracking is probably attributed to intensive sun-light irradiation as

can be seen in Fig. 6.7 (1). To increase robustness of classification, we optimized a second rectangular ROI with parameters $c^{\mathrm{h},x}$, $c^{\mathrm{h},y}$, $w^{\mathrm{h},x}$, $w^{\mathrm{h},y}$ for the nose-pointing direction. Thereafter, we fused the intersection point of the driver's gaze $p_{t'}^{\mathrm{g},x}, p_{t'}^{\mathrm{g},y}$ with its rectangular ROI as well as the intersection point of the nose-pointing direction $p_{t'}^{\mathrm{h},x}, p_{t'}^{\mathrm{h},y}$. This was done using the logistic regression model

$$p(x_{t'}^{\mathrm{z}} = 1 | p_{t'}^{\mathrm{g},x}, p_{t'}^{\mathrm{g},y}, p_{t'}^{\mathrm{h},x}, p_{t'}^{\mathrm{h},y}) = \exp\big(\lambda_1^{\mathrm{class}} + \lambda_2^{\mathrm{class}} \max(|p_{t'}^{\mathrm{g},x} - c^{\mathrm{g},x}| - w^{\mathrm{g},x}, |p_{t'}^{\mathrm{g},y} - c^{\mathrm{g},y}| - w^{\mathrm{g},y}) \\ \lambda_3^{\mathrm{class}} \max(|p_{t'}^{\mathrm{h},x} - c^{\mathrm{h},x}| - w^{\mathrm{h},x}, |p_{t'}^{\mathrm{h},y} - c^{\mathrm{h},y}| - w^{\mathrm{h},y})\big). \quad (6.1)$$

Anecdotical evidence of the classification using the logistic regression model is presented in Fig. 6.7 (2).

Estimation of Eyes-Off Duration

Given the classification of the driver's gaze next the time passed since the driver's gaze was on the road for the last time, i.e. the eyes-off duration (EOD) $d_{t'}$, was estimated.

This seems simple, but note that a single wrong classification of the gaze has a massive impact on the estimated EOD. For example, can a single wrong classification into gaze on road falsely strongly reduce the estimated EOD. [6] proposed approaches for preprocessing of eye-tracking data based on outlier detection and removal as well as total-variation based smoothing and interpolation. However, these were developed for offline processing and turned out to be ineffective in an online setting. Instead we employed a chain-structured conditional random field to estimate EOD: We first defined the variable $\delta_{t'} \in \mathbb{Z}$ to be the time steps passed since the last gaze switch using its sign to indicate whether the gaze is on the road or off the road

$$\delta_{t'} = \begin{cases} +\min_{\{k:\, k \geq 0, u_{t'-k}^{\mathrm{z}} = 1\}}(k) \text{ if } x_{t'}^{\mathrm{z}} = 1 \\ -\min_{\{k:\, k \geq 0, u_{t'-k}^{\mathrm{z}} = 1\}}(k) \text{ if } x_{t'}^{\mathrm{z}} = 0 \end{cases}. \quad (6.2)$$

The employed random field model specifies the probability of an element $\delta_{t'+1}$ given $\delta_{t'}$ and $p_{t'+1}^{\mathrm{g},x}, p_{t'+1}^{\mathrm{g},y}, p_{t'+1}^{\mathrm{h},x}, p_{t'+1}^{\mathrm{h},y}$ as

$$p(\delta_{t'+1} | \delta_{t'}, p_{t'+1}^{\mathrm{g},x}, p_{t'+1}^{\mathrm{g},y}, p_{t'+1}^{\mathrm{h},x}, p_{t'+1}^{\mathrm{h},y}) = \exp\big(\psi(\delta_{t'}, p_{t'+1}^{\mathrm{g},x}, p_{t'+1}^{\mathrm{g},y}, p_{t'+1}^{\mathrm{h},x}, p_{t'+1}^{\mathrm{h},y}) + \psi(\delta_{t'+1}, \delta_t)\big). \quad (6.3)$$

In this context the factor $\psi(\delta_{t'}, p_{t'}^{\mathrm{g},x}, p_{t'}^{\mathrm{g},y}, p_{t'}^{\mathrm{h},x}, p_{t'}^{\mathrm{h},y})$ was defined by means of the classification model (6.1)

$$\exp\big(\psi(\delta_{t'}, p_{t'}^{\mathrm{g},x}, p_{t'}^{\mathrm{g},y}, p_{t'}^{\mathrm{h},x}, p_{t'}^{\mathrm{h},y})\big) = \begin{cases} p(1 | p_{t'}^{\mathrm{g},x}, p_{t'}^{\mathrm{g},y}, p_{t'}^{\mathrm{h},x}, p_{t'}^{\mathrm{h},y}) & \text{if } \delta_{t'} \geq 0 \\ 1 - p(1 | p_{t'}^{\mathrm{g},x}, p_{t'}^{\mathrm{g},y}, p_{t'}^{\mathrm{h},x}, p_{t'}^{\mathrm{h},y}) & \text{else} \end{cases} \quad (6.4)$$

and the factor $\psi(\delta_{t'+1}, \delta_t)$ was given by

$$\exp\big(\psi(\delta_{t'+1}, \delta_t)\big) \\ = \begin{cases} \exp\big(\lambda_1^{\mathrm{dyn}} + \boldsymbol{\lambda}_2^{\mathrm{dyn}\,\top}[|\delta_t|; |\delta_t|^2; |\delta_t|^3]\big) & \text{if } \delta_{t'+1} = -1, |\delta_t| \leq \lambda_3^{\mathrm{dyn}}, \delta_{t'} \geq 0 \\ 1 - \exp\big(\lambda_1^{\mathrm{dyn}} + \boldsymbol{\lambda}_2^{\mathrm{dyn}\,\top}[|\delta_t|; |\delta_t|^2; |\delta_t|^3]\big) & \text{else if } \delta_{t'+1} = \delta_t + 1, |\delta_t| \leq \lambda_3^{\mathrm{dyn}}, \delta_{t'} \geq 0 \\ \exp\big(\lambda_4^{\mathrm{dyn}}\big) & \text{else if } \delta_{t'+1} = -1, |\delta_t| > \lambda_3^{\mathrm{dyn}}, \delta_{t'} \geq 0 \\ 1 - \exp\big(\lambda_4^{\mathrm{dyn}}\big) & \text{else if } \delta_{t'+1} = \delta_t, |\delta_t| > \lambda_3^{\mathrm{dyn}}, \delta_{t'} \geq 0 \\ \exp\big(\lambda_5^{\mathrm{dyn}} + \boldsymbol{\lambda}_6^{\mathrm{dyn}\,\top}[|\delta_t|; |\delta_t|^2; |\delta_t|^3]\big) & \text{else if } \delta_{t'+1} = +1, |\delta_t| \leq \lambda_7^{\mathrm{dyn}}, \delta_{t'} < 0 \\ 1 - \exp\big(\lambda_5^{\mathrm{dyn}} + \boldsymbol{\lambda}_6^{\mathrm{dyn}\,\top}[|\delta_t|; |\delta_t|^2; |\delta_t|^3]\big) & \text{else if } \delta_{t'+1} = \delta_t - 1, |\delta_t| \leq \lambda_7^{\mathrm{dyn}}, \delta_{t'} < 0 \\ \exp\big(\lambda_8^{\mathrm{dyn}}\big) & \text{else if } \delta_{t'+1} = +1, |\delta_t| > \lambda_7^{\mathrm{dyn}}, \delta_{t'} < 0 \\ 1 - \exp\big(\lambda_8^{\mathrm{dyn}}\big) & \text{else if } \delta_{t'+1} = \delta_t - 1, |\delta_t| > \lambda_7^{\mathrm{dyn}}, \delta_{t'} \geq 0 \\ 0 & \text{else} \end{cases}. \quad (6.5)$$

The rationale behind the chosen transition factor was to use a complex model for short glances as to distinguish between natural glance behavior and failure of eye-tracking. For long glances the model was made less informative. The model parameters $\boldsymbol{\lambda}^{\text{dyn}}$ were inferred from the same set of manually annotated data which was used to optimize the gaze classification.

For inference of the distribution of $p(\delta_{t'}|\boldsymbol{p}^{\text{g},x}_{0:t}, \boldsymbol{p}^{\text{g},y}_{0:t}, \boldsymbol{p}^{\text{h},x}_{0:t}, \boldsymbol{p}^{\text{h},y}_{0:t})$ we employed a particle filter [14]. Given a particle $\delta_{t'-1}$ and $p^{\text{g},x}_t, p^{\text{g},y}_t, p^{\text{h},x}_t, p^{\text{h},y}_t$ the a-posterior distribution of $\delta_{t'}$ is given in analytic form

$$
\begin{aligned}
&\text{if } \delta_{t'-1} > 0, \delta_{t'} = \delta_{t'-1} + 1 \text{ or } \delta_{t'} = -1\text{:} \\
&p(\delta_t|p^{\text{g},x}_t, p^{\text{g},y}_t, p^{\text{h},x}_t, p^{\text{h},y}_t, \delta_{t'-1}) = \exp\left(\psi(\delta_{t'}, p^{\text{g},x}_{t'}, p^{\text{g},y}_{t'}, p^{\text{h},x}_{t'}, p^{\text{h},y}_{t'})\right) \exp\left(\psi(\delta_t, \delta_{t'-1})\right) \\
&\qquad \big[\exp\left(\psi(-1, p^{\text{g},x}_{t'}, p^{\text{g},y}_{t'}, p^{\text{h},x}_{t'}, p^{\text{h},y}_{t'})\right) \exp\left(\psi(-1, \delta_{t'-1})\right) \\
&\qquad\quad + \exp\left(\psi(\delta_{t'-1}+1, p^{\text{g},x}_{t'}, p^{\text{g},y}_{t'}, p^{\text{h},x}_{t'}, p^{\text{h},y}_{t'})\right) \exp\left(\psi(\delta_{t'-1}+1, \delta_{t'-1})\right)\big]^{-1}
\end{aligned} \tag{6.6}
$$

$$
\begin{aligned}
&\text{if } \delta_{t'-1} < 0, \delta_{t'} = \delta_{t'-1} - 1 \text{ or } \delta_{t'} = +1\text{:} \\
&p(\delta_t|p^{\text{g},x}_t, p^{\text{g},y}_t, p^{\text{h},x}_t, p^{\text{h},y}_t, \delta_{t'-1}) = \exp\left(\psi(\delta_{t'}, p^{\text{g},x}_{t'}, p^{\text{g},y}_{t'}, p^{\text{h},x}_{t'}, p^{\text{h},y}_{t'})\right) \exp\left(\psi(\delta_t, \delta_{t'-1})\right) \\
&\qquad \big[\exp\left(\psi(+1, p^{\text{g},x}_{t'}, p^{\text{g},y}_{t'}, p^{\text{h},x}_{t'}, p^{\text{h},y}_{t'})\right) \exp\left(\psi(+1, \delta_{t'-1})\right) \\
&\qquad\quad + \exp\left(\psi(\delta_{t'-1}-1, p^{\text{g},x}_{t'}, p^{\text{g},y}_{t'}, p^{\text{h},x}_{t'}, p^{\text{h},y}_{t'})\right) \exp\left(\psi(\delta_{t'-1}-1, \delta_{t'-1})\right)\big]^{-1}
\end{aligned} \tag{6.7}
$$

Hence, empirically we observed that already 20 particles $\{\boldsymbol{\delta}^{i=1:20}_{t'}\}$ suffice to maintain a good approximation of the distribution of $\delta_{t'}$. Fig. 6.7 (3) shows an example of the 20 particle estimates of $\delta_{t'}$.

6.3.3 Processing of CAN-BUS Data

Next, it is explained how the relevant quantities measured in the vehicle's CAN-BUS were filtered and processed. These steps are required to generate the POMDP model of the driving situation.

CAN-BUS Data

To compute the policy for definition of appropriate glance behavior, we first generated the joint task POMDP model corresponding to the current driving situation. The kinematic model of the primary task of driving is given by the model

$$
\mathbf{x}^{\text{p}}_t = \mathbf{A}(v_t)\mathbf{x}^{\text{p}}_t + \mathbf{B}(v_t)u^{\text{p}}_t + \mathbf{a}(v_t, \kappa_t) + \boldsymbol{\epsilon}^{\text{p}}_t, \quad \boldsymbol{\epsilon}^{\text{p}}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\epsilon^{\text{p}}})
$$

and an initial state $\mathbf{x}^{\text{p}}_0 = [y_{0\equiv t'}\ \dot{y}_{0\equiv t'}\ \phi_{0\equiv t'}\ \alpha_{0\equiv t'}]^\top$. Furthermore, the matrices $\mathbf{A}(v_t), \mathbf{B}(v_t), \mathbf{a}(v_t, \kappa_t)$ depend on the vehicles steering wheel to yaw-rate constant $\bar{c}_1$ (see Sec. 3.3.1) which needs to be identified.

We obtained the data related to the lane-tracking from a CAN-BUS connected to the MPC2 camera, while the vehicle's driving speed, longitudinal acceleration, yaw-rate, steering angle and steering wheel velocity were obtained from the vehicle's electronic stability program computer via a central CAN-BUS.

Filtering and Prediction

A first processing step estimated the constant $\bar{c}_1$ by online gradient descent minimization of the error in predicting the yaw-rate

$$
\min_{\bar{c}_1} \|\bar{c}_1 v_{t'} \alpha_{t'} - \dot{\bar{\phi}}_{t'}\|^2. \tag{6.8}
$$

Furthermore, a Kalman-filter using the same parameterization as employed in the preprocessing of the data of driving experiment I in Sec. 4.6.2 was used to filter the data from the lane-tracking. Finally, a second Kalman filter was used to filter the vehicle's driving speed and longitudinal acceleration.

Generating the POMDP model of the primary task does not only require current driving speed and road curvature but also its future values along the horizon T. The lane-tracking camera estimates the road curvature from the visible lane boundaries ahead of the vehicle, hence the current curvature $\kappa_{t'}$ was used as an estimate of the entire horizon $\boldsymbol{\kappa}_{0:T} = \kappa_{t'}$. The future speed profile was predicted based on the filtered longitudinal acceleration. Specifically, we used the prediction model

$$v_t = \max(v_{t'} + t/25\,\mathrm{s}\,\dot{v}_{t'}, 0) \tag{6.9}$$

where $1/25\,\mathrm{s}$ is the inter-sample time and where it was assumed that once the vehicle stands it remains at zero driving speed.

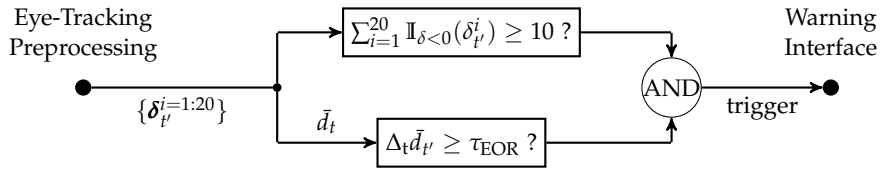

Figure 6.4: Illustration of the implementation of the eyes-on-road (EOR) algorithm. The particle estimates of the eyes-off duration $\{\boldsymbol{\delta}_t^{i=1:20}\}$ are first used to check whether the driver's gaze is off the road. If additionally, the average EOD in seconds $\Delta_t \bar{d}_t$ exceeded a threshold τ_{EOR} a warning was triggered.

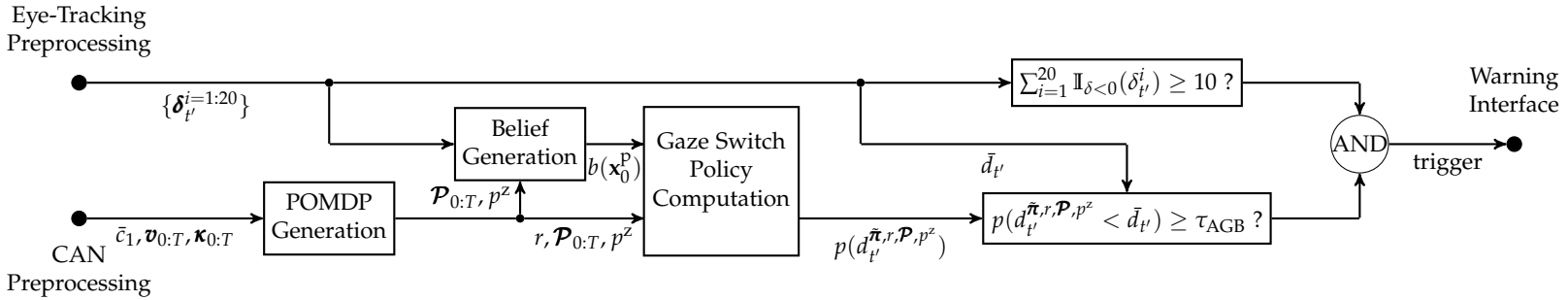

Figure 6.5: Illustration of the implementation of appropriate glance behavior (AGB). The quantities returned from the CAN data preprocessing were used to generate the joint task POMDP model $r, \mathcal{P}_{0:T}, p^z$. Thereafter the driver's belief of the primary task states $b(\mathbf{x}_t^{\text{p}})$ was estimated based on the POMDP model and the particle estimates of the EOD. Given the driver's belief and the POMDP model the maximum causal entropy policy for gaze switching $\tilde{\boldsymbol{\pi}}_{0:T}$ was computed. If the driver's gaze was estimated to be off the road and if the probability of a smaller EOD under the policy $p(d_{t'}^{\tilde{\boldsymbol{\pi}},r,\mathcal{P},p^z} < \bar{d}_{t'})$ was higher than a threshold τ_{AGB} a warning was triggered.

6.3.4 Eyes-On-Road Implementation

We implemented the eyes-on-road algorithm in the following way: To trigger a warning, it was first required that at least half of the particles $\{\boldsymbol{\delta}_{t'}^{i=1:20}\}$ indicated that the driver's gaze was off the road

$$\sum_{i=1}^{20} \mathbb{I}_{\delta<0}(\delta_{t'}^{i}) \geq 10. \tag{6.10}$$

Second, the eyes-off duration $d_{t'}$ was estimated by the average eyes-off duration $\bar{d}_{t'}$ over the particles indicating gaze off the road,

$$\bar{d}_{t'} = \sum_{i=1}^{20} \mathbb{I}_{\delta<0}(\delta_{t'}^{i})\, |\delta_{t'}^{i}|. \tag{6.11}$$

If additionally the estimated EOD scaled by the inter-sample time $\Delta_t \bar{d}_{t'}$ exceeded a predefined threshold τ_{EOR} of seconds, a warning was triggered. The estimated EOD is depicted as a blue line in Fig. 6.7 (5). Furthermore the entire implementation is illustrated in Fig. 6.4.

6.3.5 Appropriate Glance Behavior Implementation

The implementation of warning based on computing appropriate glance behavior was done in the following way: First, the joint task POMDP model was generated. Thereafter the covariance of the Gaussian belief of the primary task states of the driver was estimated. This was based on the particle estimates of the off-road duration as well as the POMDP model. Given both the current belief of the driver and the POMDP model the maximum causal entropy policy was computed. A warning was finally triggered if both the particles indicated the driver's gaze being off the road and if under the maximum causal entropy policy a return of gaze to the road was sufficient likely. The implementation is outlined in Fig. 6.5.

POMDP Generation

Given the estimate of the steering angle to yaw-rate constant $\bar{c}_1$ and the predicted speed profile $\boldsymbol{v}_{0:T}$ as well as the curvature profile $\boldsymbol{\kappa}_{0:T}$ the joint task POMDP model was generated. We combined the kinematic model of the driving task with the simple secondary task model (Sec. 3.3.3) and the sensor model restriction (Sec. 3.5.2). That is the following POMDP was generated:

$$r(x_t, u_t) = \theta_1 (y_t)^2 + \theta_2 (\dot{y}_t)^2 + \theta_3 (\alpha_t)^2 + \theta_4 (u_t^{\text{p}})^2 + \theta_5 u_t^{\text{z}} + \theta_6 (1 - x_t^{\text{z}}) \tag{6.12}$$

$$\mathcal{P}_t(x_{t+1}|x_t, u_t) \text{ def. by } \begin{cases} \mathbf{x}_{t+1}^{\text{p}} & = \mathbf{A}(v_t)\mathbf{x}_t^{\text{p}} + \mathbf{B}(v_t)u_t^{\text{p}} + \mathbf{a}(v_t, \kappa_t) + \boldsymbol{\epsilon}_t^{\text{p}}, \\ & \quad \boldsymbol{\epsilon}_t^{\text{p}} \sim \mathcal{N}(\mathbf{0}, \text{diag}((\sigma_y^{\text{x}})^2, (\sigma_{\dot{y}}^{\text{x}})^2, (\sigma_{\phi}^{\text{x}})^2, 0)) \\ x_{t+1}^{\text{z}} & = x_t^{\text{z}} \oplus u_t^{\text{z}} \end{cases} \tag{6.13}$$

$$p^{\text{z}}(z_t|x_t) \text{ def. by } \begin{cases} \mathbf{z}_t & = \mathbf{H}\mathbf{x}_t^{\text{p}} + \boldsymbol{\epsilon}_t^{z}(x_t^{\text{z}}), \\ & \quad \mathbf{H} = \text{diag}(1, 0, 1, 1), \quad \boldsymbol{\epsilon}_t^{z}(x_t^{\text{z}}) \sim \mathcal{N}(\mathbf{0}, \text{diag}((\sigma_y^{\text{z}})^2(x_t^{\text{z}}), 0, (\sigma_{\phi}^{\text{z}})^2(x_t^{\text{z}}), 0)) \end{cases}. \tag{6.14}$$

In this context we employed a planning horizon T of 50 steps which corresponds to look-a-head time of 2 s.

Belief Generation

An important part of the joint task POMDP is the driver's initial belief of the primary task states $b(\mathbf{x}_0^{\text{p}}) = \mathcal{N}(\boldsymbol{\mu}_0^{\text{p}}, \boldsymbol{\Sigma}_0^{\text{p}})$. Specifically, the covariance $\boldsymbol{\Sigma}_0^{\text{p}}$ of the initial belief is required to compute the MCE policy for gaze switching.

In the applied variant of the joint task POMDP $\mathcal{P}, p^{\text{z}}$ this covariance matrix was fully determined by the EOD $d_{0\equiv t'}$, i.e. $\boldsymbol{\Sigma}_0^{\text{p}}(d_{t'})$. The EOD was estimated by means of the particles $\{\boldsymbol{\delta}_{t'}^{i=1:20}\}$ in our warning

system architecture. From these particles we obtained an estimate of the covariance of the belief of the primary task states in the following way: We maintained the individual covariances $\mathbf{\Sigma}_t^{\mathrm{p}}(\delta_{t'}^i)$ associated with every particle $\delta_{t'}^i$. An estimate of the covariance of the driver's belief $\overline{\mathbf{\Sigma}^{\mathrm{p}}}_0(d_{t'})$ was obtained by taking the expectation over all covariances $\mathbf{\Sigma}_t^{\mathrm{p}}(\delta_{t'}^i)$ associated with particles $\delta_{t'}^i$ that indicated that the driver's gaze was off the road:

$$\overline{\mathbf{\Sigma}^{\mathrm{p}}}_0(d_{t'}) = \sum_{i=1}^{20} \mathbb{I}_{\delta<0}(\delta_{t'}^i)\, \mathbf{\Sigma}_t^{\mathrm{p}}(\delta_{t'}^i). \tag{6.15}$$

This specific estimator was used in the present work as it showed improved robustness compared to the covariance $\mathbf{\Sigma}_t^{\mathrm{p}}(\bar{d}_{t'})$ resulting from the estimated eyes-off duration.

Gaze Switch Policy Computation

The generated POMDPs (6.12)-(6.14) are instances of the joint task POMDP under sensor model restriction. As such the maximum causal entropy policy $\tilde{\boldsymbol{\pi}}_{0:T}$ of the POMDPs can be obtained by Algo. 7. For the purpose of computing appropriate glance behavior, we further optimized the procedure by hard coding the simple secondary task model.

