Yet Another Distributional Bellman Equation

Nicole Bäuerle[a1], **Tamara Göll**[a], **Anna Jaśkiewicz**[b]

[a]Department of Mathematics, Karlsruhe Institute of Technology, Karlsruhe, Germany, email: *nicole.baeuerle@kit.edu*, *tamara.goell@kit.edu*

[b]Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Wrocław, Poland, email: *anna.jaskiewicz@pwr.edu.pl*

October 16, 2025

Keywords. Dynamic programming, Markov Decision Process, Bellman Equation, Non-Markovian Process

Abstract. We consider non-standard Markov Decision Processes (MDPs) where the target function is not only a simple expectation of the accumulated reward. Instead, we consider rather general functionals of the joint distribution of terminal state and accumulated reward which have to be optimized. For finite state and compact action space, we show how to solve these problems by defining a lifted MDP whose state space is the space of distributions over the true states of the process. We derive a Bellman equation in this setting, which can be considered as a distributional Bellman equation. Well-known cases like the standard MDP and quantile MDPs are shown to be special examples of our framework. We also apply our model to a variant of an optimal transport problem.

1. Introduction

Classical MDP theory is concerned with the maximization of expected accumulated reward $\mathbb{E}R_N$ over N discrete stages or an infinite time horizon where the transition law of the process can be controlled. Thus, we are interested in the first moment of the random variable R_N representing the accumulated reward and aim to maximize this. For the standard theory, solving these problems with the help of a Bellman equation, a recursive equation for the optimal value of the problem depending on the time horizon, see Puterman (2014); Hernández-Lerma and Lasserre (1996); Bäuerle and Rieder (2011).

However, later there has been increased interest in criteria which also account for risk in the sense of deviation from the mean or in higher moments of the accumulated reward. Hence, in criteria which address other aspects of the underlying return distribution. A widely discussed criterion for example is the risk-sensitive reward $-\mathbb{E}\exp(-\lambda R_N)$ which has first been addressed by Howard and Matheson (1972); Jaquette (1976). It can be considered as a criterion taking all moments of the accumulated reward into account since

$$-\mathbb{E}\exp(-\lambda R_N) = -\sum_{k=0}^{\infty} \frac{(-\lambda)^k \mathbb{E}(R_N^k)}{k!}.$$

Risk-sensitive Markov decision problems can again be solved using a Bellman equation, but the value function is time-dependent even in the stationary infinite horizon setting (see the aforementioned references). For Q-learning in this setting, see Borkar

[1]Corresponding author

(2002); Mihatsch and Neuneier (2002); Borkar (2010); Shen et al. (2014). Further extensions include more general expected utility (Chung and Sobel (1987); Bäuerle and Rieder (2014)), the application of risk measures (Ruszczyński (2010); Bäuerle and Ott (2011); Shen et al. (2013); Bäuerle and Glauner (2021); Moghimi and Ku (2025)) or the consideration of mean-variance problems (Mannor and Tsitsiklis (2011); Cui et al. (2014); Bäuerle and Jaśkiewicz (2025); Bäuerle et al. (2025)). In the first two cases, it is again possible to solve the problem with a Bellman equation on an extended state space, where an auxiliary variable like the 'accumulated reward so far' has to be added to the natural state of the process. For mean-variance problems, one can take advantage of the fact that the variance can be represented as an optimization problem. A recent overview can be found in Bäuerle and Jaśkiewicz (2024). Besides this, other papers considered quantile optimization in the sense that $\mathbb{P}(R_N \geq t)$ for fixed $t \in \mathbb{R}$ has to be optimized, see e.g. Filar et al. (1995); Wu and Lin (1999); Chow et al. (2015); Gilbert et al. (2017); Li et al. (2022). In this case, the problem can again be solved using a state space extension. In particular, the latter three papers are also concerned with finding efficient algorithms to solve MDPs with quantile criteria. In Marthe et al. (2024) the authors discuss among others which utility problems may be solved by dynamic programming.

On the other hand, there is a stream of literature which deals with distributional reinforcement learning. In this theory, the aim is to learn the distribution of the accumulated returns under a stationary Markovian policy (Bellemare et al. (2017, 2023); Rowland et al. (2019)) or equivalently to learn the quantile function (Dabney et al. (2018b,a)) and use parametric families to achieve this efficiently. This is similar to Q-learning for a fixed policy. Optimization enters the scene by choosing greedy actions according to an appropriate functional which relates to the target quantity. However, in general this theory applies to problems where the classical MDP theory can be used (maybe on an extended state space) and where the optimal policy can be found within the deterministic stationary policies (Lyle et al. (2019) compares traditional RL (reinforcement learning) to distributional RL).

In this paper, we combine the distributional approach with a very general optimization target involving the joint distribution of current state and accumulated reward. More precisely, if F_N^σ is the joint distribution of the accumulated reward R_{N-1} and current state X_N under policy σ, we are interested in maximizing $H(F_N^\sigma)$ where H maps distributions to real values. Hence, H could map the distribution on the expectation of R_{N-1}, yielding a classical MDP, to a quantile, to a probability or to the distance to a given distribution. This gives a very flexible target, comprising, in principle, all previously considered cases. Also situations with constraints may be handled (see e.g. Altman (2021); Basak and Shapiro (2001)). Of course, and this is crucial to note, we cannot expect that we can restrict the search for optimal policies to Markovian policies, nor can we restrict to deterministic policies in general. However, we can show that optimal policies depend only on the current distribution of state and accumulated reward so far. Thus, they depend on the history of the process, but only through a certain sufficient statistic. We can also identify cases where the optimal decision rule is not randomized. This is achieved by introducing a 'lifted' MDP with states given by distributions. The approach is inspired by recent papers considering MDPs with distributions as states (see e.g. Bäuerle and Jaśkiewicz (2025)). In order to keep the exposition simple, we restrict here to finite state and action spaces and in Sec. 3 to finite state and compact action space. The recent paper Pires et al. (2025) considers a related question, however is concentrating on infinite horizon discounted reward. They consider a possibly infinite state space, but a finite action space. In their main theorem, the target function H has to satisfy some indifference properties and they do not define a 'lifted' MDP, hence there is no typical value function. However, they provide a large number of different applications ranging from Conditional Value at Risk optimization to Deep η networks.

The *contributions* of our paper are: (i) We introduce in a proper mathematical way a very general optimization problem involving the joint distribution of the accumulated reward and current state of a Markov Decision process. (ii) We introduce a 'lifted' MDP to solve the problem recursively. We do this for finite state and finite/compact action spaces. (iii) We show that cases previously treated in the literature like classical MDP and quantile MDP are included as special cases. (iv) For the infinite horizon problem, we discuss existence and reduction of optimal policies. (v) We give a new application of approximating a given distribution by a random walk which we solve in a naive dynamic programming fashion. This is a modification of an optimal mass transportation problem.

The *outline* of the paper is as follows: In the next section we introduce the model with finite state and action spaces and derive the recursive solution algorithm under a continuity assumption on the objective mapping H. In Sec. 3 we consider the case of a compact action space but restrict to distributions of the terminal state plus expected reward. In order to obtain optimal policies, we need a further continuity property of the transition law. In Sec. 4 we briefly discuss the infinite horizon problem in our setting. Using an example, we show that optimal stationary policies cannot be expected. However, we prove the existence of optimal policies when the target function is Wasserstein-Lipschitz and can reduce them to a certain set of policies. In the last section we consider different applications: We show that the usual Bellman equation is included for the classical objective of maximizing the expected accumulated reward. We also consider a quantile MDP where we show a similar result. The last application is non-standard: We determine the transition probabilities of a Markov chain in such a way that, at a fixed time point, a weighted criterion of distance to a given distribution and expected transportation cost is minimized. We solve this problem using naive dynamic programming and discuss the results.

2. The model

By $\mathbb{R}$ ($\mathbb{N}$) we denote the set of all real numbers (positive integers). Let Y be a Borel space, i.e., a Borel subset of a complete separable metric space with its Borel σ-algebra $\mathcal{B}(Y)$. Then, $P(Y)$ stands for the set of all probability distributions on $(Y, \mathcal{B}(Y))$. We equip $P(Y)$ with the weak topology, that is, the coarsest topology for which the mapping $\mu \to \int_Y f d\mu$ is continuous for every bounded and continuous function $f : Y \to \mathbb{R}$.

Suppose we are given a classical Markov decision model with finite state and action space. We are first interested in models with finite time horizon $N \in \mathbb{N}$. More precisely, the model is described by the following items:

(i) E is the finite state space,
(ii) A is the finite action space,
(iii) $r : E \times A \to \mathbb{R}$ is the one-stage reward,
(iv) q is the transition probability from $E \times A$ to E,
(v) $g : E \to \mathbb{R}$ is the terminal reward.

Thus, let $(\Omega, \mathcal{F})$ be the measurable space with $\Omega = E \times (A \times E)^N$ and $\mathcal{F}$ the corresponding power set. We define the set of all histories by $H_n := E \times (A \times E)^n$, i.e., $h_n = (x_0, a_0, \dots, x_n)$ (for $n = 1, \dots, N$) describes the sequence of states and actions which have occurred up to time n. For $n = 0$, we have $H_0 = E$. The state process is then given by the random variables $X_n : \Omega \to E$ with $X_n(h_N) = x_n$ and the action process by the random variables $A_n : \Omega \to A$ with $A_n(h_N) = a_n$. A policy σ is a sequence $(\sigma_n)_{n=0}^{N-1}$ of history dependent decision rules, where each decision rule σ_n specifies the probability distribution $\sigma_n(\cdot|h_n)$ on the action space A for $h_n \in H_n$ and $n = 0, \dots, N-1$. We denote this set of policies by Π_N. It is well-known that a policy $\sigma = (\sigma_n)_{n=0}^{N-1} \in \Pi_N$, an initial distribution ν and the transition probability q induce a probability distribution $\mathbb{P}_\nu^\sigma$ on $(\Omega, \mathcal{F})$ (see p. 23 in Puterman (2014)) so that

$$\begin{aligned}&\mathbb{P}_\nu^\sigma(X_0 = x_0, A_0 = a_0, X_1 = x_1, \dots, X_N = x_N) =\\ &\nu(x_0)\sigma_0(a_0|x_0)q(x_1|x_0, a_0)\sigma_1(a_1|x_0, a_0, x_1) \cdot \ldots \cdot \sigma_{N-1}(a_{N-1}|h_{N-1})q(x_N|x_{N-1}, a_{N-1}).\end{aligned}$$

In classical MDP theory, we are interested in maximizing the expected reward of the system over the time horizon when the initial distribution is ν. Thus, we define for $n = 0, 1, \dots, N-1$

$$R_n := \sum_{k=0}^{n} r(X_k, A_k) \quad \text{and} \quad R_N := \sum_{k=0}^{N-1} r(X_k, A_k) + g(X_N).$$

For a fixed initial distribution ν, one then tries to solve

$$V(\nu) = \sup_{\sigma \in \Pi_N} \mathbb{E}_\nu^\sigma[R_N]$$

where $\mathbb{E}_\nu^\sigma$ denotes the expectation w.r.t. $\mathbb{P}_\nu^\sigma$. It is well-known that the optimal policy (which exists here, since state and action spaces are finite) can be found among the deterministic Markovian decision rules and that it does not depend on ν. More precisely, when we define $V_N(x) := g(x)$ and for $n = N-1, \dots, 0$

$$V_n(x) = \max_{a \in A} \big\{ r(x, a) + \sum_{x'} V_{n+1}(x') q(x'|x, a) \big\}, \tag{2.1}$$

then $V(\nu) = \sum_x V_0(x)\nu(x)$ and the maximizers in (2.1) define an optimal (deterministic) policy.