The probability $p(d_{t'}^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z} < \bar{d}_{t'})$ of shorter EOD under the MCE policy $\tilde{\boldsymbol{\pi}}$ than the current estimated EOD was obtained in the following way. If the driver's gaze was on the road, we set $p(d_{t'}^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z} < \bar{d}_{t'}) = 0$. Else if the driver's gaze was off the road the probability of a return of gaze back to the road $p(d_{t'}^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z} < \bar{d}_{t'})$ was computed by

$$p(d_{t'}^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z} < \bar{d}_{t'}) = 1 - \prod_{t''=t'-\bar{d}_{t'}+1}^{t'} \tilde{\pi}_{t''}(u_{t''}^z = 0 | t'' - \bar{d}_{t'}). \tag{6.16}$$

That is possible because the return of gaze back to the road is the complementary event of the event that always sensor control $u_{t''}^z = 1$ "keep current sensor state" has been chosen under the maximum causal entropy gaze switch policy.

6.3.6 Warning Interface

In the previous sections we described the algorithms for preprocessing of sensor data and the approaches for distraction assessment. If either EOR or AGB triggered a warning the warning interface was activated. Here, a warning icon was presented at the same display used for the visually demanding secondary tasks which is depicted in Fig. 6.6.

Figure 6.6: Visually demanding secondary task and the visual component of the distraction warning. The secondary task required reading random numbers 1 and 2 shown in green in the right lower quarter of the display. The visual component of the warning consisted of a red warning icon shown in the right upper corner of the display.

The warning icon was combined with a short high frequency warning tone which was output via the vehicles built-in loudspeakers. Warnings were output as long as the interface received a trigger from the driver attention assessment algorithms. That is, if the driver did not return his or her gaze to the

road further warnings were produced. We wish to note that other works [104, 5] used more complex warning concepts. Specifically, warnings were presented longer and escalated if the driver did not return gaze to road. We decided on the single short warning tone because it facilitates recognizing the warning timing by the user. This was important for the user test (described later in Sec. 6.4) because we were interested in the subjective assessment of the warning timing by the participants.

6.3.7 Example of Warning System

Combining the preprocessing, the driver attention assessment algorithms and the warning interface results in the final warning systems. In Fig. 6.7 we present a short snippet from the user test which illustrates the different components of the system and their important outputs:

Fig. 6.7 shows a snippet of 2.5 s length in which a participant averted her gaze from the road, received a distraction warning and returned gaze back to the road. The first row **(1)**, shows the driver's glance behavior. The time steps corresponding to the presented pictures are indicated as non-filled rectangles in all other plots. **(2)** plots the gaze angles measured by the eye-tracking system. Here, dashed lines are used to indicate the gaze heading and gaze pitch. The blue dashed line indicates gaze intersection with the optimized rectangular ROI as described in Sec. 6.3.2 and e.g. applied in [108]. The red dashed line denotes the probability of gaze on the road using the logistic regression model as employed in the present work (see (6.1)). The particle estimates of the (signed) time passed since the last gaze switch as described in Sec. 6.3.2 are shown in **(3)**. **(4)** depicts the vehicle's interior featuring the warning icon. Here frames were captured at approximately the same time as the frames of the driver's face. The warning icon appears on the third picture on the display mounted to the right of the steering wheel. Finally, the warning indices used in EOR and AGB are shown in **(5)**. The blue line indicates the EOD used to trigger warnings in EOR. The Red lines indicate the probability of return of gaze to the road scenery in AGB. Here the solid red line shows the actual warning index of the present snippet in the driving experiment whereas the dashed red lines depict warning index of AGB simulated for different driving speeds. The warning trigger is depicted as a black line. In the present snipped the warning trigger was controlled by AGB.

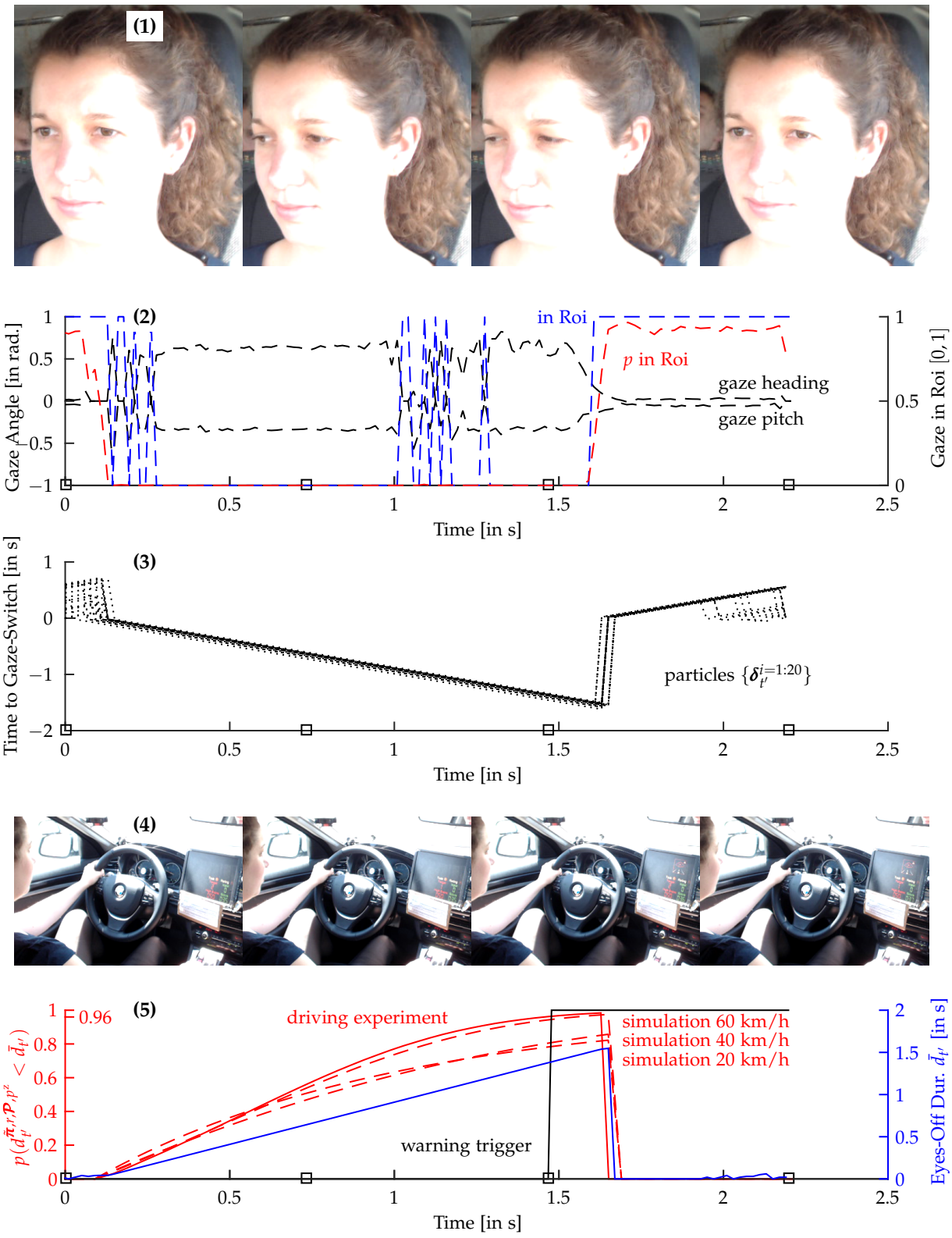

Figure 6.7: Example of the warning system (written consent of the participant was obtained).

6.4 User Study

In the last section we introduced two distraction warning systems that we integrated into a test vehicle. Both warning systems shared a common architecture comprising of sensor data preprocessing, driver attention assessment and a warning interface. In this context, both the preprocessing algorithms as well as the warning generation were exactly the same for both systems whereas the main differences were the usage of the EOR and AGB for assessment of driver attention. While EOR produces warnings if the driver's eyes-off duration exceeds a fixed threshold, warnings produced by AGB are dependent on the driving situation as modeled in the joint task POMDP. Furthermore, the driver attention assessment based on EOR is comparably simple, while AGB requires to compute rational policies for non-trivial POMDP models.

Given these differences, we conducted a user test to compare and evaluate both warning systems. Here, the following research questions were addressed:

R1 Can computing appropriate glance behavior be applied as a real-time distraction warning system?

R2 How are the warning systems accepted by users exposed to the warnings?

R3 What are the effects of the distraction warning systems on driving performances?

R4 What are the effects of the distraction warning systems on drivers' glance behavior?

In the present user test we exemplary investigated these research questions with respect to adaption of warnings to driving speed in AGB.

We wish to note, that such an adaption is in principle also possible using heuristics as e.g. employed in [62]. However, in our model of normative glance behavior adaptivity naturally results from the kinematic model of the primary driving task. Intuitively this is because of the following reason: Compared to low driving speed, at higher driving speed the same orientation in lane, e.g. an offset $\Delta\phi$ from zero due to a steering error, results in increased deviation from the lane center if not noticed by a driver who averted his or her gaze. Correspondingly, at higher driving speeds a decreased duration of glances off the road is permitted in AGB. In addition to that, in AGB also adaptation to other aspects of the joint task of secondary task engagement while driving is possible. For example, can the warning system be adapted to the driver's specific sensor characteristics in an individual driving task as discussed in Cpt. 5. Due to the issues involved in estimation of sensor models we omitted an evaluation of this property of the model of appropriate glance behavior.

In the following we will now introduce the driving experiment that was conducted to investigate the aforementioned research questions.

6.4.1 Participants

For the user test we recruited 18 (4 female, 14 male) participants from the Robert Bosch GmbH at Renningen. The age of the participants ranged from 25 to 43 years (mean $\mu = 31.5$, standard deviation $\sigma = 5.1$). As the eye-tracking system did not reliably work with glasses only drivers participated that either wore contact lenses or did not need visual aids at all. All participants possessed a driving license valid in Germany. Furthermore, kilometers driven by the participants in the last year showed significant variation ranging from 1000 to 36000 km. In addition to that, the drivers had a similar attitude towards engagement in secondary tasks as assessed by a questionnaire regarding the frequency of engagement in several common secondary tasks.

Figure 6.8: Participants squinting due to strong sun irradiation (written consent was obtained).

Of these 18 participants, 3 participants' data could not be analyzed as strong sun irradiation and participants' squinting led to insufficient eye-tracking quality. Fig. 6.8 depicts two examples of squinting participants. In addition to that, the data of one participant was corrupted due to a saving error.

6.4.2 Test Track

The user test was conducted on a closed test track of the Robert Bosch GmbH at Renningen. The track is depicted in Fig. 6.9. Here, recordings were taken on a marked lane in the northern part of the test track which had a length of approximately 700 meters.

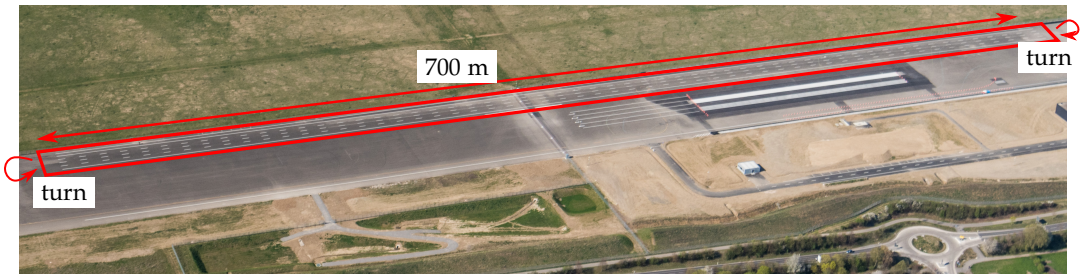

Figure 6.9: Aerial photography of the test track used for the user test. Recordings were taken driving in one of the northern marked lanes of a length of approximately 700 m.

Lane keeping on a test track comes with significantly lower risk compared to lane keeping on a motorway as in previous experiments. Consequently, lane departure on a test track may be more frequent. To motivate the participants to stay in lane we deployed orange cones next to the lane and instructed the driver not to run over the cones. Here, cones on the left and on the right were offset to ensure that they do not provide additional guidance cues in peripheral vision. The installation is shown in Fig. 6.10.

Figure 6.10: Illustration of the installation of cones on the test track.

6.4.3 Protocol

Participants first filled out a questionnaire regarding demographic aspects, driving style and attitude towards secondary tasks while driving. Thereafter, the purpose of the user test was explained. In this context, participants only knew that two different approaches for attention assessment would be evaluated but no details regarding their functionality were revealed.

Before the recordings of user test were started, the participants practiced the secondary task both in stand-still and at 40 km/h. The entire user test consisted of three blocks: First participants drove and engaged into a secondary task without a warning system active, thereafter they were treated with both warning systems block-wise and in randomized order (warning system condition). Two recordings of every warning system condition were taken at the driving speeds of 20 km/h, 40 km/h and 60 km/h (speed condition). In each trial, participants started from stand-still and first accelerated to the desired driving speed. Once the driving speed was stably reached both the secondary task and, dependent on the experimental condition, the warning system was activated by the instructor. The secondary task was automatically deactivated after 30 s. At the end of the lane participants turned the vehicle and rated the warning system if any was active.

In the user test we used the same secondary task as in the previous driving experiments (Sec. 4.6 and Sec. 5.6). We refer to Sec. 4.6 and Sec. 3.3.3 for a detailed description. To motivate the participants to engage into the secondary task, they were shown their score defined as the number of correct button presses minus the number of incorrect button presses. Furthermore, a fabricated high-score of 70 was displayed.

The drivers were asked to rate the warning system experienced in the last period of secondary task engagement with respect to three categories:

1. **Number of the received warnings** (presented in German as "Wie angemessen war die Anzahl der Warnungen?") on a Likert-type-scale of 5 items *few* ("zu wenige"), *a little few* ("etwas zu wenige"), *ideal* ("genau angemessen"), *a little many* ("etwas zu viele"), *many* ("zu viele")

2. **Timing of the received warnings** (presented in German as "Wie rechtzeitig waren die Warnungen?") on a Likert-type-scale of 5 items *soon* ("zu früh"), *a little soon* ("etwas zu früh"), *ideal* ("genau rechtzeitig"), *late* ("etwas zu spät"), *very late* ("zu spät")

3. **Usefulness of received warnings** (presented in German as "Wie hilfreich war das Warnsystem zum sicheren Fahren?") on a Likert-type-scale of 5 items *useless* ("nicht hilfreich"), *quite useless* ("wenig hilfreich"), *sort of useful* ("etwas hilfreich"), *useful* ("hilfreich"), *very useful* ("sehr hilfreich")

The participants made their ratings by additional numeric buttons $1-5$ on the same number pad that was also used for the secondary task. Fig. 6.11 depicts the number pad used for rating and the legend presented in close vicinity.

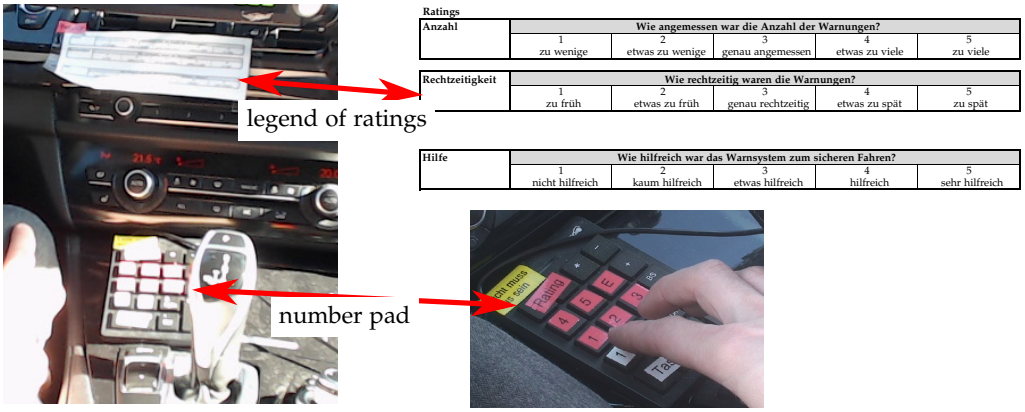

Figure 6.11: Illustration of the ratings conducted by the participants of the user test. Ratings were made using the red numeric buttons 1-5 on a number pad. A legend explained how numbers related to the ratings.

Participants were instructed to specifically rate the last period of secondary task engagement under the experienced "artificial" scenario on the test track.

6.4.4 Calibration of Warning Systems

In the user test we sought to evaluate the different algorithms EOR and AGB for distraction assessment. We wish to remind, that EOR uses a fixed threshold on the eyes-off duration while AGB results in a threshold on the eyes-off duration adapted to the driving speed. Consequently, we were specifically interested in investigating the effects of this adaptation on drivers' behavior and how it is received by the user. Therefore, we needed to ensure that both warning systems show a similar total sensitivity. That is, we did not want to compare systems of which one generally warns at a shorter eyes-off duration than the other one.

For these reasons a preliminary driving experiment was conducted to obtain calibration data. Here, we recruited 16 different participants with similar age, driving style and attitude towards secondary task as those of the user test. The behavioral data of driving without a warning system was first used to infer reward parameters and sensor model parameters of the POMDP model underlying AGB by means of SRMCE-ISWYS Algo. 18 (see Cpt. 5). Thereafter, we simulated both warning systems for a variety of thresholds τ_{EOR}, τ_{AGB}. The resulting average times between warnings are shown in Fig. 6.12.

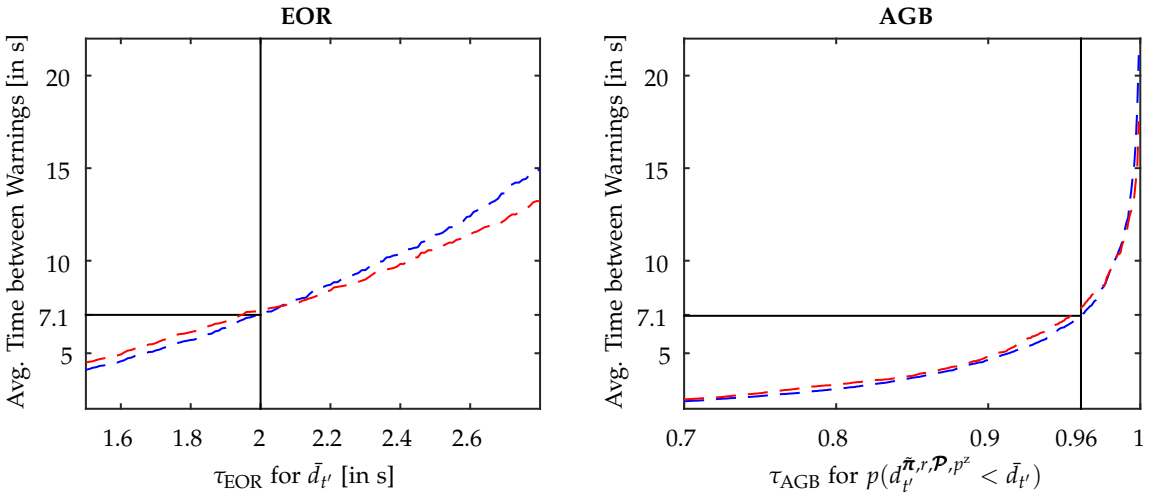

Figure 6.12: Comparison of the average times between warnings for data used for calibration and for data from the user test. Data from the calibration is depicted by dashed red lines, data from driving without warning system in the user test is indicated by dashed blue lines. Left plot shows the average time between warnings for EOR for different thresholds τ_{EOR} for the eyes-off duration $\bar{d}_t$. Right plot shows the average time between warnings for AGB for the different threshold τ_{AGB} on the probability of return of gaze to the road $p(d_t^{\tilde{\boldsymbol{\pi}},r,\boldsymbol{\mathcal{P}},p^z} < \bar{d}_t)$ under the maximum causal entropy policy. The threshold for EOR τ_{EOR} was set to 2 s and the threshold of AGB τ_{AGB} was set to 0.96 which resulted in the same avg. time between warnings of 7.1 s.

We decided on a threshold on the eye-off duration of 2 s. This was because several algorithms for attention assessment proposed in the literature build on this threshold value [105, 236]. In addition to that, this threshold was approximately the largest value were there occurred at least one warning in 75% of all trials. The threshold for the probability of a return of gaze back to the road scenery was set 0.96 as to result in the same average time between warnings of 7.1 s as EOR. The effective warning thresholds on the eyes-off duration of EOR and AGB are contrasted in Tab. 6.1.

Tabular 6.1: Effective Warning Threshold on Eyes-Off Duration of Warning Systems in User Test

Driving Speed	**Effective Warning Threshold**	
	EOR	AGB
20 km/h	$\tau = 2.0$ s	$\tau = 3.0$ s
40 km/h		$\tau = 2.0$ s
60 km/h		$\tau = 1.3$ s

Using calibration data of several participants the average times between warnings obtained on the calibration data generalized to the user test. Here, very similar results were observed when simulating the warning system on data of driving without a warning system as can be seen in in Fig. 6.12. We wish to note that indeed a larger number of participants was required for proper calibration due to large inter-individual differences in glance behavior.

6.4.5 Experimental Design and Measures

Previously, we introduced the population of participants, the protocol and the parameterization of the warning systems. In this section we will describe the experimental design and the measures obtained from the behavioral data. The design and the measures will later be used to formulate testable hypotheses in Sec. 6.4.6.

Experimental Design

The experiment had a 3 (warning system) $\times$ 3 (driving speed) repeated measures design. Warning system had three levels: No warning system active, warning system based on the eyes-off-road algorithm active and warning system based on computation of appropriate glance behavior active. The no warning system level investigated driving and glance behavior of the drivers without warning system, while the other levels assessed the effects of both variants of the warning system. The factor driving speed had the three levels of 20, 40, 60 km/h. The warning system factor was presented in blocks with the no warning system condition being the first and a random order of the variants of the warning system. The factor driving speed was fully randomized. We repeated every combination of factors and participant.

Measures

The goal of the experiment was to first demonstrate the feasibility of computing appropriate glance behavior online. Second, the experiments served to study the effects of the warning systems on driver's behavior as well as how they are received by the drivers. Specifically, we were interested in the following measures:

- **The time required to compute the glance policy underlying appropriate glance behavior**: This was done using the utilities of the CANape software to profile the dynamic link library (DLL) compiled from C code, which was generated from the original SIMULINK model in the first place.

- **User ratings of the number of warnings, timing of warnings and usefulness of warnings**: Ratings of the number and the timing of warnings were treated as metric data. This was done using the numeric values $\rho_{\text{Number,Timing}}$ from 1 for very few warnings and very soon warnings to 5 for very many warnings and very late warnings which the participants had typed (see Sec. 6.4.3). As a first step values of the repetitions were averaged. Here, we analyzed the data with respect to mean statistics. Strictly, the obtained ratings are ordinal data, i.e. the ratings only indicate order preferences. However, for Likert-type scales, as ours, a pragmatic approach is often used in human factors research. This is also justified by several experimental studies [184, 206] (see also [30], *Messtheoretische Probleme bei Rating-Skalen*). In our case metric treatment is supported by the following aspects: First, the ratings are subjective judgments of the well-defined metric

quantities number and timing of warnings. Furthermore, ratings were conducted using buttons numbered 1 to 5 which supports metric interpretation by the raters (see Fig. 6.11). In addition to the absolute values we also considered the deviation from the ideal rating of the number of the warnings and the timing of warnings $|\rho_{\text{Number,Timing}} - 3|$. Similarly, the ratings of usefulness were numerically coded and analyzed with respect to means using the corresponding standard parametric procedures.

- **Lane keeping performance measured by the Standard Deviation (STD) of the position in lane and by the Root Mean Squared Error (RMSE) of the lane position** (in terms of deviation from the lane center):
  Similar as in the case of the analysis of the data from the driving experiments these metrics were chosen as they are standard measures in distraction research [252] (Cpt. 7, *Measuring the Effects of Driver Distraction*). In this context, measures were computed using the entire period of secondary task engagement and were analyzed with respect to means using the standard parametric procedures.

- **Steering performance as measured by the root mean squared error of both the steering angle and the steering angle velocity** (in terms of deviation from zero):
  As noted before (Sec. 3.3.1) also both measures of the steering behavior are standards in human factor research and have been analyzed in several other works [252]. Compared to measures of the position in lane, measures based on the steering angle are more sensitive with respect to participant's behavior. Measures were computed using the entire period of secondary task engagement.

- **Glance behavior as measured by the mean, the median, the** 0.75**-quantile and the** 0.95**-quantile of the duration of glances off the road**:
  As noted in [245, 55] effects on glance behavior are often not visible in "central" statistics such as mean and median of the glance durations. Instead, differences are more pronounced in higher quantiles. This was also observed in previous driving experiments of this thesis in Sec. 4.6.3 and in Sec. 5.6.3. We computed these statistics using the entire period of secondary task engagement. Similar as in the cases of the other measures we analyzed these statistics with respect to means using the standard parametric procedures.

With respect to all of the measures, repetitions of an experimental condition (see Sec. 6.4.5) by a participant were averaged and the mean value was used in analysis.

6.4.6 Hypotheses

Based on the protocol, the experimental design and measures we arrive at several testable hypotheses to investigate the research questions stated in Sec. 6.4:

H1 The implementation of the computation of appropriate glance behavior obtains step times smaller than the sample time of 0.04 s (*hypothesis wrt. research question R1; in the evaluation of the MATLAB implementation already step times of* 1 *s were observed [Sec. 3.6.2] which are expected to significantly decrease in the compiled C code*).

H2 The speed adaptive warnings produced by AGB receive better ratings then those produced by EOR using a fixed threshold (*hypothesis wrt. research question R2; drivers show adaptive glance behavior [207, 205], hence it is expected that a warning system that is capable of similar adaption is better received*).

- H2.1 Mean marginal deviation (expectation wrt. to driving speeds) from "ideal" of the ratings of the number of warnings is smaller for AGB than for EOR.
- H2.2 Mean marginal deviation from "ideal" of the ratings of the timing of warnings is smaller for AGB than for EOR.
- H2.3 Mean marginal usefulness of warnings is higher for AGB than for EOR.

H3 The ratings of the warnings produced by EOR vary stronger with respect to driving speed than those produced by AGB (*hypothesis wrt. research question R2; drivers show adaptive glance behavior [207, 205], hence it is expected that ratings of the warning system EOR based on a static threshold will strongly vary with respect to the driving speed*).

- H3.1 Ratings of the number of warnings produced EOR vary stronger with respect to driving speed than those produced by AGB.
- H3.2 Ratings of the timing of warnings produced EOR vary stronger with respect to driving speed than those produced by AGB.
- H3.3 Ratings of the number of warnings produced EOR vary stronger with respect to driving speed than those produced by AGB.

H4 Both warning systems improve driving performance compared to driving without a warning system (*hypothesis wrt. research question R3; lane keeping performances is reduced by secondary task engagement which is expected to be partially mitigated by the warning systems by helping drivers to improve glance strategies*).

- H4.1 Mean marginal STD lane position under both warning systems is decreased compared to driving without a warning system.
- H4.2 Mean marginal RMSE lane position under both warning systems is decreased compared to driving without a warning system.
- H4.3 Mean marginal RMSE steering angle under both warning systems is decreased compared to driving without a warning system.
- H4.4 Mean marginal RMSE steering angle velocity under both warning systems is decreased compared to driving without a warning system.

H5 Warnings based on AGB result in improved lane keeping performance compared to the warnings produced by EOR (*hypothesis wrt. research question R3; the warnings under AGB relate to a rational glance strategy taking into account vehicle physics, hence improved effectiveness is expected*).

- H5.1 Mean marginal STD of the lane position under AGB is decreased compared to EOR.
- H5.2 Mean marginal RMSE lane position under AGB is decreased compared to EOR.
- H5.3 Mean marginal RMSE steering angle under AGB is decreased compared to EOR.
- H5.4 Mean marginal RMSE steering angle velocity under AGB is decreased compared to EOR.

H6 Both warning systems result in decreased off-road glance durations compared to driving without a warning system (*hypothesis wrt. research question R4; drivers are expected to reduce the duration of glances of the road to avoid receiving warnings [53, 5]*).

- H6.1 Mean marginal mean duration of glances off the road under driving with active warning system is decreased compared to driving without warning system.
- H6.2 Mean marginal median duration of glances off the road under driving with active warning system is decreased compared to driving without warning system.
- H6.3 Mean marginal 0.75-quantile of duration of glances off the road under driving with active warning system is decreased compared to driving without warning system.
- H6.4 Mean marginal 0.95-quantile of duration of glances off the road under driving with active warning system is decreased compared to driving without warning system.

6.5 Results

In the following section we present the results of the user test. Here, data was analyzed by means of two-way repeated ANOVAs using the Greenhouse-Geisser approximation where sphericity was violated. Post-hoc tests were conducted by means of the Tukey-test in case of homoscedasticity or the Dunnett-T3-method in case of heteroscedasticity. Furthermore, in the presentation of the results the symbol μ is used to denote the mean statistic, σ denotes the standard deviation and σ^{μ} denotes the standard error (standard deviation of the estimate of the mean). We will first report on the CPU-times required for executing the main components of the warning systems, then the subjective ratings made by the participants are considered followed by the quantities of lane keeping performance. Finally, glance behavior is considered.

6.5.1 CPU-Times

We start presenting the results by first considering those that address the research question R1. That is, we report on the CPU-times of the main algorithmic components of the warning systems. The statistics of the times required for full execution of the preprocessing of the eye-tracking data and the EOR algorithm as well as the times required for computing appropriate glance behavior are shown as box-plots in Fig. 6.13.

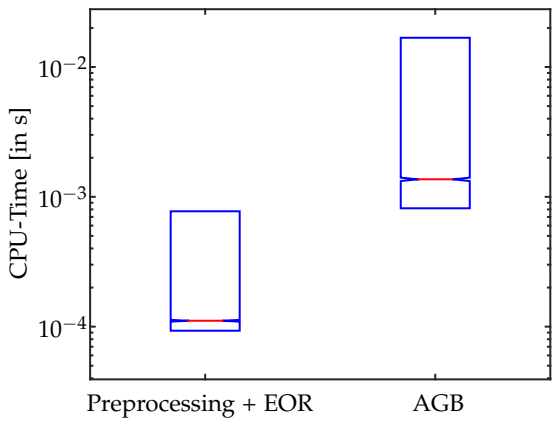

Figure 6.13: CPU-times of the main algorithmic components of the warning system. Left box depicts the distribution of CPU-times of the preprocessing of eye-tracking data and executing the EOR approach. Right box shows the CPU-times of the preprocessing of CAN data and the computations required for realizing AGB (with policy computation being the computational bottleneck). Red line in box indicates the median CPU-time, while the blue box depicts the interval from the 0.25 to the 0.75 quantile.

In the periods where the drivers engaged into the secondary task the median CPU-time of the preprocessing of eye-tracking data and executing EOR was at 1.11×10^{-4} s. The 0.25 quantile was at 9.30×10^{-5} s and the 0.75 quantile was at 7.74×10^{-4} s. In all periods of secondary task engagement there was a single excess of the sample time of 0.02 s (see Fig. 6.1 for the sample times of the components of the warning system). Executing AGB had a median CPU-time of 1.36×10^{-3}. Furthermore, the 0.25 quantile of the CPU-time was at 8.16×10^{-4} and the 0.75 quantile of the CPU-time was at 1.68×10^{-2}. Two excesses of the sample time of 0.04 s were present in the periods of secondary task engagement.

All excesses of sample time occurred when the gaze of the participant was off the road, however, the small number does not allow to draw any conclusion with respect to influencing factors. As a summary of the analysis of CPU-times we can conclude:

Hypothesis H1 is confirmed: The implementation of computing appropriate glance behavior obtains step times smaller than the sample time of 0.04 *s.*

6.5.2 Ratings

Second, we report on the ratings of number of warnings, timing of warnings and usefulness of the warning systems made by the participants in the user test. These aspects are related to the research question R2.

For descriptive purpose, the ratings for the individual categories are depicted as histograms in Fig. 6.14.

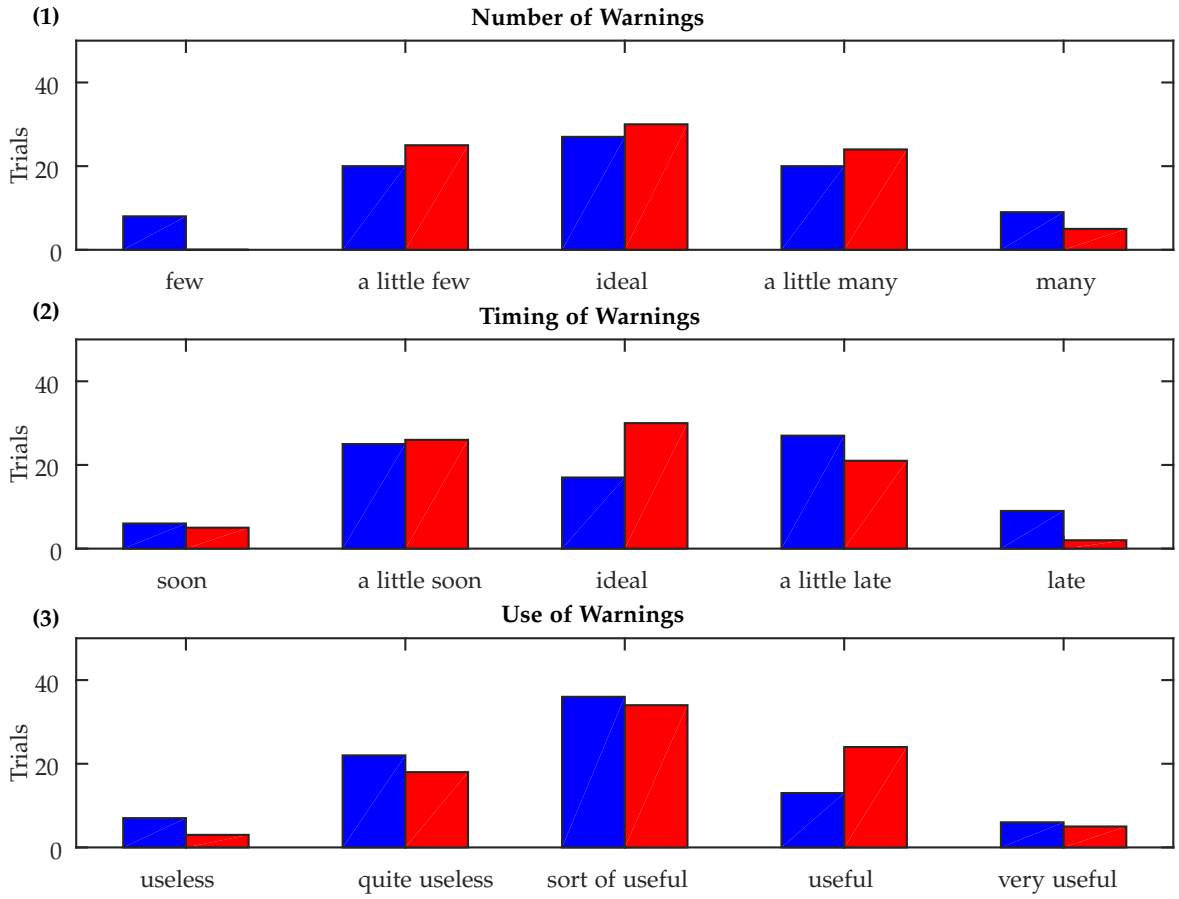

Figure 6.14: Histogram of ratings given by participants in the user test. Blue bars depict the ratings of the warning system based on ■ EOR and red bars depict the ratings of the warning system based on ■ AGB. **(1)** shows the ratings of the amount warnings, **(2)** shows the ratings of the timing of the warnings and **(3)** depicts the usefulness of the warning system for safe driving.

Generally, the distribution of ratings made for AGB was uni-modal and centered around the central item (“ideal”, “sort of useful”). In contrast, the ratings of the timing of the warnings produced by EOR showed a strong bi-modal distribution. Most often participants rated the timing of warning as either “a little soon” or “a little late”. Furthermore, the distribution of the ratings of usefulness of warnings of EOR was stronger concentrated on lower usefulness while the distribution of warnings of AGB concentrated on higher usefulness.

In the further steps of the analysis ratings were treated as metric variables according to their numeric coding (Sec. 6.4.5). We first analyzed the ratings by means of a two-way repeated measures ANOVA whose results are summarized in Tab. 6.2.

Tabular 6.2: Repeated Measures ANOVA of Ratings

Dependent Variable	**Factor**		
	System	Speed	System × Speed
Rating Number	$F(1,13) = 0.34$ $p_{\text{test}} = 0.57$	$F(2,26) = 9.96$ $p_{\text{test}} < 0.01$	$F(2,26) = 6.17$ $p_{\text{test}} < 0.01$
Deviation from "Ideal" Number	$F(1,13) = 5.51$ $p_{\text{test}} = 0.03$	$F(2,26) = 0.57$ $p_{\text{test}} = 0.57$	$F(2,26) = 1.17$ $p_{\text{test}} = 0.33$
Rating Timing	$F(1,13) = 2.13$ $p_{\text{test}} = 0.17$	$F(2,26) = 8.22$ $p_{\text{test}} < 0.01$	$F(2,26) = 5.30$ $p_{\text{test}} = 0.01$
Deviation from "Ideal" Timing	$F(1,13) = 4.84$ $p_{\text{test}} = 0.05$	$F(2,26) = 0.01$ $p_{\text{test}} = 0.99$	$F(2,26) = 1.16$ $p_{\text{test}} = 0.33$
Rating Usefulness	$F(1,13) = 4.38$, $p_{\text{test}} = 0.06$	$F(2,26) = 1.57$ $p_{\text{test}} = 0.23$	$F(2,26) = 2.23$ $p_{\text{test}} = 0.13$

Considering the deviation of the ratings of the number of warnings and the timing of warnings from "ideal" the following results were obtained: In both cases there was only a significant main effect of the warning system. Specifically, the warnings produced by AGB received a smaller marginal (integration over driving speeds) mean deviation from the ideal rating as shown in Tab. 6.3.

Tabular 6.3: Marginal Means of the Ratings

Warn. Sys.	**Rating Category**		
	Dev. from "Ideal" Number	Dev. from "Ideal" Timing	Usefulness of Warning Systems
EOR	$\mu = 0.88$, $\sigma^{\mu} = 0.09$	$\mu = 0.90$, $\sigma^{\mu} = 0.08$	$\mu = 2.87$, $\sigma^{\mu} = 0.15$
AGB	$\mu = 0.61$, $\sigma^{\mu} = 0.10$	$\mu = 0.69$, $\sigma^{\mu} = 0.09$	$\mu = 3.12$, $\sigma^{\mu} = 0.16$

Consequently, we can conclude that the number and the timing of warnings produced by AGB is better received by the drivers.

Hypotheses H2.1 and H2.2 are confirmed: Warnings produced by AGB receive significantly better ratings of timing of warnings and number of warnings than those produced by EOR

In contrast, in the ratings of usefulness there was a strong tendency towards a significant main effect of the warning system. From Tab. 6.3 it can be seen that AGB received a higher marginal mean rating of usefulness than EOR. However the p-value of $p_{\text{test}} = 0.06$ of the main effect slightly exceeded the significance level of $p_{\text{test}} = 0.05$.

Hypothesis H2.3 is not confirmed: Warnings produced by AGB receive no significantly better ratings of usefulness of warnings that those produced by EOR

The relevant statistics for hypothesis H2 are depicted in summarized form in Fig. 6.15.

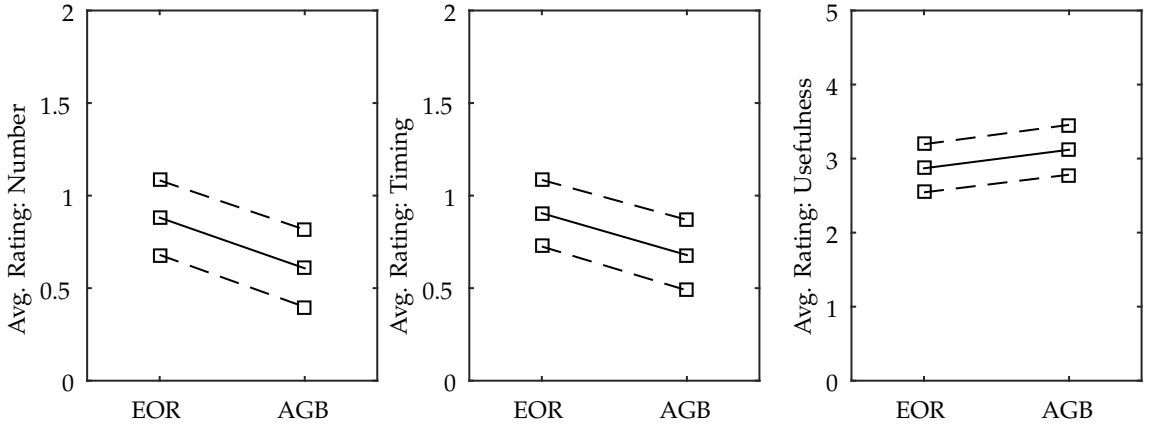

Figure 6.15: Marginal mean ratings for the warning systems. First plot shows the deviation from "ideal" of the ratings of the number of warnings, second plot depicts the deviation from "ideal" of the ratings of the timing of warnings and the third plot shows the ratings of the usefulness of the warning systems. Solid lines indicate the marginal means and dashed lines the 0.95 confidence intervals of the means of the different ratings.

To obtain a better understanding of the differences in the ratings of the different warning systems also the absolute ratings of the number of warnings and timing of warnings were analyzed. As can be seen in Tab. 6.2, with respect to these absolute ratings no significant main effect of the warnings system but a significant main effect of the driving speed factor as well as a significant interaction was present.

We analyze the interaction effect graphically: Fig. 6.16 shows the ratings of the number of warnings for the warnings produced by both warning systems at the different driving speeds.

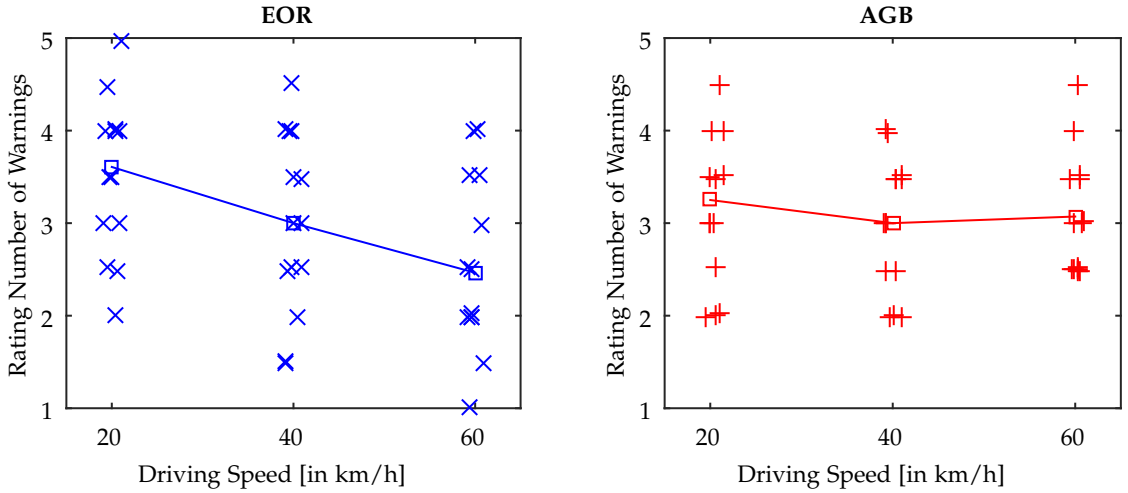

Figure 6.16: Interaction between warning system and driving speed wrt. rating of number of warnings. Left plot shows the ratings of EOR for the individual driving speeds. Right plot shows the ratings of AGB for the individual driving speeds. The ratings (average over both runs) made by the participants are indicated by $\times$ and $+$ and are randomly jittered by a factor 0.1 to improve visibility.

As can be seen in both the left and the right plot of Fig. 6.16 the number of warnings was rated "ideal" at the driving speed of 40 km/h. The number of warnings resulting from EOR show a very clear trend from "a little few warnings" to "a little many warnings" from driving speed 20 km/h to 60 km/h. In contrast, the number of warnings resulting from AGB was almost constantly rated as "ideal". These results from graphical analysis could also formally be established by comparing mean ratings of the warning systems at the different driving speeds as shown in Tab. 6.4 and Tab. 6.5.

Tabular 6.4: Differences of Rating Number wrt. Driving Speed for EOR

Driving Speed	**Difference in Rating Timing wrt. Speed for EOR**	
	to 20 km/h	to 40 km/h
40 km/h	$\mu = -0.61$, $\sigma^{\mu} = 0.22$, $p_{\text{test}} = 0.04$	
60 km/h	$\mu = -1.14$, $\sigma^{\mu} = 0.25$, $p_{\text{test}} < 0.01$	$\mu = -0.53$, $\sigma^{\mu} = 0.24$, $p_{\text{test}} = 0.11$

Tabular 6.5: Differences of Rating Number wrt. Driving Speed for AGB

Driving Speed	**Difference in Rating Number wrt. Speed for AGB**	
	to 20 km/h	to 40 km/h
40 km/h	$\mu = -0.25$, $\sigma^{\mu} = 0.17$, $p_{\text{test}} = 0.34$	
60 km/h	$\mu = -0.17$, $\sigma^{\mu} = 0.17$, $p_{\text{test}} = 0.56$	$\mu = +0.07$, $\sigma^{\mu} = 0.15$, $p_{\text{test}} = 0.88$

Here, the ratings of the number of warnings of EOR significantly differed with respect to the driving speeds whereas no significant differences were present in the ratings of AGB.

Hypothesis H3.1 is confirmed: The ratings of the number of warnings of warning system EOR vary stronger with respect to driving speed than the ratings of the warning system AGB.

The ratings of the timing of the warnings are analyzed in similar graphical fashion. Fig. 6.17 shows the ratings made by the participants for both warning systems at the different driving speeds.

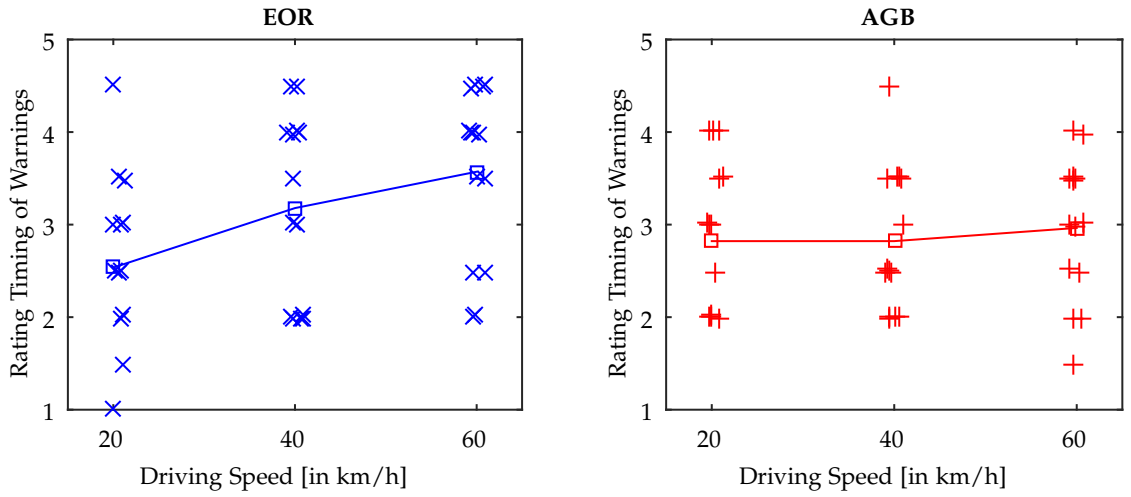

Figure 6.17: Interaction between warning system and driving speed wrt. rating of number, the timing of warnings. Left plot shows the ratings of EOR for the individual driving speeds. Right plot shows the ratings of AGB for the individual driving speeds. The ratings (average over both runs) made by the participants are indicated by $\times$ and $+$ and are randomly jittered by a factor of 0.1 to improve visibility.