In this paper however, we generalize the optimization criterion to one where the joint distribution of the accumulated reward and terminal state is involved. Since state and action spaces are finite, the random variables $R_0, \dots, R_N$ take only a finite number of possible values. In what follows, we denote by S the finite set of all possible realizations of the random variables $R_0, R_1, \dots, R_N$. For a fixed policy $\sigma \in \Pi_N$, let F_n^σ be the joint distribution of (X_n, R_{n-1}), i.e.,

$$F_n^\sigma(x, s) = \mathbb{P}_\nu^\sigma(X_n = x, R_{n-1} = s), \quad (x, s) \in E \times S, \ n \geq 1.$$

Obviously, this distribution depends on the chosen policy σ (and the initial distribution ν which is fixed and thus not part of the notation). Let now $H : P(E \times S) \to \mathbb{R}$ be an arbitrary functional. The aim is to solve the following optimization problem

$$\sup_{\sigma \in \Pi_N} H(F_N^\sigma). \tag{2.2}$$

The solution in (2.2) is called *a value function.*

Example 2.1. Let us consider some special cases of our setting. Let $F \in P(E \times S)$.

a) Of course the classical case is included by defining

$$H(F) = \sum_{x,s} (g(x) + s) F(x, s).$$

Obviously $H(F_N^\sigma) = \mathbb{E}_\nu^\sigma[R_N]$ in this case and it is just the expected accumulated reward of the system and we can use the standard Bellman equation (2.1) to solve the problem.

b) When we choose $H(F) = \sum_{x,s} F(x,s)\chi(x,s)$, where

$$\chi(x,s) := \begin{cases} 1, & \text{if } s + g(x) \geq t \\ 0, & \text{else} \end{cases}$$

for a fixed $t \in \mathbb{R}$, then $H(F_N^\sigma) = \mathbb{P}_\nu^\sigma(R_N \geq t)$ is the quantile of the terminal accumulated reward. By varying the function χ we can treat the probability that the accumulated reward ends up in different areas.

c) Let $E \subset \mathbb{R}$. If W_1 is the Wasserstein distance between the distributions, then we might consider

$$H(F) = W_1(F(\cdot, S), G) = \int_{\mathbb{R}} |F_c(t) - G_c(t)| dt,$$

where $G \in P(\mathbb{R})$ is the given target distribution. Here, F_c and G_c denote the corresponding cumulative distributions. Then the aim is to choose the policy σ such that the distribution of X_N is as close as possible to G, i.e., we wish to find $\inf_{\sigma \in \Pi_N} H(F_N^\sigma)$. The distance W_1 may be replaced by any other reasonable distance between distributions.

In order to solve problems like (2.2), we have to lift the MDP to a more general state space which is given by joint distributions of the accumulated reward and original state of the process. Moreover, actions in the lifted MDP are transition kernels from the augmented state (x,s) of the process to the action space. Thus, we define

$$\Pi^M := \{\pi : E \times S \to P(A)\},$$

where S is as before the finite set of all possible realizations of the accumulated rewards. More precisely, we formally define the *lifted MDP* which is a deterministic dynamic control problem by the following data:

(i) $P(E \times S)$ is the state space where the interpretation of a state $F_n \in P(E \times S)$ at time point n is given as the joint distribution of (X_n, R_{n-1}). Note here that since state and action spaces are finite, F_n is ultimately a discrete distribution on a finite number of points.

(ii) Π^M is the action space.

(iii) The one-stage reward is zero.

(iv) $H : P(E \times S) \to \mathbb{R}$ is the terminal reward.

(v) The transition function $T : P(E \times S) \times \Pi^M \to P(E \times S)$ is given by

$$T^\pi(F)(x', s') := T(F, \pi)(x', s') = \sum_{(x,s,a)\,:\, r(x,a) = s'-s} q(x'|x,a)\pi(a|x,s)F(x,s), \tag{2.3}$$

where $(x', s') \in E \times S$ and q and r are the data from our initial MDP defined at the beginning of this section.

The previous data defines a lifted MDP which is indeed a *deterministic* dynamic control problem. Policies in this lifted model are denoted by (f_n) where $f_n : P(E \times S) \to \Pi^M$. Though state and action spaces are finite in the original formulation, this is no longer true in the lifted MDP. Indeed, the price we have to pay for the model to be deterministic now is that the state and action spaces consist of probability distributions and transition kernels, respectively. But in the end this also implies that there is no randomization over the lifted action space necessary in order to improve the value.

We first prove the following crucial connection between the original MDP and the lifted MDP. The notation $\nu \otimes \mu$ stands for the product measure between the two probability distributions μ, ν.

Proposition 2.2. *For every policy $\sigma = (\sigma_n)_{n=0}^{N-1}$ in the original MDP model, there exists an action sequence $(\pi_0, \ldots, \pi_{N-1})$ in the lifted MDP such that*

$$\mathbb{P}_\nu^\sigma(X_N = x, R_{N-1} = s) = T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_0}(F_0)(x, s), \quad (x, s) \in E \times S,$$

where $F_0 = \nu \otimes \delta_0$ and T^π is the operator defined in (2.3).

Proof. The proof is by induction over the length of the time horizon. First consider $N = 1$. Then

$$\mathbb{P}_\nu^\sigma(X_1 = x', r(X_0, A_0) = s') = \sum_{(x,a)\,:\,r(x,a)=s'} q(x'|x, a)\sigma_0(a|x)\nu(x) = T^{\pi_0}(F_0)(x', s')$$

where $\pi_0(a|x, s) = \pi_0(a|x, 0) := \sigma_0(a|x)$. Now suppose the statement is correct up to time $n \le N - 1$. For abbreviation we denote $F_n^{(\pi_0,\ldots,\pi_{n-1})} = T^{\pi_{n-1}} \circ T^{\pi_{n-2}} \circ \ldots \circ T^{\pi_0}(F_0)$ for all $n \le N - 1$. We have to show that $\mathbb{P}_\nu^\sigma(X_{n+1} = x, R_n = s) = F_{n+1}^{(\pi_0,\ldots,\pi_n)}(x, s)$. Thus, we obtain by using the induction hypothesis in the second equation:

$$\begin{aligned}
&\mathbb{P}_\nu^\sigma(X_{n+1} = x', R_n = s') = \\
&\sum_{x,s} \mathbb{P}_\nu^\sigma(X_{n+1} = x', R_n = s'|X_n = x, R_{n-1} = s)\mathbb{P}_\nu^\sigma(X_n = x, R_{n-1} = s) = \\
&\sum_{x,s} \mathbb{P}_\nu^\sigma(X_{n+1} = x', R_n = s'|X_n = x, R_{n-1} = s)F_n^{(\pi_0,\ldots,\pi_{n-1})}(x, s) = \\
&\qquad\sum_{(x,s,a)\,:\,r(x,a)=s'-s} q(x'|x, a)\mathbb{P}_\nu^\sigma(A_n = a|X_n = x, R_{n-1} = s)F_n^{(\pi_0,\ldots,\pi_{n-1})}(x, s) = \\
&\qquad\sum_{(x,s,a)\,:\,r(x,a)=s'-s} q(x'|x, a)\pi_n(a|x, s)F_n^{(\pi_0,\ldots,\pi_{n-1})}(x, s) = T^{\pi_n}(F_n^{(\pi_0,\ldots,\pi_{n-1})})(x', s'),
\end{aligned}$$

where we define $\pi_n(a|x, s) := \mathbb{P}_\nu^\sigma(A_n = a|X_n = x, R_{n-1} = s)$. Obviously, the definition of π depends on σ. This proves the statement. □

Remark 2.3. Proposition 2.2 may appear surprising at first sight, since there is an arbitrary (history-dependent) policy on the left-hand side and a 'Markov' policy (in the sense of the lifted MDP) on the right hand side. However, it is well-known that in Markov decision processes, for any history dependent policy, we can always find a Markov policy such that the distribution of state and action at every time point (i.e., the marginal distributions) coincide, see e.g. Theorem 2 in Derman and Strauch (1966) or Corollary 2.1 in Kallenberg (2002). This is essentially what also happens in Proposition 2.1.

On the other hand, given any action sequence $(\pi_n)_{n=0}^{N-1}$ it is immediate (since the accumulated reward up to time $n - 1$ is a function of the history up to that time) that by choosing $\sigma_n(\cdot|h_n) := \pi_n(\cdot|x_n, s_{n-1})$, we can construct from any sequence $(\pi_n)_{n=0}^{N-1}$ a sequence $(\sigma_n)_{n=0}^{N-1}$ such that the equality in Proposition 2.2 holds. As a result, we can write problem (2.2) as follows:

$$\sup_{\sigma\in\Pi_N} H(F_N^\sigma) = \sup_{(\pi_0,\ldots,\pi_{N-1})} H\big(T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_0}(F_0)\big). \tag{2.4}$$

The right-hand side in (2.4) is a deterministic dynamic control problem, which can be solved with the help of the value functions J_n, where

$$\begin{aligned} J_N(F) &:= H(F), \\ J_n(F) &:= \sup_{\pi \in \Pi^M} J_{n+1}(T^\pi(F)), \quad n = 0, 1, \ldots, N-1, \end{aligned} \tag{2.5}$$

for $F \in P(E \times S)$. We make the following assumption:

(C1): The mapping $F \to H(F)$ is upper semicontinuous, that is, $\limsup_{k\to\infty} H(F_{(k)}) \leq H(F)$ for any sequence $(F_{(k)})$ converging to F in $P(E \times S)$ as $k \to \infty$, i.e., $F_{(k)}(x,s) \to F(x,s)$ for every $(x,s) \in E \times S$.