Considering the ratings for EOR in the left plot of Fig. 6.17 a clear trend from "a little early warnings" at 20 km/h to "a little late warnings" at 60 km/h can be seen. Analogously to the case of the ratings of the number of warnings the ratings of the timing of warnings of AGB shown in the right plot do not exhibit such a trend. Instead, ratings are constantly centered around "ideal timing". These observations are also supported by a statistical analysis of the mean ratings at the different driving speeds whose results are summarized in Tab. 6.6 and Tab. 6.7.

Tabular 6.6: Differences of Rating Timing wrt. Driving Speed for EOR

Driving Speed	**Difference in Rating Timing wrt. Speed for EOR**	
	to 20 km/h	to 40 km/h
40 km/h	$\mu = 0.64$, $\sigma^{\mu} = 0.17$, $p_{\text{test}} < 0.01$	
60 km/h	$\mu = 1.04$, $\sigma^{\mu} = 0.23$, $p_{\text{test}} < 0.01$	$\mu = 0.39$, $\sigma^{\mu} = 0.21$, $p_{\text{test}} = 0.21$

Tabular 6.7: Differences of Rating Timing wrt. Driving Speed for AGB

Driving Speed	**Difference in Rating Timing wrt. Speed for AGB**	
	to 20 km/h	to 40 km/h
40 km/h	$\mu = 0.00$, $\sigma^{\mu} = 0.18$, $p_{\text{test}} = 1.00$	
60 km/h	$\mu = 0.14$, $\sigma^{\mu} = 0.21$, $p_{\text{test}} = 0.78$	$\mu = 0.14$, $\sigma^{\mu} = 0.21$, $p_{\text{test}} = 0.77$

The analysis of the marginal means revealed significant differences between the ratings of the timing of warnings of EOR at the different driving speeds. With respect to the warnings resulting from AGB instead no significant differences between the driving speeds could be established.

Hypothesis H3.2 is confirmed: The ratings of the timing of warnings of warning system EOR vary stronger with respect to driving speed than the ratings of the warning system AGB.

Considering the ratings of usefulness of the warnings, no significant influence of the driving speed nor a significant warning system driving speed interaction could be established. That is, the ratings of usefulness of the warnings systems were largely constant over driving speeds.

Hypothesis H3.3 is not confirmed: The ratings of usefulness of warning system EOR does not show significantly higher variation with respect to driving speed than the ratings of usefulness of the warning system AGB.

6.5.3 Position in Lane

Third, the position in lane in the periods of secondary task engagement is considered. As in previous driving experiments, we analyzed both the STD of the lane position as well as the RMSE of the lane position which are related to research question R3.

We first depict the distributions of the STD under the different warning system conditions and driving speed conditions in Fig. 6.18

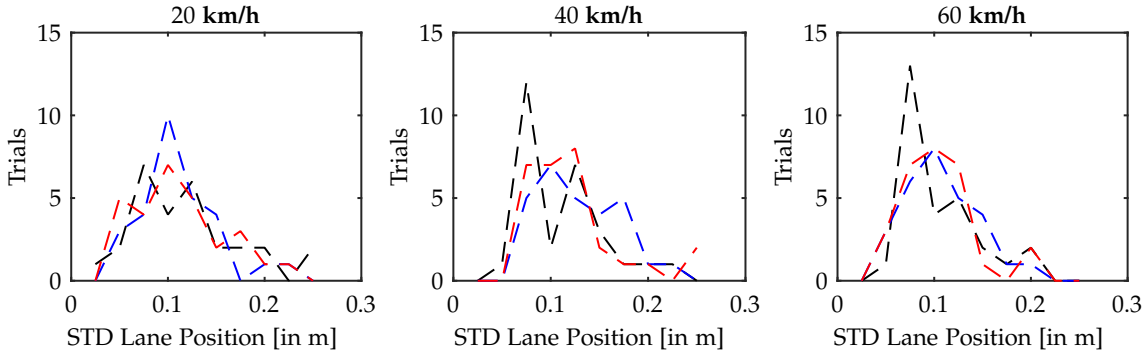

Figure 6.18: Histograms of STD of the lane position for the individual driving speeds and the warning systems. Plots show the distribution of the STD of the lane position for driving speeds 20, 40, 60 km/h from left to right. Dashed **black** lines indicate the distribution for driving without warning system, dashed **blue** lines indicate the distribution for driving with EOR, dashed **red** lines indicate the distribution for driving with AGB.

As can be seen in the plots, the distributions of the STD of the lane position are largely similar. In all experimental conditions the STD has a dominant mode at approximately 0.1 m as well as a short left tail towards 0 m and a moderate tail to 0.25 m. In addition to the optical impression also the statistics of the distribution of the STD are similar. This is can be seen in Tab. 6.8 which reports the means and standard deviations of the STD of the lane position.

Tabular 6.8: Statistics of STD Lane Position

Warn. Sys.	**STD Lane Position**		
	20 km/h	40 km/h	60 km/h
none	$\mu = 0.12, \sigma = 0.06$	$\mu = 0.11, \sigma = 0.05$	$\mu = 0.10, \sigma = 0.04$
EOR	$\mu = 0.11, \sigma = 0.04$	$\mu = 0.13, \sigma = 0.04$	$\mu = 0.11, \sigma = 0.04$
AGB	$\mu = 0.11, \sigma = 0.05$	$\mu = 0.12, \sigma = 0.06$	$\mu = 0.10, \sigma = 0.04$

In addition to the STD Fig. 6.19 shows the distribution of the RMSE of the lane position in the different experimental conditions.

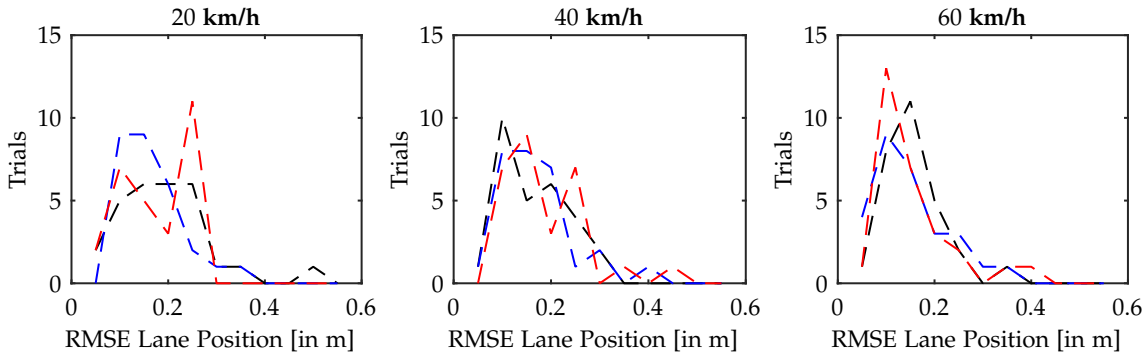

Figure 6.19: Histograms of the RMSE of the lane position for the individual driving speeds and the warning systems. Plots show the distributions of the RMSE of the lane position for driving speeds 20, 40, 60 km/h from left to right. Dashed **black** lines indicate the distribution for driving without warning system, dashed **blue** lines indicate the distribution for driving with EOR, dashed **red** lines indicate the distribution for driving with AGB.

Similar as in the case of the distribution of the STD of the lane position also the distribution of the RMSE of the lane position are largely similar. The distribution of the RMSE shows a longer tail to the right (high deviation from the lane center) than the STD. Furthermore, the individual distributions show a mode at approximately 0.1 m. The mean and standard deviation of the RMSE of the lane position are presented in Tab. 6.9.

Tabular 6.9: Statistics of RMSE Lane Position

Warn. Sys.	**RMSE Lane Position**		
	20 km/h	40 km/h	60 km/h
none	$\mu = 0.19$, $\sigma = 0.09$	$\mu = 0.16$, $\sigma = 0.07$	$\mu = 0.15$, $\sigma = 0.06$
EOR	$\mu = 0.16$, $\sigma = 0.07$	$\mu = 0.17$, $\sigma = 0.07$	$\mu = 0.15$, $\sigma = 0.07$
AGB	$\mu = 0.18$, $\sigma = 0.06$	$\mu = 0.18$, $\sigma = 0.08$	$\mu = 0.15$, $\sigma = 0.08$

Similar to the STD, also the statistics of the distribution of the RMSE of the lane position show little differences between the experimental conditions.

We statistically analyzed the influence of the factors warnings system and driving speed on both metrics of the position in lane by means of a two-way repeated measures ANOVA. The results of this procedure are summarized in Tab. 6.10.

Tabular 6.10: Repeated Measures ANOVA of Lane Position

Dependent Variable	**Factor**		
	System	Speed	System × Speed
STD Lane Position	$F(2,26) = 0.05$ $p_{\text{test}} = 0.95$	$F(2,26) = 1.53$ $p_{\text{test}} = 0.24$	$F(4,52) = 1.92$ $p_{\text{test}} = 0.14$
RMSE Lane Position	$F(2,26) = 0.60$ $p_{\text{test}} = 0.55$	$F(2,26) = 2.27$ $p_{\text{test}} = 0.12$	$F(4,52) = 1.53$ $p_{\text{test}} = 0.20$

The small differences between the distributions of both metrics of the lane position is formally established by the ANOVA: There were no significant main effect of neither the factor warning system nor the factor driving speed. Furthermore, also no significant interaction effects were present. The position in lane under secondary task interaction was not influenced by driving speed and was also not affected by the treatment with the distraction warning system.

Hypotheses H4.1, H4.2, H5.1 and H5.2 are not confirmed: Neither mean STD of the lane position nor mean RMSE of the lane position are significantly affected by the treatment with the distraction warning system. Furthermore, no significant differences between the treatment with the different warning systems are present.

6.5.4 Steering Behavior

Fourth, we report on the RMSE of the steering angle and the RMSE of the steering angle velocity. As the applied steering effort also contributes to lane keeping performance, this analysis addresses research question R3.

We first illustrate the distribution of the RMSE of the steering angle in Fig. 6.20.

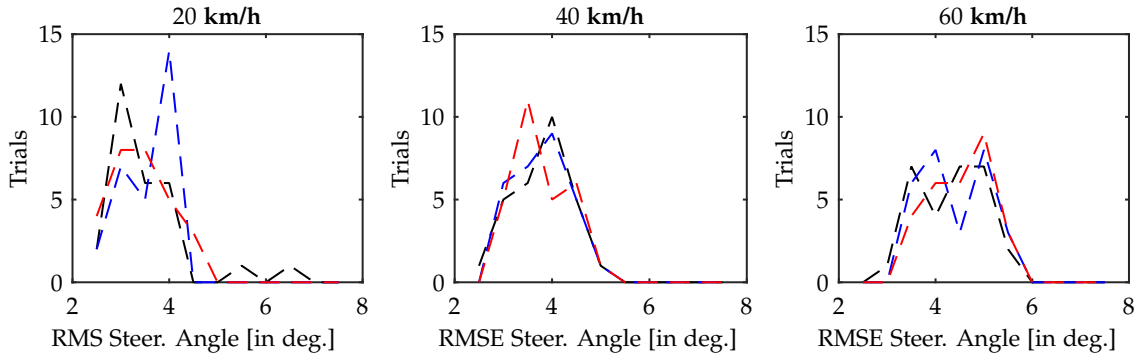

Figure 6.20: Histograms of the RMSE of the steering angle for the individual driving speeds and the warning systems. Plots show the distribution of the RMSE of the steering angle for driving speeds 20, 40, 60 km/h from left to right. Dashed **black** lines indicate the distribution for driving without warning system, dashed **blue** lines indicate the distribution for driving with EOR, dashed **red** lines indicate the distribution for driving with AGB.

The plots of Fig. 6.20 show that the distribution of the RMSE of the steering angle was shifted towards increased steering angles under increased driving speed. In contrast, the distributions under the different conditions of the warning system were similar.

Different observations were made wrt. distributions of the RMSE of the steering angle velocity which are depicted in Fig. 6.21.

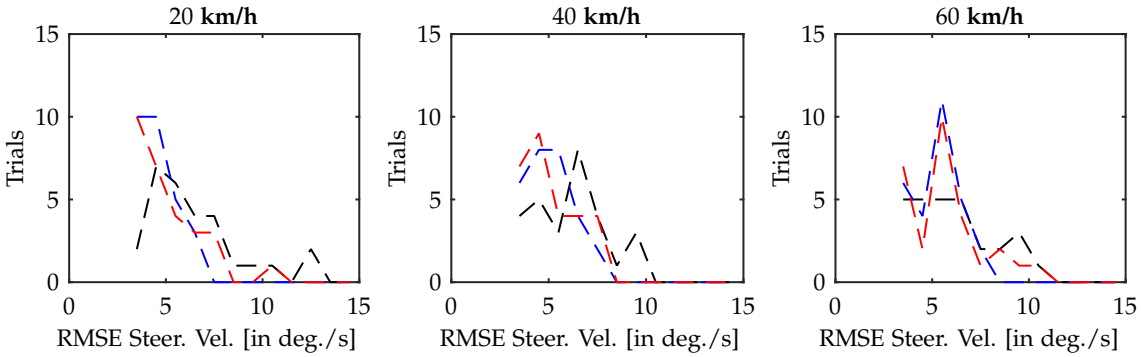

Figure 6.21: Histograms of the RMSE of the steering angle velocity for the individual driving speeds and the warning systems. Plots show the distribution of the RMS of the steering angle velocity for driving speeds 20, 40, 60 km/h from left to right. Dashed **black** lines indicate the distribution for driving without warning system, dashed **blue** lines indicate the distribution for driving with EOR, dashed **red** lines indicate the distribution for driving with AGB.

As can be seen in the plots the distribution of the RMSE of steering angle velocity is only weakly affected by the driving speed. In contrast, especially at driving speeds 20 km/h and 40 km/h more probability mass is at lower RMSE of the steering angle velocity for active warning systems EOR and AGB.

A two-way repeated measures ANOVA was conducted to statistically analyze the RMSE of the steering angle and the RMSE of the steering angle velocity. The results of the ANOVA are summarized in Tab. 6.11.

Tabular 6.11: Repeated Measures ANOVA of Steering Angle and Steering Angle Velocity

Dependent Variable	**Factor**		
	System	Speed	System × Speed
RMSE Steering Angle	$F(2,26) = 0.62$ $p_{\text{test}} = 0.54$	$F(2,26) = 37.13$ $p_{\text{test}} < 0.01$	$F(4,52) = 0.61$ $p_{\text{test}} = 0.66$
RMSE Steering Angle Velocity	$F(2,26) = 10.00$ $p_{\text{test}} < 0.01$	$F(2,26) = 1.20$ $p_{\text{test}} = 0.32$	$F(4,52) = 2.18$ $p_{\text{test}} = 0.08$

The ANOVA revealed a significant main effect of driving speed on the RMSE of the steering angle. There was no significant main effect of the warning system factor and no significant interaction. Consequently, treatment with the warning system had no effect on the steering angle as measured by the RMSE. Similar, also no difference between both warning systems was present.

Hypotheses H4.3, H5.3 are not confirmed: The distraction warning systems show no significantly reduced RMSE of the steering angle. In addition to that AGB does not significantly reduce RMSE of the steering angle compared to EOR.

To further investigate the effect of driving speed on the RMSE of the steering angle post-hoc tests were conducted. The results of the comparison of marginal (integration over warnings system factor) mean RMSE of the steering angle are shown in Tab. 6.12.

Tabular 6.12: Marginal Means of RMSE Steering Angle wrt. Driving Speed

Driving Speed	**RMSE Steering Angle**	**Diff. to** 20 km/h	**Diff. to** 40 km/h
20 km/h	$\mu = 3.50$, $\sigma^{\mu} = 0.08$		
40 km/h	$\mu = 3.81$, $\sigma^{\mu} = 0.07$	$p_{\text{test}} = 0.02$	
60 km/h	$\mu = 4.42$, $\sigma^{\mu} = 0.08$	$p_{\text{test}} < 0.01$	$p_{\text{test}} < 0.01$

The post-hoc test showed that the RMSE of the steering angle monotonously increases with driving speed.

In contrast to the steering angle a significant main effect of the warning system was present in the RMSE of the steering angle velocity. Tab. 6.13 shows the results of a post-hoc analysis of the marginal mean (integrated over driving speed) RMSE of the steering angle velocity.

Tabular 6.13: Marginal Means of RMSE Steering Angle Velocity

Warn. Sys.	**RMSE Steering Angle Velocity**	**Diff. to** none	**Diff. to** EOR
none	$\mu = 6.18$, $\sigma^{\mu} = 0.48$		
EOR	$\mu = 4.97$, $\sigma^{\mu} = 0.28$	$p_{\text{test}} = 0.01$	
AGB	$\mu = 5.20$, $\sigma^{\mu} = 0.40$	$p_{\text{test}} < 0.01$	$p_{\text{test}} = 0.66$

Comparison of the marginal means shows that the RMSE of the steering angle velocity is significantly reduced under treatment with either of the warning systems. However, there are no significant differences between both warning systems.

Hypothesis H5.4 is not confirmed: AGB does not result in significantly reduced RMSE of the steering angle velocity compared to EOR.

A typical issue that can arise in within-subject designs as ours are training effects. In the present driving experiment none of the participants was familiar with the employed test vehicle. Furthermore driving without a warning system was always first in the protocol (see Sec. 6.4.3). Hence, the main effect of the warning system could possibly also be attributed to driving experience with the vehi-

cle. Therefore, an additional post-hoc mixed-effects regression analysis with repeated measures was conducted. For this purpose, the model

$$y_{\mathrm{RMSE}\,\dot{\alpha}} = \mu_1 + \lambda x^{\mathrm{trial}} + \mu_2(x^{\mathrm{sys}}) + \mu_3(x^{\mathrm{spd}}) + \mu_4(x^{\mathrm{sys}}, x^{\mathrm{spd}}) + \epsilon_1(x^{\mathrm{par}}) + x^{\mathrm{trial}}\epsilon_2(x^{\mathrm{par}}) + \epsilon_3, \quad (6.17)$$

where $y_{\mathrm{RMSE}\,\dot{\alpha}}$ denotes the RMSE of the steering angle velocity, x^{trial} the number of the trial, x^{sys} the warning system condition, x^{spd} the speed condition and x^{par} denotes an individual participant was used. Terms $\mu_1, \lambda x^{\mathrm{trial}}, \mu_2(x^{\mathrm{sys}}), \mu_3(x^{\mathrm{spd}}), \mu_4(x^{\mathrm{sys}}, x^{\mathrm{spd}})$ are fixed effects and terms $\epsilon_1(x^{\mathrm{par}})$, $x^{\mathrm{trial}}\epsilon_2(x^{\mathrm{par}}), \epsilon_3$ are independent random effects. We report the results of F-tests on the fixed effects in Tab. 6.14.

Tabular 6.14: Regression Model of RMSE Steering Angle Velocity

Dependent Variable	**Fixed Effects**			
	Number of Trial	System	Speed	System × Speed
RMSE Steering Angle	$F(1,242) = 1.17$ $p_{\mathrm{test}} = 0.28$	$F(2,242) = 3.85$ $p_{\mathrm{test}} = 0.02$	$F(2,242) = 2.17$ $p_{\mathrm{test}} = 0.11$	$F(4,242) = 2.42$ $p_{\mathrm{test}} = 0.06$

The F-tests revealed that significant effects can only be attributed to the factor warning system in the employed regression model. The effect of the number of the trial was not significant. Fig. 6.22 depicts the prediction of RMSE of the steering angle velocity using the maximum likelihood parameters of the regression model.

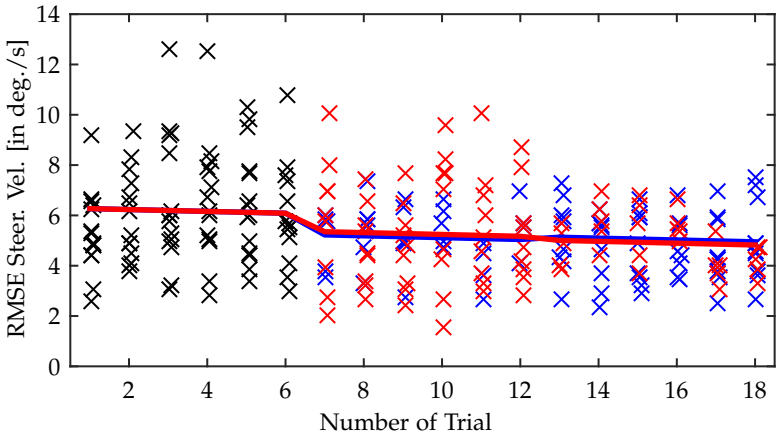

Figure 6.22: Regression analysis of potential training effects on the RMSE of the steering angle velocity. Plot depicts the RMSE of the steering angle velocity for the number of the trial. Individual trials in driving without warning system are depicted by black ×, trials with EOR are denoted by red × and trials with AGB are indicated by blue ×. Solid lines show the mean prediction over participants of the regression model for EOR first in red and for AGB first in blue.

From the regression analysis we conclude that the main effect of the warning system on the RMSE of the steering angle velocity is unlikely the result of a training effect.

Hypothesis H4.4 is confirmed: The distraction warning systems result in significantly reduced RMSE of the steering angle velocity compared driving without a warning system.

6.5.5 Glance Behavior

Finally, the driver's glance behavior is analyzed. Specifically, we investigate the effects of driving speed and warning system on the mean duration, the median duration as well as the 0.75 and the 0.95 quantile of the duration of glances off the road. This analysis is related to research question R4 of the user test.

As a first aspect of the analysis of glance behavior, Fig. 6.23 shows the distribution of the durations of glances off the road.

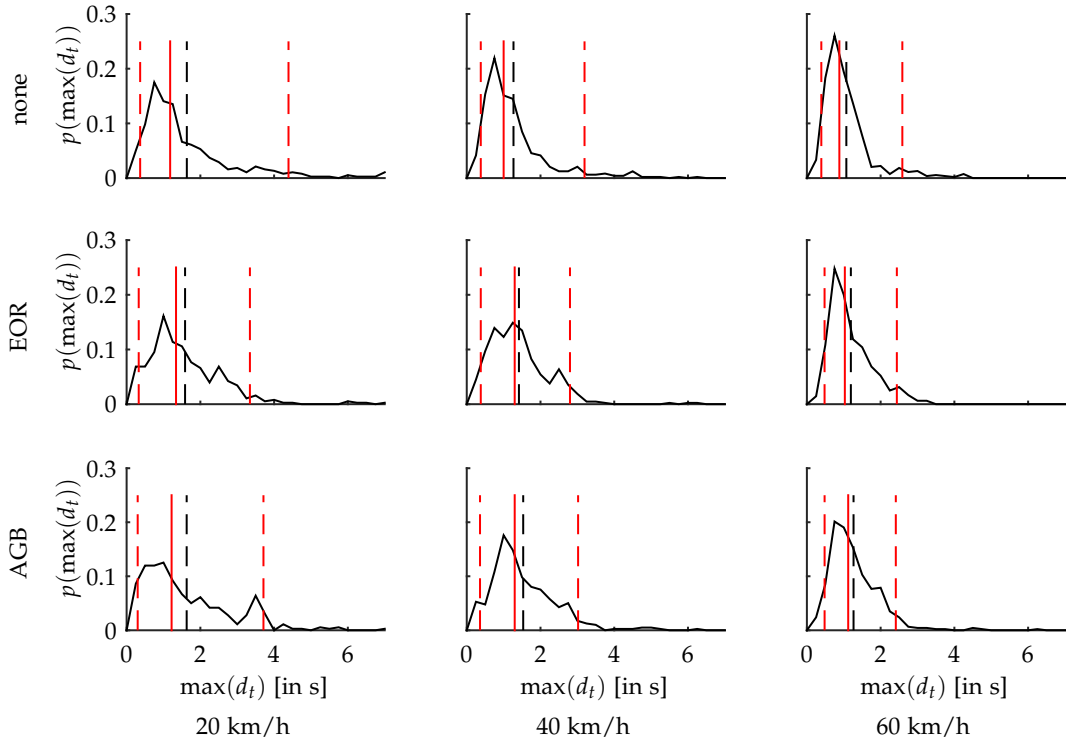

Figure 6.23: Distributions of the durations of glances off the road $\max(d_t)$ at the different driving speeds and for different warning systems. Dashed red lines indicate the $[0.05, 0.95]$ quantiles, while the solid red lines indicate the median. The mean maximum glance duration is denoted by a dashed black line.

As can be seen from the plots, the distribution of glance durations was visibly influenced by the driving speed. In all warning system conditions, the duration of long glances (right tail of the distribution) decreased. In contrast "central" statistics such as median and mean varied only slightly and are approximately at $1.5 - 1.7$ s.

In the following we list the mean and standard deviations of the mean duration in Tab. 6.15, the median duration in Tab. 6.16, the 0.75 quantile in Tab. 6.17 as well as the 0.95 quantile of the duration of glance off the road in Tab. 6.18.

Tabular 6.15: Statistics of the Mean of the Duration of Glances Off the Road

Warn. Sys.	**Mean Dur. Glances Off the Road**		
	20 km/h	40 km/h	60 km/h
none	$\mu = 2.21$, $\sigma = 1.41$	$\mu = 1.57$, $\sigma = 0.96$	$\mu = 1.23$, $\sigma = 0.60$
EOR	$\mu = 1.77$, $\sigma = 0.68$	$\mu = 1.52$, $\sigma = 0.47$	$\mu = 1.29$, $\sigma = 0.41$
AGB	$\mu = 2.07$, $\sigma = 1.03$	$\mu = 1.72$, $\sigma = 0.67$	$\mu = 1.37$, $\sigma = 0.49$

Tabular 6.16: Statistics of the Median of the Duration of Glances Off the Road

Warn. Sys.	**Median Dur. Glances Off the Road**		
	20 km/h	40 km/h	60 km/h
none	$\mu = 2.09$, $\sigma = 1.41$	$\mu = 1.54$, $\sigma = 1.05$	$\mu = 1.19$, $\sigma = 0.61$
EOR	$\mu = 1.62$, $\sigma = 0.70$	$\mu = 1.43$, $\sigma = 0.51$	$\mu = 1.25$, $\sigma = 0.49$
AGB	$\mu = 1.99$, $\sigma = 1.05$	$\mu = 1.64$, $\sigma = 0.67$	$\mu = 1.30$, $\sigma = 0.50$

Tabular 6.17: Statistics of the 0.75 Quantile of the Duration of Glances Off the Road

Warn. Sys.	0.75 **Quant. Dur. Glances Off the Road**		
	20 km/h	40 km/h	60 km/h
none	$\mu = 3.03$, $\sigma = 2.12$	$\mu = 1.91$, $\sigma = 1.19$	$\mu = 1.54$, $\sigma = 0.83$
EOR	$\mu = 2.26$, $\sigma = 0.74$	$\mu = 1.93$, $\sigma = 0.58$	$\mu = 1.62$, $\sigma = 0.51$
AGB	$\mu = 2.54$, $\sigma = 1.27$	$\mu = 2.15$, $\sigma = 0.82$	$\mu = 1.67$, $\sigma = 0.58$

Tabular 6.18: Statistics of the 0.95 Quantile of the Duration of Glances Off the Road

Warn. Sys.	0.95 **Quant. Dur. Glances Off the Road**		
	20 km/h	40 km/h	60 km/h
none	$\mu = 4.03$, $\sigma = 2.41$	$\mu = 2.73$, $\sigma = 1.48$	$\mu = 2.06$, $\sigma = 1.08$
EOR	$\mu = 3.36$, $\sigma = 1.48$	$\mu = 2.77$, $\sigma = 1.04$	$\mu = 2.13$, $\sigma = 0.61$
AGB	$\mu = 3.45$, $\sigma = 1.64$	$\mu = 2.95$, $\sigma = 1.20$	$\mu = 2.35$, $\sigma = 1.04$

To statistically analyze the effects of the individual metrics of glance behavior a two-way repeated measures ANOVA was conducted. We report the results of the ANOVA in Tab. 6.19.

Tabular 6.19: Repeated Measures ANOVA of Glance Behavior

Dependent Variable	**Factor**		
	System	Speed	System $\times$ Speed
Mean Dur. Glances Off the Road	$F(2,26) = 0.44$ $p_{\text{test}} = 0.65$	$F(2,26) = 22.98$ $p_{\text{test}} < 0.01$	$F(4,52) = 1.92$ $p_{\text{test}} = 0.09$
Median Dur. Glances Off the Road	$F(2,26) = 0.93$ $p_{\text{test}} = 0.41$	$F(2,26) = 16.98$ $p_{\text{test}} < 0.01$	$F(4,52) = 2.30$ $p_{\text{test}} = 0.07$
0.75 Quant. Dur. Glances Off the Road	$F(2,26) = 0.35$ $p_{\text{test}} = 0.71$	$F(2,26) = 25.72$ $p_{\text{test}} < 0.01$	$F(4,52) = 3.15$ $p_{\text{test}} = 0.02$
0.95 Quant. Dur. Glances Off the Road	$F(2,26) = 0.67$ $p_{\text{test}} = 0.52$	$F(2,26) = 25.51$ $p_{\text{test}} < 0.01$	$F(4,52) = 2.25$ $p_{\text{test}} = 0.07$

The ANOVA revealed no significant main effect of the warning system in neither of the considered metrics. In contrast, there was a significant main effect of the driving speed on every metric of the glance behavior. Furthermore, interaction effects between warning system and driving speed were present of which the effect was significant for the 0.75 quantile of the glance durations.

The effects of driving speed on glance behavior is investigated in detail by means of post-hoc comparison of the marginal (integration over warning systems) means of the different metrics. We present the results of the comparisons in Tab. 6.20 - 6.23.

Tabular 6.20: Marginal Means of Mean Duration of Glances Off the Road

Driving Speed	**Mean Dur. Glances Off the Road**	**Diff. to** 20 km/h	**Diff. to** 40 km/h
20 km/h	$\mu = 1.98$, $\sigma^{\mu} = 0.22$		
40 km/h	$\mu = 1.57$, $\sigma^{\mu} = 0.13$	$p_{\text{test}} < 0.01$	
60 km/h	$\mu = 1.29$, $\sigma^{\mu} = 0.11$	$p_{\text{test}} < 0.01$	$p_{\text{test}} < 0.01$

Tabular 6.21: Marginal Means of Median Duration of Glances Off the Road

Driving Speed	**Median Dur. Glances Off the Road**	**Diff. to** 20 km/h	**Diff. to** 40 km/h
20 km/h	$\mu = 1.85$, $\sigma^{\mu} = 0.22$		
40 km/h	$\mu = 1.49$, $\sigma^{\mu} = 0.12$	$p_{\text{test}} = 0.02$	
60 km/h	$\mu = 1.25$, $\sigma^{\mu} = 0.12$	$p_{\text{test}} < 0.01$	$p_{\text{test}} < 0.01$

Tabular 6.22: Marginal Means of 0.75 Quantile of Duration of Glances Off the Road

Driving Speed	0.75 **Quant. Dur. Glances Off the Road**	**Diff. to** 20 km/h	**Diff. to** 40 km/h
20 km/h	$\mu = 2.58$, $\sigma^{\mu} = 0.31$		
40 km/h	$\mu = 2.03$, $\sigma^{\mu} = 0.19$	$p_{\text{test}} < 0.01$	
60 km/h	$\mu = 1.61$, $\sigma^{\mu} = 0.15$	$p_{\text{test}} < 0.01$	$p_{\text{test}} < 0.01$

Tabular 6.23: Marginal Means of 0.95 Quantile of the Duration of Glances Off the Road

Driving Speed	0.95 **Quant. Dur. Glances Off the Road**	**Diff. to** 20 km/h	**Diff. to** 40 km/h
20 km/h	$\mu = 3.73$, $\sigma^{\mu} = 0.44$		
40 km/h	$\mu = 2.85$, $\sigma^{\mu} = 0.30$	$p_{\text{test}} < 0.01$	
60 km/h	$\mu = 2.21$, $\sigma^{\mu} = 0.22$	$p_{\text{test}} < 0.01$	$p_{\text{test}} < 0.01$

The interaction effect present in the 0.75 quantile of the glance duration is graphically analyzed. For this purpose, the means and standard errors for the individual warnings system conditions and driving speeds are show in Fig. 6.24.

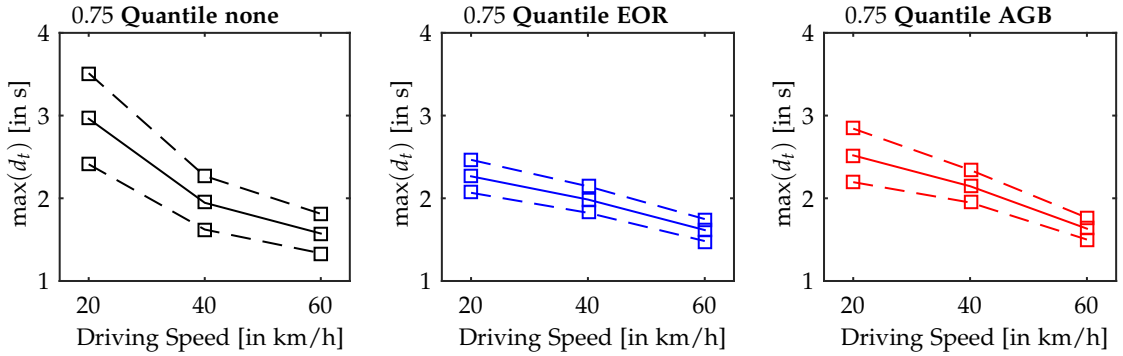

Figure 6.24: Interaction between the warning system and driving speed wrt. the 0.75 quantile of the off-road glance duration. Left plot corresponds to driving without warning system, middle plot corresponds to driving under EOR and right plot corresponds to driving under AGB. Solid lines indicate the mean and dashed lines show the 0.95 confidence interval of the mean.

The plots of the 0.75 quantile of the glance durations in the different driving conditions show a weaker influence of the driving speed factor in the conditions of active warning systems (see Fig. 6.24). However, differences between driving speeds were still significant for EOR and AGB as verified by post-hoc test on the means. Similar observations were made for the other metrics of the glance duration. As the p-value of the interaction effects of those metrics exceeded the significance-level of 0.05 a detailed analysis is omitted in this work.

Summarized, the analysis of glance behavior revealed that the warning systems did not significantly affect the glance behavior in total but rather resulted in decreased effect of the driving speed.

Hypotheses H5.1-4 are not confirmed: The distraction warning systems do not result in significantly reduced durations of glances off the road.

6.6 Discussion

Previously, a user study of the developed warning systems was introduced. In this context, a driving experiment was conducted to provide a proof of concept of the approaches and to investigate the acceptance by the user as well as the effects of the warning systems on driver behavior. In the experiment the warning systems were evaluated at different driving speeds. Following the presentation of the results of the user test, this section will discuss the findings in pursuit of the research questions stated in Sec. 6.4.

6.6.1 Feasibility of Appropriate Glance Behavior for Distraction Warning

Research question R1 asked whether the concept of appropriate glance behavior as defined in Sec. 3.4 and implemented in Sec. 6.3.5 was applicable for real-time distraction warning. This question was mainly motivated by the computational demands of computing policies in the joint task POMDP and the question of the robustness with respect to noisy sensor data.

The CPU-times demonstrate that our implementation of appropriate glance behavior is sufficiently fast to be run at a sample time of 0.04 s. In addition to that, pre-processing of eye-tracking can reliably run at a sample time of 0.02 s. Comparing the CPU-times of the compiled C code of the policy computation (see Sec. 6.5.1) with those obtained by the MATLAB implementation used in the evaluation of Sec. 3.6.2, a speed-up of a factor of more than 25 could be observed. This shows that appropriate glance behavior can be computed sufficiently fast in the scenario of lane keeping. The present warning system is the to-the-best of our knowledge only system for distraction warning based on online solution of POMDPs. Therefore, no comparison to other approaches is possible. We wish to note that other works that addressed exact solution of similar POMDP models considered offline optimization of sensor states [246]. Clearly, this is not feasible in the present application were the adaption to the specific POMDP instance modeling the current driving situation is desired. We refer to the analysis of the computational demands of policy computation prior in this work in Sec. 3.6.2 for a discussion of the aspects that enable online solution.

In the user test three participants needed to be excluded because of insufficient eye-tracking quality (see Sec. 6.4.1). Note that this was rather an issue of the eye tracking than of the employed pre-processing. If eye-tracking is lost for a long period (as in case of squinting shown in Fig. 6.8) and drivers show no significant head movements when averting gaze (which is the case for a considerable amount of drivers [64]) the developed algorithms cannot estimate the eyes-off duration. Considering the subjective ratings of usefulness made by the participants, the distraction warnings based on AGB were well received. Hence the warning system based on AGB demonstrated sufficient robustness with respect to the quality of eye-tracking in real driving of the remaining participants.

The employed sample times were not derived from specific functional requirements. Instead sample times followed largely those of the eye-tracking system and the lane-tracking system. This choice of sample times appeared to be sufficiently small to ensure effectiveness of the warning system which can be seen at the positive user ratings (see Sec. 6.5.5). We even think that the sample time of AGB can possibly be increased up to 0.1 s without significant loss of functionality.

6.6.2 Acceptance Of Distraction Warning Systems

Experienced drivers show glance behavior that is adapted to the driving situation, e.g. the driving speed [207, 205]. Furthermore, it was demonstrated that planning and decision making in assistance system, e.g. collaborative robots, can profit from explicit model of human behavior [54]. Against this background, research question R2 and its related hypotheses H2.1-H2.4 address potential differences in the acceptance of the "baseline" distraction warning system using EOR and the warning system using AGB which features explicit models of the driving situation as well as the driver's vehicle control.

The subjective ratings of number of warnings, timing of warnings and usefulness of warnings clearly demonstrate that warnings adapted to the driving speed produced by AGB are better received by the users. Due to appropriately calibrating the warning systems (see Sec. 6.4.4) average ratings of the number of warnings and timing of warnings were similar for both warning systems. That is, in average both variants of the warning system had the same sensitivity. Consequently, increased deviation of ratings from "ideal" of EOR compared to AGB was attributed to stronger variation of ratings across the driving speeds. The fixed warning threshold of 2.0 s used in EOR was judged too sensitive (too many and too early warnings) at 20 km/h and too liberal (too few and too late warnings) at 60 km/h by the users which can be seen in Fig. 6.16 and Fig. 6.17. This was not the case for the effective warning thresholds on eyes-off duration adapted to the driving speed which resulted from AGB (see Tab. 6.1).

In contrast to the ratings of number of warnings and timing of warnings the ratings of usefulness of AGB were not significantly better than those of EOR. As the increase of ratings of usefulness of AGB compared to EOR was close to significance ($p_{\text{test}} = 0.06$) this was most likely attributed to the small number of participants considered in the analysis. We expect significant effects using an increased number of participants.

Similar to the present study, subjective ratings were used in [126, 119] to assess distraction mitigation systems. However, the results of these works are not directly comparable to the present user test. This is because we aimed at a comparison of both variants of distraction warning systems while the other experiments investigate the overall acceptance of this type of warning system. Furthermore, both protocol and employed scales employed in previous work were quite different from those employed in this thesis.

6.6.3 Effects on Driving Performance

Driver distraction can result in problematically decreased driving performance. Hence, the main purpose of distraction warning systems is to mitigate these decrements by beneficially altering the drivers' gaze switching behavior. Research question R3 and the associated hypotheses H4.1-4 ask if the implemented warning systems can improve driving performance compared to driving without treatment. Hypotheses H5.1-4 additionally ask if the warning system AGB which is adapted to the driving situation shows additional benefits compared to the warning system based on EOR.

The analysis of the data of the user test revealed that the distraction warning systems do not significantly improve driving performance in terms of the position in lane and the steering effort (see Sec. 6.5.3 and Sec. 6.5.4). Both implemented warning systems do not directly influence the drivers' vehicle control but aim at reducing long glances off the road which finally can improve vehicle control performance. However, in the user test durations of glances off the road were not significantly reduced (see Sec. 6.5.5). Consequently, no improved lane keeping performance can be expected. The same observation was previously made in the evaluation in [126]. Here, lane keeping performance as measured by the STD was also not improved and no effects of the warning system on glance durations could be established. Similar results on driving performance were also obtained in the simulator studies [51, 52]. In those experiments headway keeping and steering performance did not benefit from advisory feedback to return gaze to road. The authors hypothesized that this was possible due to the comparatively moderate difficulty of the driving task. This is also a possible explanation for this user test as the driving task was not very demanding (see Sec. 6.4.3) which was also noted by some participants.

In contrast to the other metrics employed to asses lane keeping behavior, RMSE of the steering angle velocity was significantly reduced by treatment with the distraction warning system (see Tab. 6.13). This may be explainable by a calming effect on the drivers resulting them to perform less abrupt

steering movements. In [51, 52] the dynamics of steering behavior were not affected by feedback but as a different metric (steering entropy) was employed, results are not fully comparable.

As shown in [51] automatically locking secondary tasks in demanding driving situation significantly improves driving performances. In contrast to a distraction warning system this approach significantly reduced the duration of glances off the road (see the discussion in Sec. 6.6.3) which might in turn resulted in significant improvements of driving performance.

In [28] a lane keeping assistance system was adapted to the driver's attention state. This was implemented by intervening earlier when the driver engaged in a secondary task. Here, improved lane position could be obtained under similar user acceptance. In contrast to our study, in that work the maximum deviation from the lane center was used as a metric which is not a standard in driver distraction research. The glance policy computed in AGB aims at reducing the expected squared deviation from the lane center (see Sec. 3.3.1). This specific choice of objective results in a gaze switch policy which is independent of the current lane position (see Sec. 3.6.1). Consequently, comparison with [28] suggests that AGB may be more effective in improving driving performance using a different objective which results in a gaze switch policy adapted to the current lane position.

6.6.4 Effects on Glance Behavior

Distraction warning systems aim at reducing the duration of long glances off the road [52, 5, 126]. Hence research question R4 and related hypotheses H6.1-4 addressed potential effects of treatment with both variants of the warning system on glance behavior.

The analysis of the duration of glances off the road of the participants shows no significant effects of the treatment with the warning system (see Sec. 6.5.5). Instead drivers seem to largely maintain the glance strategies shown in secondary task engagement without warning system. However, the warning systems, especially EOR, tend to decrease variation of glance durations over driving speeds (see Fig. 6.24) which was significant considering the 0.75 quantile of glance behavior. This indicates that the static threshold on eyes-off duration employed in EOR reduces the drivers' adaption of glance behavior to driving speed which is problematic. Typically distraction warning systems escalate if the driver does not return his or her gaze to the road [52, 5]. To facilitate the subjective assessment of warning timing this user test employed only a single warning stage (see Sec. 6.3.6). Hence, lack of effects of the distraction warning systems on glance behavior could be explained by the drivers not feeling sufficiently urged to return gaze.

[52, 119] found positive effects on glance behavior of distraction warning systems. However, in both works only the proportion of time the drivers looked at the road was significantly reduced. In contrast, similar as in our work and in the study of [126] durations of off-road glances were not significantly reduced. In this context, it was hypothesized that adapting glance behavior might require longer treatment with the warning system. In the extended field evaluation of [5], no significant effects of distraction warning on glance behavior could be established but a trend towards reduced off-road glance durations and fewer triggered warnings was observed. We wish to note that the timing of warnings was rated significantly better in the present experiments for AGB. That is drivers generally welcomed warnings similarly adapted to the driving situation as their own glance strategies. Consequently, lack of effects on glance behavior observed in previous works [52, 5, 126] may also be explained by the fact that the employed attention assessment neglected the context of the driving situation. [119] employed a warning index adapted to the vehicles driving speed to implement a distraction warning systems. However in that work the benefits were possibly weakened by the reported issues with insufficient eye-tracking quality.

As mentioned earlier in Sec. 6.6.3, blocking engagement in the secondary task in driving situations of high demand decreased the duration of off-road glances in [52]. In [119] a similar blocking approach was used in a part of the user test. This blocking feature showed a similar effect of shortening glances off the road. In contrast to a distraction warning system blocking secondary task engagement removes the reason for the driver to avert gaze from the road. Hence, it must be expected that this cause most of the driver return and keep their gaze on the road. Consequently, the proportion as well as the duration of off-road glance is reduced.

6.7 Conclusion

The present chapter developed and evaluated a distraction warning systems based on computing appropriate glance behavior previously defined in Sec. 3.4. Online pre-processing of eye-tracking and CAN-data as well as online policy computation in the joint task POMDP was implemented in a test vehicle to trigger a visual-auditiv warning. The resulting warning system was evaluated in comparison with a state-of-the-art distraction warning system by means of a user test on a test track. Here, sufficient speed of computation as well as robustness wrt. sensor noises were demonstrated. The user test revealed that the developed warning system was significantly better received by the participants than the state-of-the art system. Detailed analysis showed that this was due to the fact that the state-of-the-art system neglected the specific driving situation which was incorporated in the other system. However, both evaluated warning systems failed to significantly alter the drivers' glance behavior and consequently no significantly improved lane keeping performance except for less abrupt steering was found.

The evaluation gives an encouraging proof of concept of our framework of appropriate glance behavior. That is, it shows that the joint task of driving and secondary task engagement can be modeled that robust and real-time computation of gaze-switch policies is possible. Furthermore the study demonstrates that these policies also result in improved distraction warning. However, the discussion of the results revealed several aspects that need further investigation: It needs to be investigated if the lack of effects of the warning systems on glance behavior can be mitigated by a more elaborated, e.g. escalating, warning design or increased difficulty of the joint task.

In second step it can be necessary to increase the complexity of the representation of the task of lane keeping, e.g. with respect to non-quadratic task objectives, used in the normative model of appropriate glance behavior. The CPU-time required for policy computation in the joint task POMDP was well below 0.04 s. Hence, in increased sample time it could be possible to locally solve more complex POMDP models by iterative approximation with the joint task model analogously to [56, 244].

The present user study was conducted on a test track to realize the experimental design. While this is sufficient for a first proof of concept ultimately the distraction warning system must be evaluated in the ecological context as done in [5]. However, to do so all typical driving situations must be modeled in an extended joint task POMDP.

7 Conclusion and Outlook

Engagement into a visually demanding secondary task during manual driving requires situationally adapted glance behavior. Research has shown that experienced drivers are capable of applying such strategies. In contrast, the state-of-the-art algorithmic approaches assess driver attention without the situational context. The scope of this thesis has been to establish new and alternative techniques for assessment of driver attention from both glance behavior and the characteristics of the driving situation. These can help to improve driver distraction warning systems with respect to both effectiveness and user acceptance.

The present chapter summarizes the main contribution of this thesis. Furthermore, we discuss both the potential and the limitations of the pursued methodology which are used to derive directions for future research.

7.1 Conclusion

This work has addressed the development and evaluation of a normative model of glance behavior for interaction with a visually demanding secondary task while driving. This served the purpose of enabling assessment of driver attention in the context of the current driving situation. To this end, a decision theoretic model of the joint task of engagement in a secondary task in lane keeping has been developed. Approaches for computing optimal and rational policies that defined the normative model have been derived. In addition to that, we have developed new techniques for estimating important model parameters. Finally, the normative model of glance behavior was integrated into a distraction warning system in a test vehicle and evaluated with respect to effectiveness and user acceptance. In the following we will summarize the main results obtained throughout this course.