Note that this condition is satisfied for all cases in Example 2.1.

We obtain the following result:

Theorem 2.4. *Assume (C1), i.e., H is upper semicontinuous.*

a) *If (J_n) is computed according to* (2.5)*, then $J_0(F_0)$ is the value function of problem* (2.2)*, i.e., $J_0(F_0) = \sup_{\sigma \in \Pi_N} H(F_N^\sigma)$.*

b) *Maximizers $(f_0^*, \ldots, f_{N-1}^*)$ in the recursion of* (2.5) *exist, i.e., for $n = 0, 1, \ldots, N-1$*

$$f_n^*(F) = \underset{\pi}{\operatorname{argmax}}\, J_{n+1}(T^\pi(F)), \; F \in P(E \times S).$$

c) *Define the following state-action sequence starting with state F_0:*

$$\begin{aligned} \pi_0^* &= f_0^*(F_0), \\ F_1 &= T^{\pi_0^*}(F_0), \\ &\vdots \\ \pi_n^* &= f_n^*(F_n), \\ F_{n+1} &= T^{\pi_n^*} \circ \ldots \circ T^{\pi_0^*}(F_0). \end{aligned}$$

Then, $(\pi_0^*, \pi_1^*, \ldots, \pi_{N-1}^*)$ determines an optimal policy $(\sigma_0^*, \ldots, \sigma_{N-1}^*)$ for problem (2.4) *as follows*

$$\sigma_n^*(a|h_n) = \pi_n^*\Big(a \Big| x_n, \sum_{k=0}^{n-1} r(x_k, a_k)\Big)$$

for $n = 0, 1, \ldots, N-1$, where $h_n = (x_0, a_0, \ldots, x_n)$.

Proof. The proof proceeds by backward induction. For part a) we prove that (J_n) computed according to (2.5) satisfies

$$J_n(F) = \sup_{(\pi_n, \ldots, \pi_{N-1})} H\big(T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_n}(F)\big)$$

for $n = 0, \ldots, N-1$. For $n = N-1$, the statement follows by definition. Suppose the statement is true for $n+1$. We prove that it is also true for n. By (2.5) and the induction hypothesis we obtain

$$\begin{aligned} J_n(F) &= \sup_{\pi_n \in \Pi^M} J_{n+1}(T^{\pi_n}(F)) \\ &= \sup_{\pi_n \in \Pi^M} \sup_{(\pi_{n+1}, \ldots, \pi_{N-1})} H\big(T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_{n+1}}(T^{\pi_n}(F))\big) \\ &= \sup_{(\pi_n, \ldots, \pi_{N-1})} H\big(T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_n}(F)\big) \end{aligned}$$

which proves the statement. It remains to show that the maximum points exist. We do this again by induction, proving that all J_n are upper semicontinuous and the maximum points exist. The argument for the induction starting at time point $N-1$ is essentially the same as for the induction step and we thus skip it. Set $m = |E| \cdot |S|$. Consider the optimization problem in (2.5) and assume that $J_{n+1} : \mathbb{R}^m \to \mathbb{R}$ is upper semicontinuous. Let $\pi \in \Pi^M$. Clearly, if $F_{(k)} \to F$ in $P(E \times S)$ and $\pi_{(k)} \to \pi$ in Π^M (i.e., $\pi_{(k)}(a|x,s) \to \pi(a|x,s)$ for every $a \in A$ and $(x,s) \in E \times S$), then

$$\sum_{(x,s,a)\,:\,r(x,a)=s'-s} q(x'|x,a)\pi_{(k)}(a|x,s)F_{(k)}(x,s) \to \sum_{(x,s,a)\,:\,r(x,a)=s'-s} q(x'|x,a)\pi(a|x,s)F(x,s)$$

for every $(x', s') \in E \times S$ as $k \to \infty$. Further, the set Π^M is compact, since $\Pi^M = P(A)^m$. By Proposition 2.4.3 in Bäuerle and Rieder (2011) there exists $\pi^* \in \Pi^M$ such that

$$\begin{aligned}
&\sup_{\pi\in\Pi^M} J_{n+1}\Big(\sum_{(x,s,a)\,:\,r(x,a)=\cdot-s} q(\cdot|x,a)\pi(a|x,s)F(x,s)\Big)\\
&= J_{n+1}\Big(\sum_{(x,s,a)\,:\,r(x,a)=\cdot-s} q(\cdot|x,a)\pi^*(a|x,s)F(x,s)\Big)
\end{aligned}$$

and the mapping

$$F \to \sup_{\pi\in\Pi^M} J_{n+1}\Big(\sum_{(x,s,a)\,:\,r(x,a)=\cdot-s} q(\cdot|x,a)\pi(a|x,s)F(x,s)\Big)$$

is upper semicontinuous, which proves the statement. Note that the connection between π^* and σ^* follows from the proof of Proposition 2.2. □

We call the last equation in (2.5) a 'distributional Bellman equation'. The term 'distributional Bellman equation' has been used before (see e.g. Bellemare et al. (2023)). However, there the term refers to a recursive computation of the cumulated reward distribution under a fixed policy, i.e., in a first step there is no optimization involved.

Remark 2.5. Sometimes it may be sufficient to consider the distribution of X_N only, without R_{N-1}. In this case, the transition function simplifies as follows:

$$\begin{aligned}
T^\pi(F)(x', S) &:= \sum_{(x,s,a)} q(x'|x,a)\pi(a|x,s)F(x,s)\\
&= \sum_{(x,a)} q(x'|x,a) \sum_s \pi(a|x,s)F(x,s)\\
&= \sum_{(x,a)} q(x'|x,a)\tilde{\pi}(a|x)F(x,S),
\end{aligned}$$

where

$$\tilde{\pi}(a|x) := \frac{\sum_s \pi(a|x,s)F(x,s)}{\sum_s F(x,s)}.$$

Note that $\tilde{\pi}(a|x)$ is a probability distribution on the action space, since $\tilde{\pi}(a|x) \geq 0$ and obviously $\sum_a \tilde{\pi}(a|x) = 1$. Moreover, $\tilde{\pi}(a|x)$ depends only on x. Hence, as the states in the lifted MDP it is sufficient to consider distributions $F \in P(E)$.

3. A special continuous case

In this section we briefly discuss the case of a compact (not necessarily finite) Borel action space where we restrict ourselves to objective functions which depend only on the law of X_N and add the expected reward up to time $N-1$. As noted in Remark 2.5, in this case it is sufficient to define the state of the process at time n in the lifted MDP as the distribution of X_n. More precisely, what we change w.r.t. the model in the previous section is that

(I): The action set A is a compact metric space.

Any given distribution ν for the initial state and a policy $\sigma \in \Pi_N$ define a unique probability measure $\mathbb{P}^\sigma_\nu$ over the space of trajectories of the states and actions. The construction is standard, see Puterman (2014) and the beginning of the previous section. We define

$$F^\sigma_n(x) := \mathbb{P}^\sigma_\nu(X_n = x), \quad x \in E.$$

Formally, the aim is to solve

$$\sup_{\sigma\in\Pi_N} \Big\{ H(F^\sigma_N) + \mathbb{E}^\sigma_\nu[R_{N-1}] \Big\} \tag{3.1}$$

for a function $H : P(E) \to \mathbb{R}$. Thus, we obtain that the lifted MDP is defined by

(i) $P(E)$ is the state space, where the interpretation of a state $F_n \in P(E)$ at time point n is given as the distribution of X_n. Note here that F_n is ultimately a discrete distribution on the finite set E. In particular we have $F_0 = \nu$.

(ii) $\Pi^M = \{\pi : E \to P(A)\}$ is the action space.

(iii) The one-stage reward is given by $\hat r : P(E) \times \Pi^M \to \mathbb{R}$ with

$$\hat r(F, \pi) := \sum_x \int_A r(x, a)\pi(da|x)F(x).$$

(iv) $H : P(E) \to \mathbb{R}$ is the terminal reward.

(v) The transition function $T : P(E) \times \Pi^M \to P(E)$ is given by

$$T^\pi(F)(x') := T(F, \pi)(x') = \sum_x \int_A q(x'|x, a)\pi(da|x)F(x) \tag{3.2}$$

where $x' \in E$.

Note that a Markov policy in the original model can be linked to a sequence of actions in the lifted MDP (analogously to Prop. 2.2). As mentioned in Remark 2.3 it is well-known that for every policy σ there exists a Markov policy π that induces the same marginal probability measure (see Lemma 2 in Piunovskiy (1997) for general models with no necessarily finite action set). Therefore, we conclude that for $\sigma = (\sigma_n)_{n=0}^{N-1}$ in the original MDP model, there exists an action sequence $\pi = (\pi_n)_{n=0}^{N-1}$ in the lifted MDP such that

$$F^\sigma_N(x) = \mathbb{P}^\sigma_\nu(X_N = x) = T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_0}(F_0)(x), \quad x \in E.$$

Thus, we have that

$$\begin{aligned} & H(F^\sigma_N) + \mathbb{E}^\sigma_\nu[R_{N-1}] \\ = \; & H\Big(T^{\pi_{N-1}} \circ T^{\pi_{N-2}} \circ \ldots \circ T^{\pi_0}(F_0)\Big) + \sum_{k=0}^{N-1} \hat r\Big(T^{\pi_{k-1}} \circ \ldots \circ T^{\pi_0}(F_0), \pi_k\Big) \end{aligned} \tag{3.3}$$

where $T^{\pi_{-1}}(F_0) := F_0$. Thus, the value iteration has the following form:

$$J_N(F) := H(F),$$

$$J_n(F) := \sup_{\pi\in\Pi^M} \left\{\hat{r}(F,\pi) + J_{n+1}(T^\pi(F))\right\}, \quad n = 0,1,\dots,N-1. \qquad (3.4)$$

In this case, we need another condition to ensure the existence of optimal policies:

(C1'): The mapping $F \to H(F)$ is upper semicontinuous, that is, $\limsup_{k\to\infty} H(F_{(k)}) \leq H(F)$ for any sequence $(F_{(k)})$ converging to F in $P(E)$ as $k \to \infty$.