Appropriate Glance Behavior in the Joint Task of Driving and Secondary Task Interaction We considered the problem of developing a normative model of glance behavior feasible for application in a real-time distraction warning system in Cpt. 3. In this context a Partially Observable Markov Decision Process model (POMDP) of the joint task comprising of a linear-affine kinematic model of the lane keeping task, a linear-Gaussian model of the driver's sensory characteristics and models of the visual demanding secondary task were developed (Sec. 3.3). Using this POMDP model allowed the first mathematical definition of situationally appropriate glance behavior in the considered driving scenario (Sec. 3.4). In contrast to the Kircher's and Ahlström's recently proposed concept of minimum required attention, which shares many ideas, our approach can directly algorithmically be realized in a real-time warning system. Furthermore, we derived new exact algorithms to obtain optimal and rational policies in the POMDP which are required in the definition of appropriate glance behavior (Sec. 3.5). Finally, two variants, SRopt (Algo. 4) and STRopt (Algo. 6), of the joint task POMDP were evaluated with respect to model realism and computational feasibility of policy computation (Sec. 3.6). The results showed that the assumptions underlying the considered POMDP models did not result in policies that conflict with empirical data. However, the computational demands required for policy computation turned out to be feasible for online-application only for SRopt. In this context, the computational complexity of the solution approaches was analyzed which allowed to identify the main computational bottlenecks which must be addressed in future research.

Inferring Drivers' Policy and Reward Cpt. 4 addressed the problem of finding a realistic model of the driver's vehicle control and secondary task interaction policy. Furthermore, we considered estimating a suitable parameterization of the objective, i.e. the reward function, of the joint task POMDP from behavioral data of experienced drivers. These quantities are required to obtain a valid definition of appropriate glance behavior. For this purpose, new exact inverse optimal control techniques were derived for the class of POMDPs the joint task model belongs to (Sec. 4.4). We first demonstrated the

potential of the inverse optimal control approach as well as the differences between different variants thereof using simulated data (Sec 4.5). Thereafter, a real-traffic driving experiment of lane keeping in presence of a visually demanding secondary task was introduced (Sec 4.6). This provided data of adaptive driver behavior suitable for evaluation. Using the experimental data we validated both the model assumptions as well as the techniques for inference of reward parameters by comparing prediction errors with those obtained by the established two-point steering model [197] and the barrier model [95]. The results showed that the developed methodology improves prediction of driver behavior especially in driving situations unseen at estimation time (Sec 4.7).

Inferring Drivers' Sensor Characteristics A driver's glance behavior is deeply rooted in the characteristics of the driver's sensing of the road scenery. Obtaining models of the sensor characteristics underlying human real-world motor behavior is very challenging. We proposed the first general framework for inference of sensor model parameter in sequential decision making and evaluated its implementation for the joint task POMDP in Cpt. 5. First, inverse optimal control in POMDPs was extended to the framework of I See What You See (ISWYS) which allows to estimate both reward and sensor model parameters (Sec. 5.4). Thereafter, we derived an exact implementation of ISWYS for the POMDP class of the joint task model. A second driving experiment on lane keeping while engaging in a secondary task was presented (Sec. 5.6). In this study, drivers' vision of the forward road scenery was experimentally manipulated by imposing gaze aversion to a display mounted at different locations in the vehicle interior. In the evaluation on the obtained data, sensor model inference using ISWYS resulted in improved prediction of glance behavior compared to inverse optimal control (Sec. 5.7). However, no benefits for predicting states related to vehicle control could be established. In addition to that, the inferred sensor noise parameters were large but lead to a comparably small uncertainty in the driver's estimate of the states related to vehicle control. This indicated that drivers use a simpler representation of the driving situation than assumed in the joint task model.

Distraction Mitigation by Computation of Appropriate Glance Behavior and Its Evaluation The goal of our efforts in modeling and algorithm development was improvement of distraction warning systems. In pursuit of this objective a distraction warning system based on computing appropriate glance behavior was developed (Sec. 6.3). The benefits of warnings adapted to the driving situation provided by the new system were evaluated in a user study on a test track (Sec. 6.4). Computing appropriate glance behavior was compared to a state-of-the art approach for attention assessment in terms of user acceptance and improvement of lane keeping performance. The results of the user test showed that the number and timing of warnings of computing appropriate glance behavior were significantly better received by the users (Sec. 6.5). This could directly be attributed to adapting warnings to the driving situation. In contrast, neither of the evaluated warning systems had a significant impact on the drivers' glance behavior. As a consequence, the distraction warning system did also not improve lane keeping performance except for less abrupt steering. It was hypothesized that this effect was due to the design of the warnings, which were probably presented not urgently enough.

7.2 Potential and Limitations of the Research Methodology

The present thesis focused on driver attention assessment in the driving task of lane keeping. The problem of determining whether the driver pays sufficient attention to the driving task in every possible driving situation is a very challenging issue. The reason is that this task requires to consider difficult decision making problems incorporating the driver's uncertainty of the states of the driving situation as well as aspects of gaze switching. This is indicated by the moderately complex POMDP model of lane keeping employed in the present work (see Sec. 3.3.4).

Our work on obtaining a realistic yet computationally feasible model of appropriate glance behavior shows which trade-off decisions must be made to obtain an approach that can be implemented in a distraction warning system (Cpt. 3). The results and insights obtained in this course can guide the development of techniques for driver attention assessment in other driving situations such as e.g. headway-keeping. Furthermore, we have shown that the important model parameters required to realize the normative model of glance behavior can successfully be inferred by techniques based on

inverse optimal control (Cpt. 4 and Cpt. 5). In this context, new inference procedures were contributed to the state-of-the-art in the machine learning research. These approaches may also be applied to parameterize models of glance behavior in other driving situations. Finally, the evaluation of the warning system gave a first and clear demonstration of the weaknesses of the current state-of-the art in driver's attention assessment in the limited scenario of lane keeping (Cpt. 6). This shows that previously proposed approaches are insufficient and that a promising direction for advancement of distraction warning systems is given by the methodology established in this thesis.

Despite the results and insights gained by the present work, it has several limitations. Most important the developed models and technique address only the scenario of lane keeping. Considering this particular driving task came with several aspects that facilitated development of a normative model of glance behavior. Importantly, the problem of lane keeping could realistically and naturally be modeled by linear-affine Gaussian dynamics combined by a quadratic objective function (Sec. 3.3.1). This model structure greatly facilitated computing rational and optimal policies by allowing convenient policy factorization (Sec. 3.5.2 and Sec. 3.5.3). Similar factorization properties could also be exploited for efficient inference of reward and sensor model parameters (Sec. 4.4 and Sec. 5.5). However, the same structure is not appropriate for modeling other driving tasks. For example the driving task of headway-keeping requires a nonlinear non-quadratic objective function. This is because evaluation of the distance to a preceding vehicle requires a monotonic function that assigns high costs to small distances and low costs to high distances. Hence, additional techniques are required to obtain rational and optimal glance policies in other driving tasks.

Even though we were able to compute gaze switch policies sufficiently quickly for application in a warning system, this came with additional model assumptions that are not entirely realistic (Sec. 3.5.2). Specifically, we assumed instantaneous saturation of information once the driver returned gaze to the road. In this thesis also algorithms for policy computation (Sec. 3.5.2) and parameter inference (Sec. 4.4.3 and Sec. 5.5.2) without this particular assumption were developed. However, this thesis could not obtain sufficiently efficient computational procedures. Consequently, the approaches turned out to be too computationally demanding for both online-application and evaluation on large amount of driving data obtained in the real-traffic experiments (Sec. 3.6.2).

The present work made the assumption that drivers are capable of optimal Bayesian inference of the states of the driving situation when averting gaze. That is, drivers possess perfect models of the vehicle and situation dynamics. Although the employed kinematic model of the task of lane keeping is comparably simple it turned out that drivers likely employ less accurate internal models (Sec. 5.7.4). In more complex driving situations it must be expected that the deviation between the true situation dynamics and the internal models is even more pronounced.

Finally, we investigated only a comparably simple secondary task in our driving experiments (Sec. 4.6 and Sec. 5.6). Furthermore, this task was only included in the joint task model in a minimal version (Sec. 4.7.1) although the proposed joint task model allows for more detailed representations. From previous research it is known that the characteristics of a secondary task, such as the costs of task interruption have a strong impact on interaction and glance behavior. The potential of the joint task model with respect to such aspects has not yet been evaluated.

7.3 Outlook

In the present work the problem of determining appropriate glance behavior in the driving scenario of lane keeping was addressed. Following a summary of the main results the previous section has highlighted the potential and the limitations of the pursued methodology. Based on the findings, the most important directions for future research are identified.

Clearly, future research should address the extension of the developed approaches to further driving tasks. Besides lane keeping, headway-keeping is the most fundamental driving task. Previous research has shown that head-way keeping is significantly impaired in presence of a visually demanding secondary task e.g. [48, 123]. Hence, the driving task of headway-keeping should be investigated next. In this driving task a challenge arises from the task objective that cannot be modeled by a quadratic function as explained earlier. Furthermore, in this scenario appropriate gaze switch and vehicle control policies also strongly depend on the knowledge/belief about the preceding vehicles future behavior,

e.g. the likelihood of abrupt stopping. Detection accuracy of such intentions based on sensor characteristics has been studied in [221] which may provide a basis for modeling this type of "situational awareness" in a partially observable Markov decision process. However, obtaining an optimal or rational gaze switch policy in a realistic model of headway keeping in presence of a visually demanding secondary task will in the end also require new solution techniques.

In the present work the first technique for inference of sensor models underlying dynamic behavior was developed. This allowed to study driver perception in gaze switching for lane keeping. It was revealed that the drivers' internal model probably deviate from the true dynamics of the task. In real-traffic driving various source of disturbances are present, hence the observations made in the driving experiment should further be studied in the laboratory. Human and primate internal representations have been previously studied using mathematical model of decision making [18, 70]. However the methodology developed in this thesis is the first can potentially be used to study internal models underlying gaze switching behavior where these are assumed to be of particular importance [255]. Hence, the proposed estimation approach may help to gain a better understanding in the internal representations that guide human daily actions beyond manual vehicle control.

Visual driver distraction has the most distinct effects on driving performance. However, also cognitive distraction can negatively affect vehicle control, for example by resulting in delayed hazard response [139]. The present work has focused exclusively on drivers' visual attention but cognitive distraction is also related to glance behavior and visual perception. For example, it has been shown that cognitive distraction comes with concentration of glances to the forward road scenery [245] and impaired processing of visual information [227, 83]. These observations may be explained by effects of cognitive distraction on internal models used in driver's visual perception. For example, an impaired internal model of the driving task would require more frequent and longer glances at the forward road scenery. Studying mathematical internal models of drivers as suggested in the previous paragraph could possibly also contribute to understanding cognitive distraction.

Finally, the role of the driver is slowly changing through the advent of automatic driving functionality. Due to safety reasons current series and pre-series partial automated systems for example the Tesla Autopilot (Tesla Inc, Palo Alto, California) require the driver to continuously monitor the system's functionality and to intervene in case of failure. That is, the driver is no longer required to control the vehicle but must still maintain sufficient attention to detect and correct system errors. Consequently, the issue of appropriate glance behavior remains relevant in the context of partially automated driving. Similar as in case of manual driving, appropriate monitoring behavior depends on the driving situation. Specifically, glance behavior must consider the situation specific likelihood of system failure. For example, the quality of detecting lane boundaries and failure of lane tracking will depend on lane visibility and lane curvature. A normative model of appropriate glance behavior for monitoring of a partially automated driving system could possibly obtained in the following way: The task of failure detection could be modeled by the approach proposed in [65]. Given a detected failure, the driver has to override the partial automated system and take over control. Such takeover scenarios in automated driving have been analyzed using control theoretic techniques in [102]. Combining the model of failure detection and control mode change could finally result in a holistic model for appropriate monitoring behavior that considers the aspects of both system failure and takeover.

List of Publications

- M. Herman, T. Gindele, J. Wagner, F. Schmitt, and W. Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.
- F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Predicting lane keeping behavior of visually distracted drivers using inverse suboptimal control. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2016.
- F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Exact maximum entropy inverse optimal control for modeling human attention scheduling and control.In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.
- M. Herman, T. Gindele, J. Wagner, F. Schmitt, and W. Burgard. Simultaneous estimation of rewards and dynamics from noisy expert demonstrations. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2016.
- M. Herman, T. Gindele, J. Wagner, F. Schmitt, C. Quignon, and W. Burgard. Learning high-level navigation strategies via inverse reinforcement learning: A comparative analysis. In *Proceedings of the Australian Joint Conference of Artificial Intelligence (AI)*, 2016.
- F. Schmitt, H.-J. Bieg, M. Herman, and C. Rothkopf. I see what you see: Inferring sensor and policy models of human real-world motor behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.
- F. Schmitt, H.-J. Bieg, D. Manstetten, and R. Stiefelhagen. Distraction mitigation by computation of appropriate glance behavior and its evaluation in a user test. *(manuscript in preparation) IEEE Transaction on Intelligent Transport Systems*, 2017.

A Appendix

A.1 Proof of Kalman Belief Update

In the the section 2.1.4, we derived the belief-MDP of linear quadratic Gaussian Problems. Here, it was claimed that

$$p_t(\boldsymbol{\mu}_{t+1}^x|\boldsymbol{\mu}_t^x,\boldsymbol{\Sigma}_t^x,\mathbf{u}_t) = \mathcal{N}(\boldsymbol{\mu}_{t+1}^x|\mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t, \mathbf{A}_t\boldsymbol{\Sigma}_t^x\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^x(\boldsymbol{\Sigma}_t^x)).$$

In this section we will derive this expression. Therefore, first note that $p(\mathbf{x}_t|\mathbf{z}_{-\infty:t}) = \mathcal{N}(\boldsymbol{\mu}_t^x,\boldsymbol{\Sigma}_t^x)$. Second, we obtain $\boldsymbol{\mu}_{t+1}^x$ as

$$\begin{aligned}\boldsymbol{\mu}_{t+1}^x &= \bar{\boldsymbol{\mu}}_{t+1}^x - \mathbf{K}_{t+1}(\mathbf{H}\bar{\boldsymbol{\mu}}_{t+1}^x - \mathbf{z}_{t+1}) = \mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t - \mathbf{K}_{t+1}(\mathbf{H}[\mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t] - \mathbf{z}_{t+1}) &\text{(A.1)}\\ &= \mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t - \mathbf{K}_{t+1}(\mathbf{H}[\mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t] - [\mathbf{H}(\mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t + \boldsymbol{\epsilon}_t^x) + \boldsymbol{\epsilon}_{t+1}^z]) &\text{(A.2)}\end{aligned}$$

As $\mathbf{H}[\mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t] - [\mathbf{H}(\mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t + \boldsymbol{\epsilon}_t^x) + \boldsymbol{\epsilon}_{t+1}^z]$ has zero expectation, it holds

$$\mathbb{E}[\boldsymbol{\mu}_{t+1}^x] = \mathbf{A}_t\boldsymbol{\mu}_t^x + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t. \tag{A.3}$$

For obtaining the covariance of $\boldsymbol{\mu}_{t+1}^x$ given $\boldsymbol{\mu}_t^x,\boldsymbol{\Sigma}_t^x,\mathbf{u}_t$ we only need to consider the random elements in (A.2). That is, we need to obtain the covariance of

$$\mathbf{K}_{t+1}[\mathbf{H}(\mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t + \boldsymbol{\epsilon}_t^x) + \boldsymbol{\epsilon}_{t+1}^z] \tag{A.4}$$

given $\boldsymbol{\mu}_t^x,\boldsymbol{\Sigma}_t^x,\mathbf{u}_t$. Plugging $p(\mathbf{x}_t|\mathbf{z}_{-\infty:t}) = \mathcal{N}(\boldsymbol{\mu}_t^x,\boldsymbol{\Sigma}_t^x)$ into the term and considering the that $\mathbf{x}_t,\boldsymbol{\epsilon}_t^x,\boldsymbol{\epsilon}_{t+1}^z$ independent of all other variables it holds

$$\begin{aligned}\mathbb{V}[\mathbf{K}_{t+1}(\mathbf{H}(\mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \mathbf{a}_t + \boldsymbol{\epsilon}_t^x) + \boldsymbol{\epsilon}_{t+1}^z)] &= \mathbf{K}_{t+1}(\mathbf{H}(\mathbf{A}_t\boldsymbol{\Sigma}_t^x\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x})\mathbf{H}^\top + \boldsymbol{\Sigma}^{\epsilon^z})\mathbf{K}_{t+1}^\top &\text{(A.5)}\\ &= \mathbf{K}_{t+1}\mathbf{H}\bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top\mathbf{K}_{t+1}^\top + \mathbf{K}_{t+1}\boldsymbol{\Sigma}^{\epsilon^z}\mathbf{K}_{t+1}^\top &\text{(A.6)}\end{aligned}$$

Note, that the a-posterior covariance $\boldsymbol{\Sigma}_{t+1}^x$ is given according to

$$\boldsymbol{\Sigma}_{t+1}^x = (\mathbf{I}^{n_x} - \mathbf{K}_{t+1}\mathbf{H})\bar{\boldsymbol{\Sigma}}_{t+1}^x = (\mathbf{I}^{n_x} - \mathbf{K}_{t+1}\mathbf{H})\bar{\boldsymbol{\Sigma}}_{t+1}^x(\mathbf{I}^{n_x} - \mathbf{K}_{t+1}\mathbf{H})^\top + \mathbf{K}_{t+1}\boldsymbol{\Sigma}^{\epsilon^z}\mathbf{K}_{t+1}^\top \quad \text{(Joseph form).} \tag{A.7}$$

As the difference of the Joseph form and the classic form for the a-posterior covariance $\boldsymbol{\Sigma}_{t+1}^x$ is the zero matrix, it holds true

$$\mathbf{K}_{t+1}\mathbf{H}\bar{\boldsymbol{\Sigma}}_{t+1}^x\mathbf{H}^\top\mathbf{K}_{t+1}^\top + \mathbf{K}_{t+1}\boldsymbol{\Sigma}^{\epsilon^z}\mathbf{K}_{t+1}^\top = \mathbf{K}_{t+1}\mathbf{H}\bar{\boldsymbol{\Sigma}}_{t+1}^x = \bar{\boldsymbol{\Sigma}}_{t+1}^x - \boldsymbol{\Sigma}_{t+1}^x, \tag{A.8}$$

what proves the claim.

A.2 Proof of Reward Gradient Recursion of Joint Task Model

In Sec. 4.4.3, it was stated that the gradients of the optimal state-control function and the soft state-control wrt. to the parameters of the primary task features are given by

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\big(&\boldsymbol{\mathfrak{M}}_t^{Q^{\star,\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) + \boldsymbol{\mathfrak{M}}_t^{Q^{\star,\boldsymbol{\theta}},2}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] \\ &+ \mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})\big),\end{aligned}$$

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\big(&\boldsymbol{\mathfrak{M}}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top) + \boldsymbol{\mathfrak{M}}_t^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] \\ &+ \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}})\big),\end{aligned}$$

where $\boldsymbol{\mathfrak{M}}_t^{Q^{\star,\boldsymbol{\theta}},1}, \boldsymbol{\mathfrak{M}}_t^{\tilde{Q}^{\boldsymbol{\theta}},1} \in \mathbb{R}^{(n_x+n_u)^2\times(n_x+n_u)^2}$, $\boldsymbol{\mathfrak{M}}_t^{Q^{\star,\boldsymbol{\theta}},2}, \boldsymbol{\mathfrak{M}}_t^{\tilde{Q}^{\boldsymbol{\theta}},2} \in \mathbb{R}^{(n_x+n_u)^2\times(n_x+n_u)}$ and $\mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}}, \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} \in \mathbb{R}^{(n_x+n_u)^2}$.
While this factorization is clear for the case of time step $t = T$:

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} Q_t^{\star,\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) &= \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \\ &= \nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\Big([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top \operatorname{blk}(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \operatorname{tr}(\boldsymbol{\Theta}_1\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T})) + \theta_3 u_T^{\mathrm{z}} + \boldsymbol{\theta}_4^\top\boldsymbol{\varphi}(x_T^{\mathrm{i}}, u_T^{\mathrm{i}})\Big) \\ &= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\mathbf{I}^{(n_x+n_u)^2} \operatorname{vec}([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}][\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top) + \mathbf{0}[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \operatorname{vec}\big(\operatorname{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}), \mathbf{0})\big)\Big) \\ &= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\boldsymbol{\mathfrak{M}}_T^{Q^{\star,\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}][\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top) + \boldsymbol{\mathfrak{M}}_T^{Q^{\star,\boldsymbol{\theta}},2}[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \mathbf{m}_t^{Q^{\star,\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}})\Big) \\ &= \mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\boldsymbol{\mathfrak{M}}_T^{\tilde{Q}^{\boldsymbol{\theta}},1} \operatorname{vec}([\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}][\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}]^\top) + \boldsymbol{\mathfrak{M}}_T^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_T^{\mathrm{p}}; u_T^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:T}, x_T^{\mathrm{i}}, u_T^{\mathrm{z}}, u_T^{\mathrm{i}})\Big),\end{aligned}$$

it still needs to be shown for the remaining time steps. The derivation will exemplary be conducted for the maximum causal entropy model. The optimal policy model can be treated analogously.

For the remaining time steps the relation of the soft state-control function gradient

$$\nabla_{\boldsymbol{\theta}} \tilde{Q}_t^{\boldsymbol{\theta}}(x_t, u_t) = \boldsymbol{\varphi}(x_t, u_t) + \mathbb{E}\big[\nabla_{\boldsymbol{\theta}} \tilde{Q}_t^{\boldsymbol{\theta}}(x_{t+1}, u_{t+1}) \big| \tilde{\pi}_t^{\boldsymbol{\theta}}(u_{t+1}|x_{t+1}), \mathcal{P}(x_{t+1}|x_t, u_t)\big]$$

will be employed.