(C2): The mappings $a \to q(x'|x,a)$ and $a \to r(x,a)$ are continuous for all $x, x' \in E$.

Note that $M := \max_{(x,a)\in E\times A} |r(x,a)| < \infty$. Then we obtain:

Theorem 3.1. *Assume (C1') and (C2).*

a) *If (J_n) is computed according to* (3.4) *with T^π as defined in* (3.2)*, then $J_0(F_0)$ is the value function of problem* (3.1)*, i.e., $J_0(F_0) = \sup_{\sigma\in\Pi_N}\left\{H(F_N^\sigma) + \mathbb{E}_\nu^\sigma[R_{N-1}]\right\}$.*

b) *Maximizers $(f_0^*,\dots,f_{N-1}^*)$ in the recursion of* (3.4) *exist, i.e., for $n = 0,1,\dots,N-1$*

$$f_n^*(F) = \operatorname*{argmax}_\pi \left\{\hat{r}(F,\pi) + J_{n+1}(T^\pi(F))\right\},\ F \in P(E).$$

c) *Define the following state-action sequence starting with state $F_0 = \nu$:*

$$\begin{aligned} \pi_0^* &= f_0^*(F_0),\\ F_1 &= T^{\pi_0^*}(F_0),\\ &\vdots\\ \pi_n^* &= f_n^*(F_n),\\ F_{n+1} &= T^{\pi_n^*} \circ \ldots \circ T^{\pi_0^*}(F_0). \end{aligned}$$

Then $(\pi_0^*,\pi_1^*,\dots,\pi_{N-1}^*)$ determines an optimal policy $(\sigma_0^*,\dots,\sigma_{N-1}^*)$ for problem (3.1) *as follows*

$$\sigma_n^*(a|h_n) = \pi_n^*(a|x_n)$$

for $n = 0,1,\dots,N-1$, where $h_n = (x_0,a_0,\dots,x_n)$.

Proof. Part a) can be deduced by backward induction. Indeed note that

$$\begin{aligned} J_{N-1}(F) &= \sup_{\pi_{N-1}} \Big\{\hat{r}(F,\pi_{N-1}) + H(T^{\pi_{N-1}}(F))\Big\} \quad \text{and}\\ J_{N-2}(F) &= \sup_{\pi_{N-2}} \Big\{\hat{r}(F,\pi_{N-2}) + J_{N-1}(T^{\pi_{N-2}}(F))\Big\}\\ &= \sup_{(\pi_{N-2},\pi_{N-1})} \Big\{\hat{r}(F,\pi_{N-2}) + \hat{r}(T^{\pi_{N-2}}(F),\pi_{N-1}) + H(T^{\pi_{N-1}}\circ T^{\pi_{N-2}}(F))\Big\}. \end{aligned}$$

Consequently,

$$\begin{aligned} J_{N-n}(F) &= \sup_{(\pi_{N-n},\dots,\pi_{N-1})} \Big\{\hat{r}(F,\pi_{N-n}) + \sum_{k=1}^{n-1} \hat{r}(T^{\pi_{N-(k+1)}}\circ\ldots\circ T^{\pi_{N-n}}(F),\pi_{N-k})\\ &\qquad + H(T^{\pi_{N-1}}\circ\ldots\circ T^{\pi_{N-n}}(F))\Big\} \quad \text{and}\\ J_0(F) &= \sup_{(\pi_0,\dots,\pi_{N-1})} \Big\{\hat{r}(F,\pi_0) + \sum_{k=1}^{N-1} \hat{r}(T^{\pi_{N-(k+1)}}\circ\ldots\circ T^{\pi_0}(F),\pi_{N-k})\\ &\qquad + H(T^{\pi_{N-1}}\circ\ldots\circ T^{\pi_0}(F))\Big\}. \end{aligned}$$

Now observe that the expression in the curly brackets equals the right-hand side in (3.3). This proves part a). For parts b) and c) we proceed as follows. Let $(F_{(k)},\pi_{(k)}) \to (F,\pi)$ in

$P(E) \times \Pi^M$ as $k \to \infty$. Recall that it means that $F_{(k)}(x) \to F(x)$ and $\pi_{(k)}(\cdot|x) \to \pi(\cdot|x)$ weakly in $P(A)$ for every $x \in X$. We show that

$$\hat{r}(F_{(k)}, \pi_{(k)}) \to \hat{r}(F, \pi) \quad \text{and} \quad T(F_{(k)}, \pi_{(k)}) \to T(F, \pi)$$

as $k \to \infty$. Indeed, notice that

$$\begin{aligned}
&\left|\hat{r}(F_{(k)}, \pi_{(k)}) - \hat{r}(F, \pi)\right| = \\
&\left|\sum_x \int_A r(x,a)\pi_{(k)}(da|x)F_{(k)}(x) - \sum_x \int_A r(x,a)\pi(da|x)F(x)\right| \\
&\le \sum_x \left|\int_A r(x,a)\pi_{(k)}(da|x) - \int_A r(x,a)\pi(da|x)\right| F_{(k)}(x) \\
&+ \sum_x \int_A r(x,a)\pi(da|x) \left|F_{(k)}(x) - F(x)\right| \\
&\le \sum_x \left|\int_A r(x,a)\pi_{(k)}(da|x) - \int_A r(x,a)\pi(da|x)\right| \\
&+ M \sum_x \left|F_{(k)}(x) - F(x)\right|.
\end{aligned}$$

Hence, the first term tends to zero by (C2) and the second term converges to zero by definition. In the same manner, we show that $T(F_{(k)}, \pi_{(k)}) \to T(F, \pi)$. Clearly, π^M is compact, since $\Pi^M = P(A) \times \cdots \times P(A)$ ($|E|$ times). Hence, the remaining parts follow as before from Proposition 2.4.3 in Bäuerle and Rieder (2011). □

4. Infinite Horizon Problems

So far we have considered problems with a finite time horizon. Let us now briefly turn to the situation with an infinite time horizon in the setting of Section 2. Hence, we deal with the same problem as in Section 2 with the difference that we introduce a discount factor $\beta \in (0, 1)$ and set $N = \infty$. A policy σ is a sequence $(\sigma_0, \sigma_1, \ldots)$ of history dependent decision rules $\sigma_n : H_n \to P(A)$. By Π_∞ we denote the set of all policies. The random reward we are interested in is now given by

$$R_\infty := \lim_{n\to\infty} R_n, \qquad \text{where} \quad R_n := \sum_{k=0}^{n} \beta^k r(X_k, A_k).$$

Clearly, R_∞ is well-defined and $R_\infty \in [-M/(1-\beta), M/(1-\beta)]$, where that $M = \max_{(x,a)\in E\times A} |r(x,a)|$. Let us consider the problem of maximizing the expected discounted reward when the initial distribution is ν, that is,

$$\sup_{\sigma\in\Pi_\infty} \mathbb{E}_\nu^\sigma[R_\infty].$$

As in the previous section $\mathbb{E}_\nu^\sigma$ stands for the expectation operator that refers to the probability measure $\mathbb{P}_\nu^\sigma$ defined uniquely on $H_\infty := (E \times A)^{\mathbb{N}}$ with σ-algebra generated by the cylinder sets, see Puterman (2014). It is well-known that the problem can be solved with the help of a fixed point equation and an optimal policy can be found among stationary deterministic Markovian policies, i.e., an optimal policy σ^* is of the form $\sigma^* = (\tilde{\sigma}, \tilde{\sigma}, \ldots)$, where $\tilde{\sigma} : E \to A$. Now for $\sigma \in \Pi_\infty$ and $B \in \mathcal{B}(\mathbb{R})$ define

$$F_\infty^\sigma(B) = \mathbb{P}_\nu^\sigma(R_\infty \in B)$$

as the distribution of R_∞ under policy σ, when the initial distribution is given by ν. Assume that $H : P(\mathbb{R}) \to \mathbb{R}$ is an arbitrary functional. The aim is to solve

$$\sup_{\sigma \in \Pi_\infty} V(\sigma), \text{ where } V(\sigma) := H(F_\infty^\sigma). \tag{4.1}$$

In this situation, it is in general not true that an optimal policy can be found in the set of stationary deterministic Markovian policies as the following example shows.

Example 4.1. Consider the degenerate Markov decision process given by: $E = \{0\}$, $A = \{0, 1\}$ and $r(x, a) = \frac{1}{2}a$, i.e., the decision alone determines the reward. Thus, the transition probabilities are given by $q(0|0, a) = 1$. Further assume that $\beta = 1/2$. This implies that R_∞ is deterministic given by

$$R_\infty = \frac{1}{2} \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^k a_k,$$

where $(a_k)_{k=0}^\infty$ is the sequence of chosen actions. Hence, $R_\infty \in [0, 1]$ and we can obtain any number in $[0, 1]$ by choosing $(a_k)_{k=0}^\infty$ accordingly. Now fix $B := \{\sqrt{2}/2\}$ and let $\sigma \in \Pi_\infty$. For $F_\infty^\sigma \in P([0, 1])$ define the function $H(F_\infty^\sigma) := F_\infty^\sigma(B) = \mathbb{P}_{\delta_0}^\sigma(R_\infty = \sqrt{2}/2)$. Clearly, $\max_{\sigma \in \Pi_\infty} H(F_\infty^\sigma) = 1$ for σ^* represented by the actions $(a_0, a_1 \ldots)$ being the dyadic representation of $\sqrt{2}/2$. But $(a_0, a_1 \ldots)$ is an infinite, non-periodic sequence. Thus, there is no hope for any optimal stationary policy. Contrary, one can study the distribution of R_∞ under the restriction of stationary deterministic Markovian policies. In such setting it makes sense to discuss the existence and properties of solutions to the distributional fixed point equation $R \overset{d}{=} X + \beta R$, see e.g., Gerstenberg et al. (2023).

Here, we consider next the question of existence of optimal policies for (4.1) and their approximation. We denote by $H = \cup_{n \in \mathbb{N}} H_n$ the set of all histories. Since E and A are finite, H is countable. Therefore, $\Pi_\infty = P(A)^H = \{\sigma : H \to P(A)\}$. By Tychonoff's theorem this set is compact in the product topology. Note that the mapping $\sigma \to \mathbb{P}_\nu^\sigma$ is continuous when the set of strategic measures is equipped with the weak topology.