The soft state-control function is given by

$$\tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}}, \boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) = [\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}]^\top \mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}}[\boldsymbol{\mu}_t^{\mathrm{p}}; u_t^{\mathrm{p}}] + m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) \tag{A.9}$$

where the involved terms are defined as

$$\begin{aligned}\mathbf{M}_t^{\tilde{Q}^{\boldsymbol{\theta}}} &= [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}[\mathbf{A}_t\ \mathbf{B}_t] + \operatorname{blk}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \\ \mathbf{m}_t^{\tilde{Q}^{\boldsymbol{\theta}}} &= 2[\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}\mathbf{a}_t + [\mathbf{A}_t\ \mathbf{B}_t]^\top \mathbf{m}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} \\ m_t^{\tilde{Q}^{\boldsymbol{\theta}},1}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t}, x_t^{\mathrm{i}}, u_t^{\mathrm{p}}, u_t^{\mathrm{z}}, u_t^{\mathrm{i}}) &= \mathbf{a}_t^\top \mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}\mathbf{a}_t + 2\mathbf{a}_t^\top \mathbf{m}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}} + \operatorname{tr}(\boldsymbol{\Theta}_1\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})) \\ &\quad + \operatorname{tr}\Big(\mathbf{M}_{t+1}^{\tilde{V}^{\boldsymbol{\theta}}}(\mathbf{A}_t\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\mathbf{z}}{}_{0:t})\mathbf{A}_t^\top + \boldsymbol{\Sigma}^{\epsilon^x} - \boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\ x_t^{\mathrm{z}}] \oplus u_t^{\mathrm{z}}]))\Big) \\ &\quad + \theta_3 u_t^{\mathrm{z}} + \boldsymbol{\theta}_4^\top\boldsymbol{\varphi}(x_t^{\mathrm{i}}, u_t^{\mathrm{i}}) \\ &\quad + \mathbb{E}\Big[m_{t+1}^{\tilde{V}^{\boldsymbol{\theta}},1}([\boldsymbol{x}^{\mathbf{z}}{}_{0:t}\ x_t^{\mathrm{z}}] \oplus u_t^{\mathrm{z}}], x_{t+1}^{\mathrm{i}}) \Big| \mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^{\mathrm{z}}, u_t^{\mathrm{z}}; x_t^{\mathrm{i}}, u_t^{\mathrm{i}})\Big].\end{aligned}$$

Consequently, the gradient of the soft state-control function wrt. the parameters of the reward of primary task result in

$$\begin{aligned}
&\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t},x_t^{\mathrm{i}},u_t^{\mathrm{p}},u_t^{\mathrm{z}},u_t^{\mathrm{i}})\\
&\quad=\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\Big([\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}]^\top\,\mathrm{blk}(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)[\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}]+\mathrm{tr}(\boldsymbol{\Theta}_1\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}))+\theta_3u_t^{\mathrm{z}}+\boldsymbol{\theta}_4^\top\boldsymbol{\varphi}(x_t^{\mathrm{i}},u_t^{\mathrm{i}})\Big)\\
&\qquad+\mathbb{E}\Big[\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\tilde{Q}_{t+1}^{\boldsymbol{\theta}}(\boldsymbol{\mu}_{t+1}^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t+1},x_{t+1}^{\mathrm{i}},u_{t+1}^{\mathrm{p}},u_{t+1}^{\mathrm{z}},u_{t+1}^{\mathrm{i}})\\
&\qquad\quad\Big|\tilde{\pi}_{t+1}^{\boldsymbol{\theta}}(u_{t+1}^{\mathrm{p}},u_{t+1}^{\mathrm{z}},u_{t+1}^{\mathrm{i}}|\boldsymbol{\mu}_{t+1}^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t+1},x_{t+1}^{\mathrm{i}}),\mathcal{P}(\boldsymbol{\mu}_{t+1}^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t+1},x_{t+1}^{\mathrm{i}}|\boldsymbol{\mu}_t^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t},x_t^{\mathrm{i}},u_t^{\mathrm{p}},u_t^{\mathrm{z}},u_t^{\mathrm{i}})\Big]
\end{aligned}\tag{A.10}$$

$$\begin{aligned}
&=\mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big(\mathbf{I}^{(n_x+n_u)^2}\,\mathrm{vec}([\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}]^\top)+\mathbf{0}[\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}]+\mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}),\mathbf{0})\right)\Big)\\
&\quad+\mathbb{E}\Big[\mathbf{P}_{n_x,n_u}^{\mathrm{blk}}(\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\,\mathrm{vec}([\boldsymbol{\mu}_{t+1}^{\mathrm{p}};u_{t+1}^{\mathrm{p}}][\boldsymbol{\mu}_{t+1}^{\mathrm{p}};u_{t+1}^{\mathrm{p}}]^\top)+\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},2}[\boldsymbol{\mu}_{t+1}^{\mathrm{p}};u_{t+1}^{\mathrm{p}}]\\
&\qquad\qquad+\mathfrak{m}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}}}(\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t+1},x_{t+1}^{\mathrm{i}},u_{t+1}^{\mathrm{z}},u_{t+1}^{\mathrm{i}}))\\
&\qquad\Big|\tilde{\pi}_{t+1}^{\boldsymbol{\theta}}(u_{t+1}^{\mathrm{p}},u_{t+1}^{\mathrm{z}},u_{t+1}^{\mathrm{i}}|\boldsymbol{\mu}_{t+1}^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t+1},x_{t+1}^{\mathrm{i}}),\mathcal{P}(\boldsymbol{\mu}_{t+1}^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t+1},x_{t+1}^{\mathrm{i}}|\boldsymbol{\mu}_t^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t},x_t^{\mathrm{i}},u_t^{\mathrm{p}},u_t^{\mathrm{z}},u_t^{\mathrm{i}})\Big]
\end{aligned}\tag{A.11}$$

Next we define the following quantity

$$\boldsymbol{S}_{t+1}=\mathbf{A}_t\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t})\mathbf{A}_t^\top+\boldsymbol{\Sigma}^{\epsilon^x}-\boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}\,x_t^{\mathrm{z}}\oplus u_t^{\mathrm{z}}])$$

This allows to express the distribution of $\boldsymbol{\mu}_{t+1}^{\mathrm{p}},u_{t+1}^{\mathrm{p}}$ given $\boldsymbol{\mu}_t^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t},x_t^{\mathrm{i}},u_t^{\mathrm{p}},u_t^{\mathrm{z}},u_t^{\mathrm{i}}$ as

$$\begin{bmatrix}\boldsymbol{\mu}_{t+1}^{\mathrm{p}}\\u_{t+1}^{\mathrm{p}}\end{bmatrix}=\mathcal{N}\Big(\begin{bmatrix}\mathbf{A}_t\boldsymbol{\mu}_t^{\mathrm{p}}+\mathbf{B}_tu_t^{\mathrm{p}}+\mathbf{a}_t\\\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}(\mathbf{A}_t\boldsymbol{\mu}_t^{\mathrm{p}}+\mathbf{B}_tu_t^{\mathrm{p}}+\mathbf{a}_t)+\tilde{\mathbf{f}}_{t+1}^{\boldsymbol{\theta}}\end{bmatrix},\begin{bmatrix}\boldsymbol{S}_{t+1}&\boldsymbol{S}_{t+1}\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}\,\top}\\\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\boldsymbol{S}_{t+1}&\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\boldsymbol{S}_{t+1}\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}\,\top}+\boldsymbol{\Sigma}^{u_t^{\mathrm{p}},\boldsymbol{\theta}}\end{bmatrix}\Big).\tag{A.12}$$

The definitions

$$\tilde{\boldsymbol{\mathfrak{F}}}_t:=\begin{bmatrix}\mathbf{I}^{n_x}\\\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\end{bmatrix},\quad\tilde{\boldsymbol{\mathfrak{T}}}_t:=\begin{bmatrix}\mathbf{A}_t&\mathbf{B}_t\\\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\mathbf{A}_t&\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\mathbf{B}_t\end{bmatrix},\quad\tilde{\mathbf{t}}_t:=\begin{bmatrix}\mathbf{a}_t\\\tilde{\mathbf{F}}_{t+1}^{\boldsymbol{\theta}}\mathbf{a}_t+\tilde{\mathbf{f}}_{t+1}^{\boldsymbol{\theta}}\end{bmatrix},$$

can be used to write (A.12) in the simpler form of

$$\begin{bmatrix}\boldsymbol{\mu}_{t+1}^{\mathrm{p}}\\u_{t+1}^{\mathrm{p}}\end{bmatrix}=\mathcal{N}\Big(\tilde{\boldsymbol{\mathfrak{T}}}_t\begin{bmatrix}\boldsymbol{\mu}_t^{\mathrm{p}}\\u_t^{\mathrm{p}}\end{bmatrix}+\tilde{\mathbf{t}}_t,\,\tilde{\boldsymbol{\mathfrak{T}}}_t\boldsymbol{S}_{t+1}\tilde{\boldsymbol{\mathfrak{T}}}_t{}^\top+\mathrm{blk}(\mathbf{0},\boldsymbol{\Sigma}^{u_t^{\mathrm{p}},\boldsymbol{\theta}})\Big).\tag{A.13}$$

Employing the previous relation (A.13) allows to evaluate the expectation in (A.11). Finally, manipulations from matrix algebra (see e.g. [175]) and reordering result in the desired equation

$$\begin{aligned}
&\nabla_{\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2}\tilde{Q}_t^{\boldsymbol{\theta}}(\boldsymbol{\mu}_t^{\mathrm{p}},\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t},x_t^{\mathrm{i}},u_t^{\mathrm{p}},u_t^{\mathrm{z}},u_t^{\mathrm{i}})=\mathbf{P}_{n_x,n_u}^{\mathrm{blk}}\Big((\mathbf{I}^{(n_x+n_u)^2}+\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\tilde{\boldsymbol{\mathfrak{T}}}_t\otimes\tilde{\boldsymbol{\mathfrak{T}}}_t)\,\mathrm{vec}([\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}][\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}]^\top)\\
&+(\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},2}\tilde{\boldsymbol{\mathfrak{T}}}_t+\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}(\tilde{\boldsymbol{\mathfrak{T}}}_t\otimes\tilde{\mathbf{t}}_t+\tilde{\mathbf{t}}_t\otimes\tilde{\boldsymbol{\mathfrak{T}}}_t))[\boldsymbol{\mu}_t^{\mathrm{p}};u_t^{\mathrm{p}}]\\
&+\mathrm{vec}\left(\mathrm{blk}(\boldsymbol{\Sigma}^{\mathrm{p}}(\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}),\mathbf{0})\right)+\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\,\mathrm{vec}\left(\tilde{\mathbf{t}}_t\tilde{\mathbf{t}}_t{}^\top+\tilde{\boldsymbol{\mathfrak{F}}}_t\big(\mathbf{A}_t\boldsymbol{\Sigma}_t^{\mathrm{p}}(\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t})\mathbf{A}_t^\top+\boldsymbol{\Sigma}^{\epsilon^x}\big)\tilde{\boldsymbol{\mathfrak{F}}}_t{}^\top+\mathrm{blk}(\mathbf{0},\boldsymbol{\Sigma}^{u_t^{\mathrm{p}},\boldsymbol{\theta}})\right)+\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},2}\tilde{\mathbf{t}}_t\\
&+\mathbb{E}\big[\mathfrak{M}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}},1}\,\mathrm{vec}\left(-\tilde{\boldsymbol{\mathfrak{F}}}_t\boldsymbol{\Sigma}_{t+1}^{\mathrm{p}}([\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}\,x_t^{\mathrm{z}}\oplus u_t^{\mathrm{z}}])\tilde{\boldsymbol{\mathfrak{F}}}_t{}^\top\right)\\
&\qquad+\mathfrak{m}_{t+1}^{\tilde{Q}^{\boldsymbol{\theta}}}([\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}\,x_t^{\mathrm{z}}\oplus u_t^{\mathrm{z}}],x_{t+1}^{\mathrm{i}},u_{t+1}^{\mathrm{z}},u_{t+1}^{\mathrm{i}})|\tilde{\pi}_t(u_{t+1}^{\mathrm{z}},u_{t+1}^{\mathrm{i}}|[\boldsymbol{x}^{\boldsymbol{z}}{}_{0:t}\,x_t^{\mathrm{z}}\oplus u_t^{\mathrm{z}}],x_{t+1}^{\mathrm{i}}),\mathcal{P}^{\mathrm{i}}(x_{t+1}^{\mathrm{i}}|x_t^{\mathrm{z}},u_t^{\mathrm{z}};x_t^{\mathrm{i}},u_t^{\mathrm{i}})]\Big).
\end{aligned}$$

List of Figures

List of Tables

Bibliography

[1] Straßenverkehrsordnung (StVO), Deutschland. `https://www.gesetze-im-internet.de/stvo_2013/index.html`, 03 2013. Entry into force March 2013 (BGBl. I p. 367); Accessed: 2017-07-24.

[2] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. ACM, 2004.

[3] L. Acerbi, W. J. Ma, and S. Vijayakumar. A framework for testing identifiability of bayesian models of perception. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1026–1034, 2014.

[4] L. Acerbi, D. M. Wolpert, and S. Vijayakumar. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, 8(11):e1002771, 2012.

[5] C. Ahlström, K. Kircher, and A. Kircher. A gaze-based driver distraction warning system and its effect on visual behavior. *IEEE Transactions on Intelligent Transportation Systems,*, 14(2):965–973, 2013.

[6] C. Ahlström, K. Kircher, A. Rydström, A. Nåbo, S. Almgren, and D. Ricknäs. Effects of visual, cognitive and haptic tasks on driving performance indicators. In *4th International Conference on Applied Human Factors and Ergonomics (AHFE)*, pages 673–682. CRC Press, 2012.

[7] T. P. Alkim, G. Bootsma, and S. P. Hoogendoorn. Field operational test "the assisted driver". In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 1198–1203. IEEE, 2007.

[8] H. Almahasneh, W.-T. Chooi, N. Kamel, and A. S. Malik. Deep in thought while driving: An eeg study on drivers' cognitive distraction. *Transportation Research Part F: Traffic Psychology and Behaviour*, 26:218–226, 2014.

[9] P. Alriksson and A. Rantzer. Sub-optimal sensor scheduling with error bounds. In *Proceedings of the IFAC World Congress*, 2005.

[10] A. Amditis, L. Andreone, K. Pagle, G. Markkula, E. Deregibus, M. R. Rue, F. Bellotti, A. Engelsberg, R. Brouwer, B. Peters, et al. Towards the automotive hmi of the future: overview of the aide-integrated project results. *IEEE Transactions on Intelligent Transportation Systems*, 11(3):567–578, 2010.

[11] B. D. O. Anderson and J. B. Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[12] J. F. Antin, T. A. Dingus, M. C. Hulse, and W. W. Wierwille. An evaluation of the effectiveness and efficiency of an automobile moving-map navigational display. *International Journal of Man-Machine Studies*, 33(5):581–594, 1990.

[13] E. Arroyo, S. Sullivan, and T. Selker. Carcoach: A polite and effective driving choach. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pages 357–362. ACM, 2006.

[14] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[15] K. J. Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

[16] M. Bain and C. Sommut. A framework for behavioural cloning. *Machine Intelligence*, 15:103, 1999.

[17] C. L. Baker and O. J. Braddick. Eccentricity-dependent scaling of the limits for short-range apparent motion perception. *Vision Research*, 25(6):803–812, 1985.

[18] S. Baron and J. E. Berliner. The effects of deviate internal representations in the optimal model of the human operator. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, volume 15, pages 1055–1057. IEEE, 1976.

[19] S. Baron and D. L. Kleinman. The human as an optimal controller and information processor. *IEEE Transactions on Man-Machine Systems*, 10(1):9–17, 1969.

[20] S. Baron, D. L. Kleinman, and W. H. Levison. An optimal control model of human response part II: prediction of human performance in a complex task. *Automatica*, 6(3):371–383, 1970.

[21] C. Basu, Q. Yang, D. Hungerman, M. Singhal, and A. D. Dragan. Do you want your autonomous car to drive like you? In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 417–425. ACM, 2017.

[22] R. Bellman. A Markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957.

[23] B. Belousov, G. Neumann, C. A. Rothkopf, and J. Peters. Catching heuristics are optimal control policies. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[24] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22, 2014.

[25] D. P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995.

[26] G. J. Blaauw. *Car driving as a supervisory control task*. PhD thesis, Delft University of Technology, 1984.

[27] G. J. Blaauw, H. Godthelp, and P. Milgram. Optimal control model applications and field measurements with respect to car driving. *Vehicle System Dynamics*, 13(2):93–111, 1984.

[28] C. Blaschke, F. Breyer, B. Färber, J. Freyer, and R. Limbacher. Driver distraction based lane-keeping assistance. *Transportation research part F: traffic psychology and behaviour*, 12(4):288–299, 2009.

[29] M. Bloem and N. Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 2014.

[30] J. Bortz and N. Döring. *Forschungsmethoden und evaluation*. Springer-Verlag, 2013.

[31] A. Boularias, J. Kober, and J. Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 182–189, 2011.

[32] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[33] R. Broström, J. Engström, A. Agnvall, and G. Markkula. Towards the next generation intelligent driver information system (IDIS): The volvo car interaction manager concept. In *Proceedings of the ITS World Congress*, page 32, 2006.

[34] D. P. Brumby, D. D. Salvucci, and A. Howes. Dialing while driving? A bounded rational analysis of concurrent multi-task behavior. In *Proceedings of the International Conference on Cognitive Modeling*, pages 121–126, 2007.

[35] D. P. Brumby, D. D. Salvucci, and A. Howes. Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1629–1638. ACM, 2009.

[36] R. H. S. Carpenter. *Movements of the Eyes, 2nd Rev*. Pion Limited, 1988.

[37] A. Ceder. Drivers' eye movements as related to attention in simulated traffic flow conditions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 19(6):571–581, 1977.

[38] X. Chen and B. D. Ziebart. Predictive inverse optimal control for linear-quadratic-Gaussian systems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 165–173, 2015.

[39] Y. Chen and X. Ye. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, 2011.

[40] J. Choi and K.-E. Kim. Inverse reinforcement learning in partially observable environments. *The Journal of Machine Learning Research*, 12:691–730, 2011.

[41] F. Cnossen, T. Meijman, and T. Rothengatter. Adaptive strategy changes as a function of task demands: a study of car drivers. *Ergonomics*, 47(2):218–236, 2004.

[42] D. J. Cole. A path-following driver–vehicle model with neuromuscular dynamics, including measured and simulated responses to a step in steering angle overlay. *Vehicle System Dynamics*, 50(4):573–596, 2012.

[43] D. J. Cole, A. J. Pick, and A. M. C. Odhams. Predictive and linear quadratic methods for potential application to modelling driver steering control. *Vehicle System Dynamics*, 44(3):259–284, 2006.

[44] SAE On-Road Automated Vehicle Standards Committee et al. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard J3016*, pages 01–16, 2014.

[45] J. Daunizeau, H. E. M. Den Ouden, M. Pessiglione, S. J. Kiebel, K. J. Friston, and K. E. Stephan. Observing the observer (II): deciding when to decide. *PLoS One*, 5(12):e15555, 2010.

[46] J. Daunizeau, H. E. M. Den Ouden, M. Pessiglione, S. J. Kiebel, K. E. Stephan, and K. J. Friston. Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS One*, 5(12):e15554, 2010.

[47] M. Deisenroth and C. E. Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 465–472, 2011.

[48] T. A. Dingus, M. C. Hulse, J. F. Antin, and W. W. Wierwille. Attentional demand requirements of an automobile moving-map navigation system. *Transportation Research Part A: General*, 23(4):301–315, 1989.

[49] T. A. Dingus, D. V. McGehee, N. Manakkal, S. K. Jahns, C. Carney, and J. M. Hankey. Human factors field evaluation of automotive headway maintenance/collision warning devices. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):216–229, 1997.

[50] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):596–614, 2011.

[51] B. Donmez, L. N. Boyle, and J. D. Lee. The impact of distraction mitigation strategies on driving performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(4):785–804, 2006.

[52] B. Donmez, L. N. Boyle, and J. D. Lee. Safety implications of providing real-time feedback to distracted drivers. *Accident Analysis & Prevention*, 39(3):581–590, 2007.

[53] B. Donmez, L. N. Boyle, and J. D. Lee. Mitigating driver distraction with retrospective and concurrent feedback. *Accident Analysis & Prevention*, 40(2):776–786, 2008.

[54] A. D. Dragan. Robot planning with mathematical models of human state and action. *arXiv preprint arXiv:1705.04226*, 2017.

[55] J. Engström, E. Johansson, and J. Östlund. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):97–120, 2005.

[56] T. Erez and W. D. Smart. A scalable method for solving high-dimensional continuous POMDPs using local approximation. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

[57] T. Erez, J. J. Tramper, W. D. Smart, and S. C. A. M. Gielen. A POMDP model of eye-hand coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2011.

[58] T. Ersal, H. J. A. Fuller, O. Tsimhoni, J. L. Stein, and H. K. Fathy. Model-based analysis and classification of driver distraction under secondary tasks. *IEEE Transactions on Intelligent Transportation Systems*, 11(3):692–701, 2010.

[59] O. Evans, A. Stuhlmüller, and N. D. Goodman. Learning the preferences of ignorant, inconsistent agents. pages 323–329, 2016.

[60] J.-C. Falmagne. *Elements of psychophysical theory*. Number 6. Oxford University Press on Demand, 2002.

[61] J. M. Findlay and I. D. Gilchrist. *Active vision: The psychology of looking and seeing*. Number 37. Oxford University Press, 2003.

[62] L. Fletcher and A. Zelinsky. Driver inattention detection based on eye-gaze road event correlation. *The International Journal of Robotics Research*, 28(6):774–801, 2009.

[63] NHTSA National Center for Statistics and Analysis. Distracted driving 2015 (traffic safety facts research note. report no. dot hs 812 381). Technical report, National Highway Traffic Safety Administration, Washington, DC, 2017.

[64] L. Fridman, J. Lee, B. Reimer, and T. Victor. 'owl' and 'lizard': patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 10(4):308–313, 2016.

[65] E. G. Gai and R. E. Curry. A model of the human observer in failure detection tasks. *IEEE Transactions on Systems, Man, and Cybernetics*, (2):85–94, 1976.

[66] Y. Gao, J. Peters, A. Tsourdos, S. Zhifei, and E. Meng Joo. A survey of inverse reinforcement learning techniques. *International Journal of Intelligent Computing and Cybernetics*, 5(3):293–311, 2012.

[67] H. Godthelp. *Studies on human vehicle control*. PhD thesis, Institute for Perception TNO, 1984.

[68] H. Godthelp. Vehicle control during curve driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 28(2):211–221, 1986.

[69] H. Godthelp, P. Milgram, and G. J. Blaauw. The development of a time-related measure to describe driving strategy. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 26(3):257–268, 1984.

[70] M. Golub, S. Chase, and M. Y. Byron. Learning an internal dynamics model from control demonstration. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 606–614, 2013.

[71] D. A. Gordon. Experimental isolation of the driver's visual input. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 8(2):129–138, 1966.

[72] C. Gote, M. Flad, and S. Hohmann. Driver characterization & driver specific trajectory planning: an inverse optimal control approach. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3014–3021. IEEE, 2014.

[73] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. `http://stanford.edu/~boyd/graph_dcp.html`.

[74] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[75] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(3):478–500, 2010.

[76] M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz. Visual memory and motor planning in a natural task. *Journal of Vision*, 3(1):6–6, 2003.

[77] M. Herman, T. Gindele, J. Wagner, F. Schmitt, and W. Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[78] P. Hermannstädter and B. Yang. Driver distraction assessment using driver modeling. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3693–3698, 2013.

[79] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011.

[80] T. Horberry, J. Anderson, M. A. Regan, T. J. Triggs, and J. Brown. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention*, 38(1):185–191, 2006.

[81] W. J. Horrey and M. F. Lesch. Driver-initiated distractions: Examining strategic adaptation for in-vehicle task initiation. *Accident Analysis & Prevention*, 41(1):115–122, 2009.

[82] W. J. Horrey, M. F. Lesch, and A. Garabet. Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. *Journal of Safety Research*, 40(1):7–12, 2009.

[83] W. J. Horrey and C. D. Wickens. Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(1):196–205, 2006.

[84] W. J. Horrey, C. D. Wickens, and K. P. Consalus. Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied*, 12(2):67, 2006.

[85] D.-A. Huang, A. M. Farahmand, K. M. Kitani, and J. A. Bagnell. Approximate maxent inverse optimal control and its application for mental simulation of human interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 15, page 29th, 2015.

[86] A. D. Hutson. Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *Journal of Applied Statistics*, 26(3):343–353, 1999.

[87] T. Ishida and T. Matsuura. The effect of cellular phone use on driving performance. *IATSS research*, 25(2):6–14, 2001.

[88] J. J. Jain and C. Busso. Assessment of driver's distraction using perceptual evaluations, self assessments and multimodal feature analysis. In *Proceedings of the Biennial workshop on DSP for in-vehicle systems*, 2011.

[89] C. P. Janssen and D. P. Brumby. Strategic adaptation to performance objectives in a dual-task setting. *Cognitive Science*, 34(8):1548–1560, 2010.

[90] C. P. Janssen and D. P. Brumby. Strategic adaptation to task characteristics, incentives, and individual differences in dual-tasking. *PLoS One*, 10(7):e0130009, 2015.

[91] C. P. Janssen, D. P. Brumby, J. Dowell, N. Chater, and A. Howes. Identifying optimum performance trade-offs using a cognitively bounded rational analysis model of discretionary task interleaving. *Topics in Cognitive Science*, 3(1):123–139, 2011.

[92] S. T. Jawaid and S. L. Smith. Submodularity and greedy algorithms in sensor scheduling for linear dynamical systems. *Automatica*, 61:282–288, 2015.

[93] R. N. Jazar. *Vehicle dynamics: Theory and application*. Springer Science & Business Media, 2013.

[94] L. Johnson, B. Sullivan, M. Hayhoe, and D. Ballard. A soft barrier model for predicting human visuomotor behavior in a driving task. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 689–691. Citeseer, 2013.

[95] L. Johnson, B. Sullivan, M. Hayhoe, and D. Ballard. Predicting human visuomotor behaviour in a driving task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1636):20130044, 2014.

[96] N. Kaempchen, B. Schiele, and K. Dietmayer. Situation assessment of an autonomous emergency brake for arbitrary vehicle-to-vehicle collision scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 10(4):678–687, 2009.

[97] D. Kahneman. *Attention and effort*. Prentice-Hall, Englewood Cliffs, NJ, 1973.