We proceed with the following assumption

(C3): The mapping H is Lipschitz-continuous w.r.t. the Wasserstein W_1-distance, i.e., there exists $K > 0$ such that for all $F, G \in P(\mathbb{R})$

$$|H(F) - H(G)| \leq K \cdot W_1(F, G).$$

Note that (C3) implies that H is continuous w.r.t. weak convergence. This follows since weak convergence plus the convergence of the first moments is equivalent to convergence in the W_1-topology, see Theorem 6.9 in Villani (2008). Define for $\sigma \in \Pi_\infty$ by F_N^σ the distribution of R_N under $\mathbb{P}_\nu^\sigma$:

$$F_N^\sigma = \mathbb{P}_\nu^\sigma(R_N \in \cdot)$$

In this case, only the first N decision rules of the sequence are important. Let

$$V_N(\sigma) := H(F_N^\sigma), \quad V_N := \sup_{\sigma \in \Pi_\infty} H(F_N^\sigma). \tag{4.2}$$

Further, let

$$\begin{aligned} A_N &:= \{\sigma \in \Pi_\infty : V_N(\sigma) = \sup_{\sigma' \in \Pi_\infty} H(F_N^{\sigma'})\}, \\ A_\infty &:= \{\sigma \in \Pi_\infty : V(\sigma) = \sup_{\sigma' \in \Pi_\infty} H(F_\infty^{\sigma'})\}. \end{aligned}$$

Note that under (C3) and our discussion the sets A_N, A_∞ are non-empty. We obtain the following result where we use the definition

$$LsA_N := \{\sigma \in \Pi_\infty : \sigma \text{ is an accumulation point of a sequence } (\sigma^N) \text{ with } \sigma^N \in A_N\}.$$

Theorem 4.2. *Under (C3) we obtain:*

a) *The infinite horizon problem can be approximated by the finite horizon problems*

$$\lim_{N\to\infty} \sup_{\sigma\in\Pi_\infty} H(F_N^\sigma) = \sup_{\sigma\in\Pi_\infty} \lim_{N\to\infty} H(F_N^\sigma) = \sup_{\sigma\in\Pi_\infty} H(F_\infty^\sigma).$$

b) *There exists an optimal policy* $\sigma^* = (\sigma_0^*, \sigma_1^*, \dots) \in \Pi_\infty$ *for problem* (4.1) *and* $\emptyset \neq LsA_N \subset A_\infty$.

Proof. Parts a), b): Note that we have for fixed $\sigma \in \Pi_\infty$ and $n \geq m$:

$$H(F_n^\sigma) - H(F_m^\sigma) \leq K W_1(F_n^\sigma, F_m^\sigma) \leq K \frac{\beta^{m+1} M}{1-\beta},$$

where $\frac{\beta^{m+1}M}{1-\beta} \to 0$ for $m \to \infty$. The latter inequality follows since

$$W_1(F_n^\sigma, F_m^\sigma) \leq \mathbb{E}_\nu^\sigma |R_n - R_m| = \mathbb{E}_\nu^\sigma \left(\sum_{k=m+1}^{n} \beta^k |r(X_k, A_k)| \right) \leq \frac{\beta^{m+1} M}{1-\beta}.$$

Thus, the results are implied by Theorem A.1.5 in Bäuerle and Rieder (2011). □

Theorem 4.2 states that the value of the infinite horizon problem $\sup_{\sigma\in\Pi_\infty} H(F_\infty^\sigma)$ can be approximated by the value of the finite horizon problem $\sup_{\sigma\in\Pi_\infty} H(F_N^\sigma)$ for large N. This is not too surprising since $\beta \in (0,1)$ and the tail of the reward vanishes. Since the initial distribution ν if fixed, by Theorem 2 in Derman and Strauch (1966), the policy σ^* can be replaced by a Markov policy in the original MDP model.

5. Special Cases and Applications

We now discuss some special choices for H which are meaningful. In particular, by setting H in the right way we rediscover some well-known results in the literature. In order to simplify the presentation, we use the finite state-action framework, i.e., we can work with sums instead of integrals.

5.1. **Classical MDP.** In the classical MDP theory, we want to maximize $\mathbb{E}_{x_0}^\sigma R_N$ which is obtained when we choose

$$H(F) = \sum_{x,s} (g(x) + s) F(x,s).$$

We show next that our general algorithm in the discrete case from (2.5) yields the established Bellman optimality equation as a special case. Let us briefly recall the standard theory and denote by $V_n : E \to \mathbb{R}$ the value function in the classical case. Set $V_N = g$ and due to (2.1)

$$V_n(x) = \max_a \big\{ r(x,a) + \sum_{x'} V_{n+1}(x') q(x'|x,a) \big\}. \tag{5.1}$$

The interpretation is

$$V_n(x) = \sup_\sigma \mathbb{E}_x^\sigma \Big[\sum_{k=n}^{N-1} r(X_k, A_k) + g(X_N) \Big],$$

i.e., V_n is the maximal expected reward of the system from time n onwards when we start at time n in state x. We begin with investigating our algorithm at time point $N-1:$

$$\begin{aligned}
J_{N-1}(F) &= \sup_{\pi\in\Pi^M} H(T^\pi(F))\\
&= \sup_{\pi\in\Pi^M} \sum_{x',s'} (g(x')+s')T^\pi(F)(x',s')\\
&= \sup_{\pi\in\Pi^M} \sum_{x',s'} (g(x')+s') \sum_{(x,s,a):r(x,a)=s'-s} q(x'|x,a)\pi(a|x,s)F(x,s)\\
&= \sup_{\pi\in\Pi^M} \sum_{x'} \sum_{(x,s,a)} (g(x')+s+r(x,a))q(x'|x,a)\pi(a|x,s)F(x,s)\\
&= \sup_{\pi\in\Pi^M} \Big\{ \sum_{x'} \sum_{(x,s,a)} (g(x')+r(x,a))q(x'|x,a)\pi(a|x,s)F(x,s)\\
&\quad + \sum_{x'} \sum_{(x,s,a)} sq(x'|x,a)\pi(a|x,s)F(x,s) \Big\}\\
&= \sup_{\pi\in\Pi^M} \sum_{x'} \sum_{(x,s,a)} (g(x')+r(x,a))q(x'|x,a)\pi(a|x,s)F(x,s) + \sum_s sF(E,s)\\
&= \sum_s sF(E,s) + \sup_{\pi\in\Pi^M} \sum_{x'} \sum_{(x,s,a)} (g(x')+r(x,a))q(x'|x,a)F(x,s)\pi(a|x,s)\\
&= \sum_s sF(E,s) + \sum_{x,s} F(x,s) \sup_a \sum_{x'} (g(x')+r(x,a))q(x'|x,a)\\
&= \sum_s sF(E,s) + \sum_x F(x,S) \sup_a \sum_{x'} (g(x')+r(x,a))q(x'|x,a)\\
&= \sum_s sF(E,s) + \sum_x F(x,S) \sup_a \{r(x,a) + \sum_{x'} g(x')q(x'|x,a)\}.
\end{aligned}$$

From this observation we obtain the following conjecture for time point n:

$$J_n(F) = \sum_s sF(E,s) + \sum_x F(x,S)V_n(x),$$

where V_n is the value function of the classical Bellman equation in (5.1). We prove this conjecture by induction over the time horizon. For $n=N$ we obtain:

$$J_N(F) = H(F) = \sum_{x,s}(g(x)+s)F(x,s) = \sum_s sF(E,s) + \sum_x g(x)F(x,S).$$

For $n = N-1$ the statement is what we computed before. Now suppose the statement is true for $N, N-1, \ldots, n+1$. We show that it is also true for time point n.

$$\begin{aligned}
J_n(F) &= \sup_{\pi\in\Pi^M} J_{n+1}(T^\pi(F))\\
&= \sup_{\pi\in\Pi^M} \Big\{ \sum_{s'} s'T^\pi(F)(E,s') + \sum_{x'} T^\pi(F)(x',S)V_{n+1}(x') \Big\}\\
&= \sup_{\pi\in\Pi^M} \Big\{ \sum_{s',x'} s'T^\pi(F)(x',s') + \sum_{s',x'} T^\pi(F)(x',s')V_{n+1}(x') \Big\}\\
&= \sup_{\pi\in\Pi^M} \Big\{ \sum_{s',x'} s' \sum_{(x,s,a):r(x,a)=s'-s} q(x'|x,a)\pi(a|x,s)F(x,s)
\end{aligned}$$

$$\begin{aligned}
&+\sum_{s',x'}\sum_{(x,s,a):r(x,a)=s'-s} q(x'|x,a)\pi(a|x,s)F(x,s)V_{n+1}(x')\Big\}\\
&= \sup_{\pi\in\Pi^M}\Big\{\sum_{x'}\sum_{(x,s,a)}(r(x,a)+s)q(x'|x,a)\pi(a|x,s)F(x,s)\\
&\quad+\sum_{x'}\sum_{(x,s,a)} q(x'|x,a)\pi(a|x,s)F(x,s)V_{n+1}(x')\Big\}\\
&= \sup_{\pi\in\Pi^M}\Big\{\sum_{(x,s,a)} r(x,a)\pi(a|x,s)F(x,s)+\sum_{(x,s)} sF(x,s)\\
&\quad+\sum_{x'}\sum_{(x,s,a)} q(x'|x,a)\pi(a|x,s)F(x,s)V_{n+1}(x')\Big\}\\
&= \sum_s sF(E,s)+\sup_{\pi\in\Pi^M}\sum_{(x,s,a)}\pi(a|x,s)F(x,s)\{r(x,a)+\sum_{x'} q(x'|x,a)V_{n+1}(x')\}\\
&= \sum_s sF(E,s)+\sum_{(x,s)} F(x,s)\sup_a\{r(x,a)+\sum_{x'} q(x'|x,a)V_{n+1}(x')\}\\
&= \sum_s sF(E,s)+\sum_{(x,s)} F(x,s)V_n(x)=\sum_s sF(E,s)+\sum_x F(x,S)V_n(x).
\end{aligned}$$

Thus, when we look at the last two lines, it is possible to recover the classical Bellman equation in our algorithm. In particular, the optimal strategy does not depend on F and is deterministic. The interpretation of $J_n(F_n)$ for $F_n = T^{\pi^*_{n-1}} \circ \ldots \circ T^{\pi^*_0}(F_0)$ defined in Theorem 2.4 with $F_0 = \delta_{x_0} \otimes \delta_0$ is as follows:

$$\begin{aligned}
J_n(F_n) \quad &= \quad \sup_{(\sigma_0,\ldots,\sigma_{n-1})} \mathbb{E}_{x_0}^{(\sigma_0,\ldots,\sigma_{n-1})}\Big[\sum_{k=0}^{n-1} r(X_k,A_k)\Big] + \\
&\quad \sum_x F_n(x,S) \sup_{(\sigma_n,\ldots,\sigma_{N-1})} \mathbb{E}_x^{(\sigma_n,\ldots,\sigma_{N-1})}\Big[\sum_{k=n}^{N-1} r(X_k,A_k)+g(X_N)\Big].
\end{aligned}$$

Hence, the algorithm separates at time point n the future from the past.

5.2. **The case of quantile optimization.** We are interested in the problem

$$\sup_\sigma \mathbb{P}_\nu^\sigma(R_N \geq t)$$

for a fixed $t \in \mathbb{R}$. In other words, we want to maximize the probability that the accumulated reward exceeds the threshold t. Hence, we can write the optimization criteria as

$$\mathbb{P}_\nu^\sigma(R_{N-1}+g(X_N)\geq t)=\sum_{x,s} F_N^\sigma(x,s)V_N(x,s),$$

where

$$V_N(x,s) := \begin{cases} 1 & \text{if } s+g(x)\geq t\\ 0 & \text{else.}\end{cases}$$

Problems like this have been investigated in Filar et al. (1995); Wu and Lin (1999); Bäuerle and Ott (2011); Chow et al. (2015); Gilbert et al. (2017); Li et al. (2022). It is known that the value for $\nu = \delta_{x_0}$ is given by $V_0(x_0, 0)$ where V_0 can be computed from the recursion

$$V_n(x,s) = \sup_a \sum_{x'} V_{n+1}(x', s + r(x,a))q(x'|x,a), \quad n = 0, 1, \ldots, N-1. \tag{5.2}$$

Let us consider our algorithm again at time point $N-1$:

$$\begin{aligned}
J_{N-1}(F) &= \sup_{\pi \in \Pi^M} H(T^\pi(F)) \\
&= \sup_{\pi \in \Pi^M} \sum_{x',s'} V_N(x',s')T^\pi(F)(x',s') \\
&= \sup_{\pi \in \Pi^M} \sum_{x',s'} V_N(x',s') \sum_{(x,s,a):r(x,a)+s=s'} q(x'|x,a)\pi(a|x,s)F(s,x) \\
&= \sup_{\pi \in \Pi^M} \sum_{x'} \sum_{x,s,a} V_N(x', s + r(x,a))q(x'|x,a)\pi(a|x,s)F(s,x) \\
&= \sum_{x,s} F(s,x) \sup_{\pi \in \Pi^M} \sum_{x'} \sum_{a} V_N(x', s + r(x,a))q(x'|x,a)\pi(a|x,s) \\
&= \sum_{x,s} F(s,x) \sup_a \sum_{x'} V_N(x', s + r(x,a))q(x'|x,a) =: \sum_{x,s} F(s,x)V_{N-1}(x,s).
\end{aligned}$$

This gives rise to the following conjecture:

$$J_n(F) = \sum_{x,s} F(x,s)V_n(x,s),$$

where V_n is given in (5.2). Here we start with V_N defined above. For $n = N$, the statement is true by definition of V_N. Now suppose the statement is true for $N, N-1, \ldots, n+1$. We show that it is also true for the time point n.

$$\begin{aligned}
J_n(F) &= \sup_\pi J_{n+1}(T^\pi(F)) \\
&= \sup_{\pi \in \Pi^M} \sum_{x',s'} V_{n+1}(x',s')T^\pi(F)(x',s') \\
&= \sup_{\pi \in \Pi^M} \sum_{s',x'} V_{n+1}(x',s') \sum_{(x,s,a):r(x,a)=s'-s} q(x'|x,a)\pi(a|x,s)F(x,s) \\
&= \sup_{\pi \in \Pi^M} \sum_{x'} \sum_{x,s,a} V_{n+1}(x', s + r(x,a))q(x'|x,a)\pi(a|x,s)F(s,x) \\
&= \sum_{x,s} F(s,x) \sup_{\pi \in \Pi^M} \sum_{x'} \sum_{a} V_{n+1}(x', s + r(x,a))q(x'|x,a)\pi(a|x,s) \\
&= \sum_{x,s} F(s,x) \sup_a \sum_{x'} V_{n+1}(x', s + r(x,a))q(x'|x,a) = \sum_{x,s} F(s,x)V_n(x,s),
\end{aligned}$$

which completes the induction step. Thus, we can see that the optimal strategy does not depend on F and is deterministic. The recursion reduces to the recursion for V_n which appears in a similar way in Wu and Lin (1999); Bäuerle and Ott (2011). The interpretation for V_n is

$$V_n(x,s) = \sup_\sigma \mathbb{P}^\sigma_\nu(R_{N-1} + g(X_N) \geq t | X_n = x, R_{n-1} = s).$$

5.3. **Optimal transport.** We next consider a non-standard application. We want to determine the transition probabilities of a random walk in such a way that at a fixed time point a given distribution is best approximated in a certain metric, and at the same time the cost of transporting the probability mass is minimal. The theoretical framework is that of Section 3 with some minor extensions. Instead of maximizing a reward, we now formulate the problem as one of minimizing cost. This can be considered as an optimal transport problem. The optimal transport problem was first discussed by Monge (1781) and then refined by Kantorovich (1948), see Kantorovich (2006) for a reprint of the original article. A lot of researchers considered the optimal transport problem and many important results were established, see for example Rachev and Rüschendorf (2006b,a) or Villani (2008) for monographs on the topic. The optimal transport between discrete-time stochastic processes has connections to dynamic programming, see Terpin et al. (2024); Backhoff et al. (2017); Moulos (2021). But we will pursue a different direction here. To be more specific, we consider the following data. Let $N \in \mathbb{N}$ be a fixed time horizon, G be a given distribution on E. The terminal cost function $H : P(E) \to \mathbb{R}$ is given by

$$H(F) := \int_{\mathbb{R}} |F_c(t) - G_c(t)| dt,$$

which is the Wasserstein W_1 distance between G and the distribution F of the terminal state X_N of a controlled random walk. Here, F_c and G_c are cumulative distributions of F and G, respectively. More precisely, we define the approximating MDP (random walk) by

(i) $E = \{1, \dots, K\}$, $K \in \mathbb{N}$, is the state space,
(ii) $A = \{(a^1, a^2) \in [0,1]^2 : a^1 + a^2 \leq 1\}$ is the action space,
(iii) the transition probability from $E \times A$ to E for $x \in \{2, \dots, K-1\}$ is given by

$$q(x'|x, a^1, a^2) = \begin{cases} a^1, & x' = x+1, \\ a^2, & x' = x-1, \\ 1 - a^1 - a^2, & x' = x, \end{cases}$$

and

$$q(x'|0, a^1, a^2) = \begin{cases} a^1, & x' = 1, \\ 1 - a^1, & x' = 0, \end{cases} \qquad q(x'|K, a^1, a^2) = \begin{cases} a^2, & x' = K-1, \\ 1 - a^2, & x' = K, \end{cases}$$

(iv) $r_n(x, a^1, a^2) = c_n(a_1 + a_2)$ where $0 < c_0 \leq c_1 \leq \ldots \leq c_N \leq 1$ are constants.

The transition law is that for a random walk (X_n) which can get from state x only in state $x+1$ or $x-1$ or stay in x. The corresponding probabilities a^1, a^2 are subject to the decision. The choice of a^1, a^2 incurs some cost. Note that we have a non-stationary problem here and the transportation cost is non-decreasing in time. Thus, early transport is cheaper than later transport. The objective function is then

$$\inf_{\sigma \in \Pi_N} \Big\{ H(F_N^\sigma) + \mathbb{E}_\nu^\sigma [R_{N-1}] \Big\}, \tag{5.3}$$

where $R_{N-1} = \sum_{k=0}^{N-1} c_k (A_k^1 + A_k^2)$ with $A_k = (A_k^1, A_k^2)$. This is a weighted objective of distance to the desired distribution and expected transportation costs. Note that the assumptions (C1) and (C2) of Section 2 are satisfied and that the action space is compact. Thus, (X_n) is a random walk on E, where the up and down probabilities can be chosen and may depend on the history of the process. This kind of processes are called 'elephant random walk' (see Gut and Stadtmüller (2021)), since elephants have a long memory. In line with Section 3, it is here enough to consider the recursion for the distribution of the first component X_n, hence we write $F_n^\sigma(x) = \mathbb{P}_\nu^\sigma(X_n = x)$, $x \in E$. Recall that $F_0^\sigma = \nu$. Since the costs and transition functions have special structure, i.e., they are linear in the action variables, it is sufficient to consider the set of non-randomized actions in the lifted MDP, that is, the set of all mappings $\pi : E \to A$. With a little abuse of notation we denote the set by the same symbol Π^M as in the previous sections. By a slight extension of Theorem 3.1, the value functions in this application are for $n = 0, 1, \dots, N-1$ given by

$$\begin{aligned} J_N(F) &:= H(F), \\ J_n(F) &:= \inf_{\pi \in \Pi^M} \left\{ \hat{r}_n(F, \pi) + J_{n+1}(T^\pi(F)) \right\}, \end{aligned} \tag{5.4}$$

where

$$\begin{aligned} \hat{r}_n(F, \pi) &= \sum_x r_n(x, \pi(x)) F(x) \\ &= c_n \sum_x F(x)(\bar{a}^1(x) + \bar{a}^2(x)), \end{aligned}$$

with $\pi(x) = (\bar{a}^1(x), \bar{a}^2(x))$. In particular, $\pi(1) = (1, 0)$ and $\pi(K) = (0, 1)$ (i.e., $\bar{a}^2(1) = \bar{a}^1(K) = 0$) and

$$\begin{aligned} T^\pi(F)(x') &= \sum_x q(x'|x, \pi(x)) F(x) = \sum_x q(x'|x, \bar{a}^1(x), \bar{a}^2(x)) F(x) \\ &= \bar{a}^1(x'-1) F(x'-1) + \bar{a}^2(x'+1) F(x'+1) + (1 - \bar{a}^1(x') - \bar{a}^2(x')) F(x'). \end{aligned}$$

We start with a given distribution F_0 on E and have to compute $J_0(F_0)$. The value function $J_n(F)$ describes the minimal sum of the expected transportation cost and remaining distance of a terminal distribution of the random walk to the target, starting at time n till the terminal time N.