[98] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[99] S. Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems (NIPS)*, 2:1531–1538, 2002.

[100] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[101] R. E. Kalman. When is a linear control system optimal? *Journal of Basic Engineering*, 86(1):51–60, 1964.

[102] M. Kaustubh, D. Willemsen, and M. Mazo. The modeling of transfer of steering between automated vehicle and human driver using hybrid control framework. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 808–814. IEEE, 2016.

[103] S. D. Keen and D. J. Cole. Application of time-variant predictive control to modelling driver steering skill. *Vehicle System Dynamics*, 49(4):527–559, 2011.

[104] K. Kircher and C. Ahlström. Issues related to the driver distraction detection algorithm attend. In *1st International Conference on Driver Distraction and Inattention*, 2009.

[105] K. Kircher, C. Ahlstrom, and A. Kircher. Comparison of two eye-gaze based real-time driver distraction detection algorithms in a small-scale field operational test. In *Proceeding of the International Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pages 16–23, 2009.

[106] Katja Kircher and Christer Ahlström. Minimum required attention: a human-centered approach to driver inattention. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, page 0018720816672756, 2016.

[107] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–214. Springer, 2012.

[108] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Technical Report DOT HS 810 594, National Highway Traffic Safety Administration, Washington D.C., 2006.

[109] D. Kleinman, S. Baron, and W. H. Levison. An optimal control model of human response part I: Theory and validation. *Automatica*, 6(3):357–369, 1970.

[110] D. C. Knill. Mixture models and the probabilistic structure of depth cues. *Vision Research*, 43(7):831–854, 2003.

[111] D. C. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[112] M. Kocvara and M. Stingl. PENNON. In *High Performance Algorithms and Software for Nonlinear Optimization*, pages 303–321. Springer, 2003.

[113] M. Kondo and A. Ajimine. Driver's sight point and dynamics of the driver-vehicle-system related to it. Technical report, SAE Technical Paper, 1968.

[114] K. P. Körding and D. M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

[115] V. Krishnamurthy. Algorithms for optimal scheduling and management of hidden markov model sensors. *IEEE Transactions on Signal Processing*, 50(6):1382–1397, 2002.

[116] J. Kubitzki and W. Fastenmeier. Ablenkung durch moderne informations- und kommunikationstechniken und soziale interaktion bei autofahrern. Technical report, AZT Automotive GmbH Allianz Zentrum für Technik, Institut Mensch-Verkehr-Umwelt, Makam Research, 2016.

[117] M. Kuderer, S. Gulati, and W. Burgard. Learning driving styles for autonomous vehicles from demonstration. In *Proceedings of the IEEE International Conference on Robotics & Automation (ICRA)*, volume 134, 2015.

[118] M. Kuderer, H. Kretzschmar, and W. Burgard. Teaching mobile robots to cooperatively navigate in populated environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3138–3143, 2013.

[119] T. Kujala, H. Karvonen, and J. Mäkelä. Context-sensitive distraction warnings–effects on drivers? Visual behavior and acceptance. *International Journal of Human-Computer Studies*, 90:39–52, 2016.

[120] T. Kujala, J. Mäkelä, I. Kotilainen, and T. Tokkonen. The attentional demand of automobile driving revisited: Occlusion distance as a function of task-relevant event density in realistic driving scenarios. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(1):163–180, 2016.

[121] T. Kujala and D. D. Salvucci. Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control. *International Journal of Human-Computer Studies*, 79:66–78, 2015.

[122] C.-P. Lam, A. Y. Yang, K. Driggs-Campbell, R. Bajcsy, and S. S. Sastry. Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5776–5783. IEEE, 2015.

[123] D. Lamble, M. Laakso, and H. Summala. Detection thresholds in car following situations and peripheral vision: Implications for positioning of visually demanding in-car displays. *Ergonomics*, 42(6):807–815, 1999.

[124] M. F. Land, D. N. Lee, et al. Where we look when we steer. *Nature*, 369(6483):742–744, 1994.

[125] J. D. Lee, D. V. McGehee, T. L. Brown, and M. L. Reyes. Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(2):314–334, 2002.

[126] J. D. Lee, J. Moeckli, T. L. Brown, S. C. Roberts, C. Schwarz, L. Yekhshatyan, E. Nadler, Y. Liang, T. Victor, D. Marshall, et al. Distraction detection and mitigation through driver feedback. Technical Report DOT HS 811 547A, National Highway Traffic Safety Administration, Washington D.C., 2013.

[127] J. D. Lee, K. L. Young, and M. A. Regan. Defining driver distraction. *Driver distraction: Theory, effects, and mitigation*, page 31, 2008.

[128] J. Y. Lee, M. Gibson, and J. D. Lee. Secondary task boundaries influence drivers' glance durations. In *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 273–280. ACM, 2015.

[129] J. Y. Lee, M. C. Gibson, and J. D. Lee. Error recovery in multitasking while driving. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 5104–5113. ACM, 2016.

[130] K. Lee and H. Peng. Evaluation of automotive forward collision warning and collision avoidance algorithms. *Vehicle System Dynamics*, 43(10):735–751, 2005.

[131] R. J. Leigh and D. S. Zee. *The neurology of eye movements*, volume 90. Oxford University Press, USA, 2015.

[132] N. Lerner. Deciding to be distracted. In *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pages 499–505, 2005.

[133] S. Levine and V. Koltun. Continuous inverse optimal control with locally optimal examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 41–48, 2012.

[134] S. Levine and V. Koltun. Guided policy search. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1–9, 2013.

[135] N. Li and C. Busso. Predicting perceived visual and cognitive distractions of drivers with multimodal features. *IEEE Transactions on Intelligent Transportation Systems*, PP(99):1–15, 2014.

[136] N. Li, J. J. Jain, and C. Busso. Modeling of driver behavior in real world scenarios using multiple noninvasive sensors. *IEEE Transactions on Multimedia*, 15(5):1213–1225, 2013.

[137] Y. Liang, W. J. Horrey, and J. D. Hoffman. Reading text while driving: Understanding drivers' strategic and tactical adaptation to distraction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(2):347–359, 2015.

[138] Y. Liang, J. Lee, and M. Reyes. Nonintrusive detection of driver cognitive distraction in real time using bayesian networks. *Transportation Research Record: Journal of the Transportation Research Board*, (2018):1–8, 2007.

[139] Y. Liang and J. D. Lee. Combining cognitive and visual distraction: Less than the sum of its parts. *Accident Analysis & Prevention*, 42(3):881–890, 2010.

[140] Y. Liang and J. D. Lee. A hybrid bayesian network approach to detect driver cognitive distraction. *Transportation research part C: Emerging technologies*, 38:146–155, 2014.

[141] Y. Liang, J. D. Lee, and L. Yekhshatyan. How dangerous is looking away from the road? algorithms predict crash risk from glance patterns in naturalistic driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(6):1104–1116, 2012.

[142] Y. Liang, M. L. Reyes, and J. D. Lee. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):340–350, 2007.

[143] Y. Liao, S. E. Li, W. Wang, Y. Wang, G. Li, and B. Cheng. Detection of driver cognitive distraction: a comparison study of stop-controlled intersection and speed-limited highway. *IEEE Transactions on Intelligent Transportation Systems*, 17(6):1628–1637, 2016.

[144] C.-T. Lin, S.-A. Chen, T.-T. Chiu, H.-Z. Lin, and L.-W.i Ko. Spatial and temporal eeg dynamics of dual-task driving performance. *Journal of Neuroengineering and Rehabilitation*, 8(1):11, 2011.

[145] C.-T. Lin, L.-W. Ko, and T.-K. Shen. Computational intelligent brain computer interaction and its applications on driving cognition. *IEEE Computational Intelligence Magazine*, 4(4), 2009.

[146] T. Liu, Y. Yang, G.-B. Huang, Y. K. Yeo, and Z. Lin. Driver distraction detection using semi-supervised machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 17(4):1108–1120, 2016.

[147] Y.-C. Liu. Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveller information systems. *Ergonomics*, 44(4):425–442, 2001.

[148] L. Ljung. *System identification: Theory for the user*. Prentice Hall, Upper Saddle River and NJ [u.a.], 29th edition, 2006.

[149] C. C. MacAdam. Application of an optimal preview control for simulation of closed-loop automobile driving. *IEEE Transactions on Systems, Man and Cybernetics*, 1981.

[150] C. C. MacAdam. Understanding and modeling the human driver. *Vehicle System Dynamics*, 40(1-3):101–134, 2003.

[151] J. C. McCall and M. M. Trivedi. Video-based lane estimation and tracking for driver assistance: Survey, system, and evaluation. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):20–37, 2006.

[152] L. Meier, J. Peschon, and R. M. Dressler. Optimal control of measurement subsystems. *IEEE Transactions on Automatic Control*, 12(5):528–536, 1967.

[153] B. Metz and H.-P. Krueger. Measuring visual distraction in driving: The potential of head movement analysis. *IET Intelligent Transport Systems*, 4(4):289–297, 2010.

[154] B. Metz, A. Landau, and M. Just. Exposure to secondary tasks in germany: Results from naturalistic driving data. In *Proceedings of the International Conference on Driver Distraction and Inattention*, 2013.

[155] B. Metz, S. Schoch, M. Just, and F. Kuhn. How do drivers interact with navigation systems in real life conditions?: Results of a field-operational-test on navigation systems. *Transportation Research Part F: Traffic Psychology and Behaviour*, 24:146–157, 2014.

[156] B. Metz, N. Schömig, and H.-P. Krüger. Attention during visual secondary tasks in driving: Adaptation to the demands of the driving task. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(5):369–380, 2011.

[157] M. Miyaji, H. Kawanaka, and K. Oguri. Effect of pattern recognition features on detection for driver's cognitive distraction. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 605–610. IEEE, 2010.

[158] Y. Mo, R. Ambrosino, and B. Sinopoli. Sensor selection strategies for state estimation in energy constrained wireless sensor networks. *Automatica*, 47(7):1330–1338, 2011.

[159] M. Monfort, B. M. Lake, B. D. Ziebart, P. Lucey, and J. Tenenbaum. Softstar: Heuristic-guided probabilistic inference. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2746–2754, 2015.

[160] R. R. Mourant and T. H. Rockwell. Mapping eye-movement patterns to the visual scene in driving: An exploratory study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 12(1):81–87, 1970.

[161] R. R. Mourant and T. H. Rockwell. Strategies of visual search by novice and experienced drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 14(4):325–335, 1972.

[162] E. Muhrer and M. Vollrath. The effect of visual and cognitive distraction on driver's anticipation in a simulated car following scenario. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(6):555–566, 2011.

[163] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(4):607–626, 2009.

[164] C. J. Nash, D. J. Cole, and R. S. Bigler. A review of human sensory dynamics for application to models of driver steering and speed control. *Biological Cybernetics*, pages 1–26, 2016.

[165] J. A. Nelder and R. J. Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[166] G. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 295–302, 2007.

[167] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2000.

[168] B. Okal and K. O. Arras. Learning socially normative robot navigation behaviors with bayesian inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2889–2895. IEEE, 2016.

[169] P. Ondruska and I. Posner. The route not taken: Driver-centric estimation of electric vehicle range. In *Proceedings of the AAAI International Conference on International Conference on Automated Planning and Scheduling (ICAPS)*, pages 413–420. AAAI Press, 2014.

[170] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

[171] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

[172] N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[173] K. R. Pattipati, D. L. Kleinman, and A. R. Ephrath. A dynamic decision model of human task selection performance. *IEEE Transactions on Systems, Man, and Cybernetics*, (2):145–166, 1983.

[174] Y. Peng, L. N. Boyle, and S. L. Hallmark. Driver's lane keeping ability with eyes off road: Insights from a naturalistic study. *Accident Analysis & Prevention*, 50:628–634, 2013.

[175] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.

[176] A. Phatak, H. Weinert, I. Segall, and C. N. Day. Identification of a modified optimal control model for the human operator. *Automatica*, 12(1):31–41, 1976.

[177] W. Piechulla, C. Mayser, H. Gehrke, and W. König. Reducing drivers' mental workload by means of an adaptive man–machine interface. *Transportation Research Part F: Traffic Psychology and Behaviour*, 6(4):233–248, 2003.

[178] M. Plöchl and J. Edelmann. Driver models in automobile dynamics application. *Vehicle System Dynamics*, 45(7-8):699–741, 2007.

[179] J. Pohl, W. Birk, and L. Westervall. A driver-distraction-based lane-keeping assistance system. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 221(4):541–552, 2007.

[180] S. Pohlmann and U. Traenkle. Orientation in road traffic. Age-related differences using an in-vehicle navigation system and a conventional map. *Accident Analysis & Prevention*, 26(6):689–702, 1994.

[181] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[182] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *Proceedings of the International Joint Conference on Artifical Intelligence (IJAI)*, 2007.

[183] T. A. Ranney, W. Garrott, and M. Goodman. Nhtsa driver distraction research: past, present, and future. In *Proceedings of the International Technical Conference on the Enhanced Safety of Vehicles*, volume 2001, pages 9–p. National Highway Traffic Safety Administration, 2001.

[184] J. L. Rasmussen. Analysis of likert-scale data: A reinterpretation of gregoire and driver. *Psychological Bulletin*, 105(1,167-170), 1989.

[185] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 729–736. ACM, 2006.

[186] H. E. Rauch, C. T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

[187] M. Rezaei and R. Klette. Look at the driver, look at the road: No distraction! No accident! In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–136. IEEE, 2014.

[188] R. Risack, P. Klausmann, W. Krüger, and W. Enkelmann. Robust lane recognition embedded in a real-time driver assistance system. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 1998.

[189] R. Risack, N. Mohler, and W. Enkelmann. A video-based lane keeping assistant. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 356–361. IEEE, 2000.

[190] R. T. Rockafellar. *Convex analysis*. Princeton University Press, 2015.

[191] C. A. Rothkopf and D. H. Ballard. Credit assignment in multiple goal embodied visuomotor behavior. *Embodied and Grounded Cognition*, 217, 2010.

[192] C. A. Rothkopf and D. H. Ballard. Modular inverse reinforcement learning for visuomotor behavior. *Biological Cybernetics*, 107(4):477–490, 2013.

[193] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14):16–16, 2007.

[194] C. A. Rothkopf and C. Dimitrakakis. Preference elicitation and inverse reinforcement learning. In *Proceedings of the European Conference in Machine Learning and Knowledge Discovery in Databases (ECML)*, pages 34–48. Springer, 2011.

[195] D. Sadigh, S. Sastry, S. A. Seshia, and A. D. Dragan. Planning for autonomous cars that leverages effects on human actions. In *Proceedings of the Robotics: Science and Systems Conference (RSS)*, 2016.

[196] D. D. Salvucci. A multitasking general executive for compound continuous tasks. *Cognitive Science*, 29(3):457–492, 2005.

[197] D. D. Salvucci and R. Gray. A two-point visual control model of steering. *Perception-London*, 33(10):1233–1248, 2004.

[198] A. Sathyanarayana, P. Boyraz, and J. H. L. Hansen. Information fusion for robust 'context and driver aware' active vehicle safety systems. *Information Fusion*, 12(4):293–303, 2011.

[199] A. Sathyanarayana, P. Boyraz, Z. Purohit, R. Lubag, and J. H. L. Hansen. Driver adaptive and context aware active safety systems using CAN-bus signals. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 1236–1241. IEEE, 2010.

[200] S. Schaal, A. Ijspeert, and A. Billard. Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1431):537–547, 2003.

[201] F. Schmitt, H.-J. Bieg, M. Herman, and C. Rothkopf. I see what you see: Inferring sensor and policy models of human real-world motor behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[202] F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Exact maximum entropy inverse optimal control for modeling human attention scheduling and control. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

[203] F. Schmitt, H.-J. Bieg, D. Manstetten, M. Herman, and R. Stiefelhagen. Predicting lane keeping behavior of visually distracted drivers using inverse suboptimal control. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[204] F. Schmitt, H.-J. Bieg, D. Manstetten, and R. Stiefelhagen. Distraction mitigation by computation of appropriate glance behavior and its evaluation in a user test. *(manuscript in preparation) IEEE Transaction on Intelligent Transport Systems*, 2017.

[205] N. Schömig, B. Metz, and H.-P. Krüger. Anticipatory and control processes in the interaction with secondary tasks while driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(6):525–538, 2011.

[206] C. A. Schriesheim and L. Novelli Jr. A comparative test of the interval-scale properties of magnitude estimation and case iii scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement*, 49(1):59–74, 1989.

[207] J. W. Senders, A. B. Kristofferson, W. H. Levison, C. W. Dietrich, and J. L. Ward. The attentional demand of automobile driving. *Highway Research Record*, (195), 1967.

[208] T. B. Sheridan. *Telerobotics, automation, and human supervisory control*. 1992.

[209] T. B. Sheridan. Driver distraction from a control theory perspective. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(4):587–599, 2004.

[210] M. Shimakage, S. Satoh, K. Uenuma, and H. Mouri. Design of lane-keeping control with steering torque input. *JSAE review*, 23(3):317–323, 2002.

[211] M. Shimosaka, T. Kaneko, and K. Nishi. Modeling risk anticipation and defensive driving on residential roads with inverse reinforcement learning. In *Proceedings of the IEEE Conference on Intelligent Transport Systems (ITSC)*, 2014.

[212] H. A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

[213] M. Simon, E. A. Schmidt, W. E. Kincses, M. Fritzsche, A. Bruns, C. Aufmuth, M. Bogdan, W. Rosenstiel, and M. Schrauf. EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. *Clinical Neurophysiology*, 122(6):1168–1178, 2011.

[214] M. Sivak. The information that drivers use: is it indeed 90% visual? *Perception*, 25:1081–1089, 1996.

[215] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[216] E. J. Sondik. *The Optimal Control of Partially Observable Markov Decision Processes*. PhD thesis, Stanford University, 1971.

[217] A. Sonnleitner, M. Simon, W. E. Kincses, A. Buchner, and M. Schrauf. Alpha spindles as neurophysiological correlates indicating attentional shift in a simulated driving task. *International Journal of Psychophysiology*, 83(1):110–118, 2012.

[218] A. Sonnleitner, M. S. Treder, M. Simon, S. Willmann, A. Ewald, A. Buchner, and M. Schrauf. EEG alpha spindles and prolonged brake reaction times during auditory distraction in an on-road driving study. *Accident Analysis & Prevention*, 62:110–118, 2014.

[219] F. Soyka, P. R. Giordano, K. Beykirch, and H. H. Bülthoff. Predicting direction detection thresholds for arbitrary translational acceleration profiles in the horizontal plane. *Experimental Brain Research*, 209(1):95–107, 2011.

[220] N. Sprague and D. Ballard. Eye movements for reward maximization. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[221] J. E. Stellet, J. Schumacher, W. Branz, and J. M. Zöllner. Performance bounds on change detection with application to manoeuvre recognition for advanced driver assistance systems. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 1112–1119. IEEE, 2015.

[222] J. E. Stellet, F. Straub, J. Schumacher, W. Branz, and J. M. Zöllner. Estimating the process noise variance for vehicle motion models. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 1512–1519, 2015.

[223] Jan Erik Stellet, Patrick Vogt, Jan Schumacher, Wolfgang Branz, and J Marius Zöllner. Analytical derivation of performance bounds of autonomous emergency brake systems. In *Intelligent Vehicles Symposium (IV), 2016 IEEE*, pages 220–226. IEEE, 2016.

[224] G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, 1977.

[225] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585, 2006.

[226] H. Strasburger, I. Rentschler, and M. Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13–13, 2011.

[227] D. L. Strayer, F. A. Drews, and W. A. Johnston. Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, 9(1):23, 2003.

[228] B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. M. Hayhoe. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of Vision*, 12(13):19, 2012.

[229] H. Summala. Forced peripheral vision driving paradigm: Evidence for the hypothesis that car drivers learn to keep in lane with peripheral vision. *Vision in Vehicles*, 6:51–60, 1998.

[230] H. Summala, T. Nieminen, and M. Punto. Maintaining lane position with peripheral vision during in-vehicle tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 38(3):442–451, 1996.

[231] R. S Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT Press Cambridge, 1998.

[232] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 99, pages 1057–1063, 1999.

[233] J. A. Swets. *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press, 2014.

[234] U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1449–1456, 2007.

[235] F. Tango and M. Botta. Real-time detection system of driver distraction using machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):894–905, 2013.

[236] A. Tawari, S. Martin, and M. M. Trivedi. Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):818–830, 2014.

[237] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer. Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 115–120. IEEE, 2014.

[238] Inc. The MathWorks. Matlab statistics toolbox™ user's guide, 2014.

[239] E. Tivesten and M. Dozza. Driving context influences drivers' decision to engage in visual–manual phone tasks: Evidence from a naturalistic driving study. *Journal of Safety Research*, 53:87–96, 2015.

[240] E. Todorov and W. Li. A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the American Control Conference (ACC)*, pages 300–306. IEEE, 2005.

[241] K. Torkkola, N. Massey, and C. Wood. Driver inattention detection through intelligent analysis of readily available sensors. In *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pages 326–331. IEEE, 2004.

[242] M. Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the International Conference on Machine Learning (ICML)*, page 132, 2009.

[243] A. R. Valente Pais, D. M. Pool, A. M. De Vroome, M. M. Van Paassen, and M. Mulder. Pitch motion perception thresholds during passive and active tasks. *Journal of Guidance, Control, and Dynamics*, 35(3):904–918, 2012.

[244] J. Van Den Berg, S. Patil, and R. Alterovitz. Motion planning under uncertainty using iterative local optimization in belief space. *The International Journal of Robotics Research*, 31(11):1263–1278, 2012.

[245] T. W. Victor, J. L. Harbluk, and J. A. Engström. Sensitivity of eye-movement measures to in-vehicle task difficulty. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):167–190, 2005.

[246] M. P. Vitus, W. Zhang, A. Abate, J. Hu, and C. J. Tomlin. On efficient sensor scheduling for linear dynamical systems. *Automatica*, 48(10):2482–2493, 2012.

[247] B. Wandtner, M. Schumacher, and E. A. Schmidt. The role of self-regulation in the context of driver distraction: A simulator study. *Traffic Injury Prevention*, 17(5):472–479, 2016.

[248] S. Wang, Y. Zhang, C. Wu, F. Darvas, and W. A. Chaovalitwongse. Online prediction of driver distraction based on brain activity patterns. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):136–150, 2015.

[249] T. Wertheim. Über die indirekte sehschärfe. *Zeitschrift für Psychologie & Physiologie der Sinnesorgane*, 7, 1894.

[250] M. Wollmer, C. Blaschke, T. Schindl, B. Schuller, B. Farber, S. Mayer, and B. Trefflich. Online driver distraction detection using long short-term memory. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):574–582, 2011.

[251] L. Yekhshatyan and J. D. Lee. Changes in the correlation between eye and steering movements indicate driver distraction. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):136–145, 2013.

[252] K. Young, J. D. Lee, and M. A. Regan. *Driver distraction: Theory, effects, and mitigation*. CRC Press, 2008.

[253] K. L. Young and P. M. Salmon. Examining the relationship between driver distraction and driving errors: A discussion of theory, studies and methods. *Safety Science*, 50(2):165–174, 2012.

[254] K. Zeeb, A. Buchner, and M. Schrauf. Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident Analysis & Prevention*, 92:230–239, 2016.

[255] H. Zhao and W. H. Warren. On-line and model-based approaches to the visual control of action. *Vision Research*, 110:190–202, 2015.

[256] J. Zheng, S. Liu, and L. M. Ni. Robust bayesian inverse reinforcement learning with sparse behavior noise. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2198–2205, 2014.

[257] B. D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

[258] B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1255–1262, 2010.

[259] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1433–1438, 2008.

[260] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proceedings of the International Conference on Ubiquitous Computing*, pages 322–331. ACM, 2008.

[261] Brian Ziebart, Anind Dey, and J Andrew Bagnell. Probabilistic pointing target prediction via inverse optimal control. In *Proceedings of the ACM international conference on Intelligent User Interfaces*, pages 1–10. ACM, 2012.