Remark 5.1 (Connection to classical optimal transport)**.** Note that when all $c_k \equiv 1$, then $J_0(F_0)$ is exactly the Wasserstein distance to the distribution G, also known as the earth-mover distance. In order to see this, recall that the Wasserstein W_1 distance between two discrete distributions

$$\mu = \sum_{j=1}^K \omega_j \delta_j \quad \text{and} \quad \nu = \sum_{i=1}^K \vartheta_i \delta_i,$$

where $\omega_j, \vartheta_i \geq 0$, for all $i, j = 1, \dots, K$ and $\sum_{j=1}^K \omega_j = \sum_{i=1}^K \vartheta_i = 1$ can be computed from the following linear program (see Kantorovich (2006))

$$\begin{aligned} \min \quad & \sum_{j=1}^K \sum_{i=1}^K q_{j,i} |j - i| \\ \text{s.t.} \quad & \sum_{i=1}^K q_{j,i} = \omega_j, \quad j = 1, \dots, K, \\ & \sum_{j=1}^K q_{j,i} = \vartheta_i, \quad i = 1, \dots, K, \\ & q_{j,i} \geq 0, \quad j = 1, \dots, K, \ i = 1, \dots, K. \end{aligned} \tag{P}$$

This linear program can be interpreted as a problem of transporting probability mass from the distribution μ to the distribution ν with minimal cost, where the cost is given as the sum of single masses times transportation distance. In our problem formulation we can - within one step - only transport mass from one state to its neighbors, i.e., from 4 to 3 and 5, but not to 6. However, obviously for a given transport $q = q_{j,i} > 0$ in (P) with $j < i$ we can split the transport in $i-j$ transports of the form $q_{j,j+1} = q, q_{j+1,j+2} = q, \ldots, q_{i-1,i} = q$. The cost will be identical, because $q_{j,i}|j-i| = q_{j,j+1} + q_{j+1,j+2} + \ldots + q_{i-1,i}$. Thus, consider for a moment only transports with the additional restriction $q_{j,i} = 0$ for $i \notin \{j+1, j-1\}$. The corresponding cost of such a transport can be expressed by $\sum_{j=0}^{K-1} q_{j,j+1} + \sum_{j=1}^{K} q_{j,j-1}$. Now consider one time step of our model, say from n to $n+1$ and let $X_n \sim \mu$ and $X_{n+1} \sim \nu$. We can realize the same transport in our model by choosing $\bar{a}^1(j) = \frac{q_{j,j+1}}{\omega_j}$ and $\bar{a}^2(j) = \frac{q_{j,j-1}}{\omega_j}$. Note that the (conditional) transition probabilities implied by a transport $(q_{j,i})$ in (P) are given by $\mathbb{P}(X_{n+1} = i | X_n = j) = \frac{q_{j,i}}{w_j}$. The one-step cost of the transport if $c_k \equiv 1$ is

$$\mathbb{E}(A_n^1 + A_n^2) = \sum_{j=1}^{K} \omega_j \left(\frac{q_{j,j+1}}{\omega_j} + \frac{q_{j,j-1}}{\omega_j}\right) = \sum_{j=0}^{K-1} q_{j,j+1} + \sum_{j=1}^{K} q_{j,j-1}$$

and hence the cost are the same in both models. Thus, the optimal transport of (P) can be implemented by decomposing it in neighboring transports and the terminal cost given by the Wasserstein distance to the target. The remaining cost is exactly given by the Wasserstein distance in the end.

However, this argument breaks if $c_k \not\equiv 1$. In our problem formulation there is a time aspect: It is cheaper to move mass early. In particular if $K = 4$ and $F_0 = (\frac{1}{2}, 0, 0, \frac{1}{2})$ and $G = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ and $N \geq 2$, then it is easy to see that $W_1(F_0, G) = 1$ and $J_0(F_0) = \frac{1}{2}(c_1 + c_2)$. However, if $F_0 = (0, \frac{1}{2}, 0, \frac{1}{2})$ and $G = (\frac{1}{2}, 0, \frac{1}{2}, 0)$ then again $W_1(F_0, G) = 1$, but $J_0(F_0) = c_1$. Hence, we get the same W_1- distance but different results in our optimization problem.

Theorem 5.2. *The following statements hold for the optimal policy:*

a) *If at time point n not all mass at x is moved, then xw ill only receive mass at later time points and stay a sink. Formally, for any $x \in E$*

$$\bar{a}_n^1(x) + \bar{a}_n^2(x) < 1 \quad \Rightarrow \quad \bar{a}_m^1(x) + \bar{a}_m^2(x) = 0, \, m \geq n+1.$$

b) *A mass which has already been moved will keep its direction of moving.*

Proof. Part a): Assume that $\bar{a}_n^1(x) + \bar{a}_n^2(x) < 1$ but $\bar{a}_m^1(x) + \bar{a}_m^2(x) > 0$ for some $m \geq n+1$. Note that the cost from the coordinate x equals

$$c_n F_n(x)(\bar{a}_n^1(x) + \bar{a}_n^2(x))$$

and $F_{n+1}(x) \geq F_n(x)(1 - \bar{a}_n^1(x) - \bar{a}_n^2(x))$. Let $m \geq n+1$ be the smallest number for which $a_m^1(x) + a_m^2(x) > 0$, which implies that $F_m(x) \geq F_{n+1}(x)$. Suppose that it is optimal in period m is to transfer at least the mass $F_n(x)(\alpha_1 \bar{a}_n^1(x) + \alpha_2 \bar{a}_n^2(x))$, where $\alpha_1, \alpha_2 \in [0, 1)$ are numbers such that

$$0 < \alpha_1 \bar{a}_n^1(x) + \alpha_2 \bar{a}_n^2(x) \leq 1 - \bar{a}_n^1(x) - \bar{a}_n^2(x).$$

The cost from x in period m is

$$c_m F_m(x)(\bar{a}_m^1(x) + \bar{a}_m^2(x)) \geq c_m F_n(x)(\alpha_1 \bar{a}_n^1(x) + \alpha_2 \bar{a}_n^2(x)).$$

But if the mass were transported at n, then the cost would be

$$c_n F_n(x)(\alpha_1 \bar{a}_n^1(x) + \alpha_2 \bar{a}_n^2(x)) \leq c_m F_n(x)(\alpha_1 \bar{a}_n^1(x) + \alpha_2 \bar{a}_n^2(x)).$$

Hence, it is not optimal to transfer the mass $F_n(x)(\alpha_1\bar{a}_n^1(x)+\alpha_2\bar{a}_n^2(x))$ at period m. It is cheaper to transfer it at period n. Then, $\bar{a}_n^1(x)$ and $\bar{a}_n^2(x)$ would not be optimal. Therefore, it must hold $a_m^1(x)=a_m^2(x)=0$ for $m\geq n+1$.

Even if x receives new mass at later time points this will not be moved, because without loss of generality we can assume that older mass is moved first.

Part b): Moving a mass in one direction and back at a later time results in the same distribution and just produces costs. Hence, this cannot be optimal. □

Note. that the optimal policy does not depend on the precise value of the variables c_n. However, the assumption $0<c_0\leq c_1\leq\ldots\leq c_N\leq 1$ implies the structural properties in Theorem 5.2. From these observations, we obtain the following algorithm: Suppose the state (distribution) at time n is given by $(F_n(1),\ldots,F_n(K))$ and $(G(1),\ldots,G(K))$ is the probability mass function of G. Define at point $x\in E$:

$$\Delta F_n^l(x):=\sum_{j=1}^{x-1}F_n(j)-\sum_{j=1}^{x-1}G(j),\quad \Delta F_n^u(x):=\sum_{j=x+1}^{K}F_n(j)-\sum_{j=x+1}^{K}G(j).$$

If $\Delta F_n^l(x)<0$, then in comparison to G mass is missing left of x, if $\Delta F_n^l(x)=0$, then in comparison to G the mass left of x is sufficient, if $\Delta F_n^l(x)>0$, then in comparison to G there is too much mass left of x. The expression $\Delta F_n^u(x)$ has a similar explanation concerning the mass right of x.

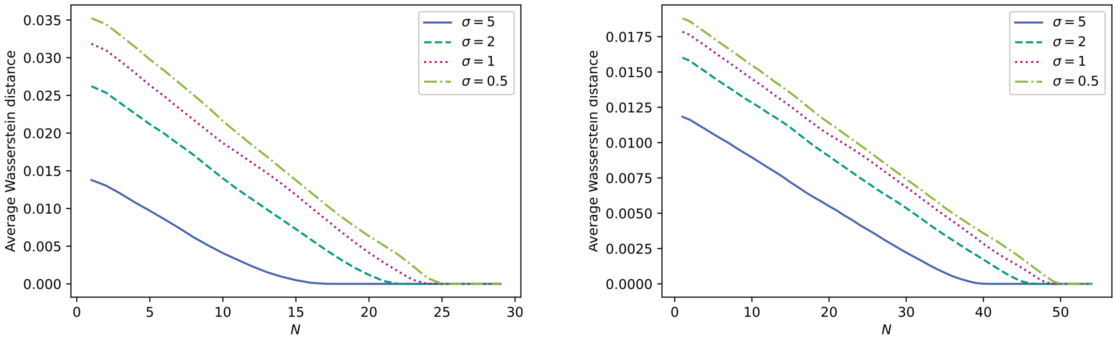
The optimal amount of mass which is moved can now be determined locally at each point $x\in E$, just by knowing $\Delta F_n^l(x)$ and $\Delta F_n^u(x)$. At every stage n, we have to go through all points $x\in E$ and do the following:

Algorithm 1 Optimal mass transportation algorithm

```
Require: $n \geq 0$ and $(F_n(1),\ldots,F_n(K))$
  while $n < N$ do for all $x \in E$
    if $\Delta F_n^l(x) \geq 0$ and $\Delta F_n^u(x) \geq 0$ then
      $a_n^1(x) = a_n^2(x) = 0$.
    else if $\Delta F_n^l(x) \geq 0$ and $\Delta F_n^u(x) < 0$ then
      $a_n^1(x) = \min\{F_n(x), -\Delta F_n^u(x)\}$.
      $a_n^2(x) = 0$
    else if $\Delta F_n^l(x) < 0$ and $\Delta F_n^u(x) \geq 0$ then
      $a_n^2(x) = \min\{F_n(x), -\Delta F_n^l(x)\}$
      $a_n^1(x) = 0$
    else if $\Delta F_n^l(x) < 0$ and $\Delta F_n^u(x) < 0$ then
      $a_n^1(x) = -\Delta F_n^u(x)$
      $a_n^2(x) = -\Delta F_n^l(x)$.
    end if
    Update $F_{n+1}$
    $n \leftarrow n + 1$
  end while
```

In the situation of the last case ($\Delta F_n^l(x)<0$ and $\Delta F_n^u(x)<0$) it follows that $F_{n+1}(x)=G(x)$ and there automatically will be enough mass at point x to realize the move.

Also note that this situation can be extended to approximating a stochastic process in continuous time by a Markov chain such that the distance between the marginal distributions is minimized.



(a) $K = 50$

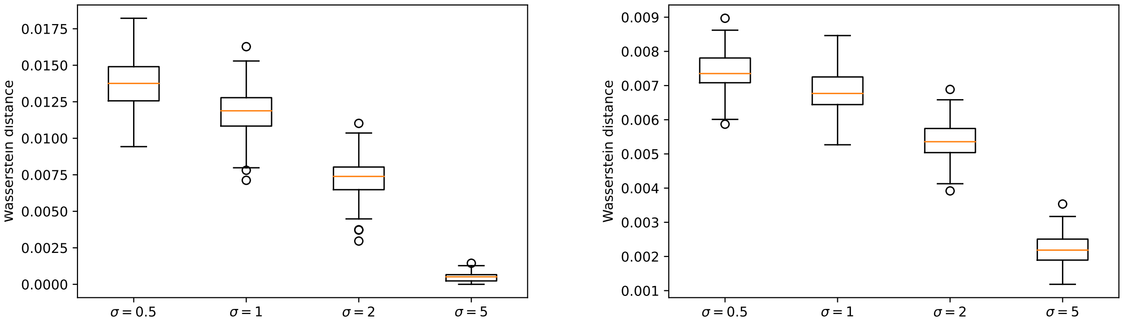
(b) $K = 100$

FIGURE 1. Average Wasserstein distance of F_N from Algorithm 1 and the target distribution G (rescaled normal distribution) over $m = 100$ initial samples for $\sigma \in \{0.5, 1, 2, 5\}$ and $K \in \{50, 100\}$.

Let us conclude this special case by illustrating some numerical results obtained from implementing Algorithm 1. First, we used a rescaled normal distribution on $E = \{1, \ldots, K\}$ as the target distribution, i.e., $G = (G(1), \ldots, G(K))$ for $G(j) = \varphi_{K/2,\sigma^2}(j) / \sum_{\ell=1}^{K} \varphi_{K/2,\sigma^2}(\ell)$, where $\varphi_{K/2,\sigma^2}$ denotes the density function of a normal distribution with mean $K/2$ and variance σ^2, where we chose $\sigma \in \{0.5, 1, 2, 5\}$. For the initial distribution, we used $m = 100$ samples of probability distributions on E obtained from sampling K i.i.d. random variables uniformly distributed on $\{0, 1, 2, \ldots, 10\}$ and appropriate scaling. For $K \in \{50, 100\}$, we applied the algorithm for $N \in \{1, \ldots, K/2 + 5\}$ and determined the average Wasserstein distance of the resulting distribution to the target distribution G over the $m = 100$ sampled initial distributions. This seems more interesting and informative than the value J_0 of the optimization problem itself, since this is difficult to interpret. The results are shown in Figure 1. We observe a linear decrease of the Wasserstein distance in terms of the number of time steps N. Depending on the choice of σ, we notice some index $N_0 = N_0(\sigma)$ such that $H(F_N) \approx 0$ for $N \geq N_0$. It appears to be decreasing in terms of the standard deviation σ, but even for the smallest choice of $\sigma = 0.5$, only $K/2$ time steps are necessary to achieve an accurate approximation of the target distribution. Note that using $N \geq K - 1$ time steps allows us to achieve any target distribution starting from any initial distribution due to the construction of the algorithm with the 'worst case' (i.e., the case with a necessary number of $N = K - 1$ time steps) being $F_0 = (1, 0, \ldots, 0)$ and $G = (0, \ldots, 0, 1)$.

Figure 2 additionally shows boxplots generated from the Wasserstein distances of the target distribution G for $K \in \{50, 100\}$ and $\sigma \in \{0.5, 1, 2, 5\}$ and the distribution F_{30} obtained after performing the first 30 steps of Algorithm 1 for the $m = 100$ initial samples. We can again observe the smallest Wasserstein distance for the largest value of σ. Moreover, we notice that the dispersion of the Wasserstein distances is smallest for the largest σ as well.

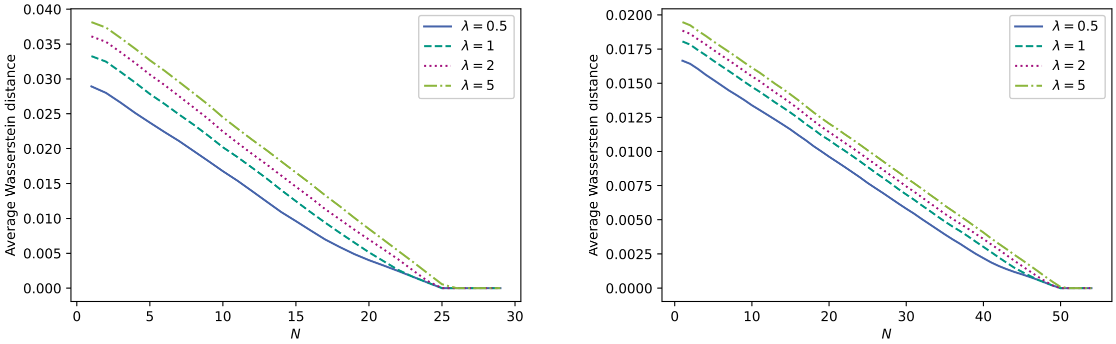
For a second example, we used a rescaled shifted exponential distribution as the target distribution, i.e., $G = (G(1), \ldots, G(K))$ for $G(j) = f_\lambda(j) / \sum_{\ell=1}^{K} f_\lambda(\ell)$, where $f_\lambda(x) = \lambda \mathrm{e}^{-\lambda(x - K/2)}$, $x > K/2$, denotes the density of an exponential distribution with parameter $\lambda > 0$ which was shifted to the right by $K/2$. We chose $\lambda \in \{0.5, 1, 2, 5\}$ for our numerical example. The initial distributions are sampled in the same way as before. The resulting



(a) $K = 50$, $n = 15$

(b) $K = 100$, $n = 30$

FIGURE 2. Boxplots of the Wasserstein distance of F_n from Algorithm 1 and the target distribution G (rescaled normal distribution) using the data from the $m = 100$ initial samples for $\sigma \in \{0.5, 1, 2, 5\}$.



(a) $K = 50$

(b) $K = 100$

FIGURE 3. Average Wasserstein distance of F_N from Algorithm 1 and the target distribution G (rescaled shifted exponential distribution) over $m = 100$ initial samples for $\lambda \in \{0.5, 1, 2, 5\}$ and $K \in \{50, 100\}$.

average Wasserstein distance of F_N and G over the $m = 100$ initial samples F_0 is shown in Figure 3. Similarly to the case of a rescaled normal target distribution, we notice a linear decay of the average Wasserstein distance. Moreover, the index $N_0 = N_0(\lambda)$ with $H(F_N) \approx 0$ for $N \geq N_0$ seems to be located at $K/2$ for any choice of λ. However, for $N < N_0$, the average Wasserstein distance is larger for larger values of λ.

5.4. **Further Applications.** The approach we describe here offers maximal flexibility in shaping reward distributions to fit some targeted distributions. This is for example interesting in portfolio optimization where constraints have to be met (e.g., Value-at-Risk constraints, see Basak and Shapiro (2001)) or the return distribution is shaped according to the risk preferences of the agent. There have been some applied studies in this direction Brayman et al. (2023) and some theoretical findings about the connection between target criterion and trading strategy Cox et al. (2014) as well as solution techniques He and Zhou (2011). This aspect could be interesting for personalized robo-advising Capponi et al. (2022) where AI tools choose portfolio strategies which match with the agents' preferences.

Another application, as indicated in the last subsection, is an optimal approximation of a continuous-time stochastic process (X_t) (e.g. Brownian motion) by a discrete-time Markov chain $(\hat{X}_n)$ in the sense that at pre-determined time points $t_1 < t_2 < \ldots < t_n$ the weighted distance $\sum_{j=1}^n \alpha_j W_1(X_{t_j}, \hat{X}_j)$ for $\alpha_j \geq 0$ has to be minimized. In this situation the transition probabilities of the discrete-time Markov chain have to be chosen. There may be constraints about the domain of the transition kernel, i.e., transitions to only certain values may be feasible.

More demanding applications include for example the optimal control of crowd behavior. Such models are often formulated in terms of mean-field systems Carmona et al. (2018) where the individual behavior is modeled in relation to the empirical distribution of the other individuals. Individuals might be persons, cars, animals and so on. Often a certain distribution of those individuals is preferred for example to avoid congestion, to guarantee smooth exiting of a building, to obtain an optimal shape of a flock of birds etc. (see Sect. 1 in Carmona et al. (2018)). Thus, the aim is here to find a policy such that the distribution of the individuals is optimally shaped.

6. Conclusion

In this paper, we studied Markov Decision Processes where the objective consists of functionals of the distribution of the accumulated reward. We showed that these kind of problems can be formulated as a dynamic program by defining a lifted MDP. The corresponding Bellman equation yields an algorithm for solving these problems. We have seen that this approach comprises many well-studied problems and allows a different point of view on the traditional Bellman equation.

References

Altman, E. (2021). *Constrained Markov decision processes.* Routledge.

Backhoff, J., Beiglbock, M., Lin, Y., and Zalashko, A. (2017). Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562.

Basak, S. and Shapiro, A. (2001). Value-at-risk-based risk management: optimal policies and asset prices. *The review of financial studies*, 14(2):371–405.

Bäuerle, N. and Glauner, A. (2021). Minimizing spectral risk measures applied to markov decision processes. *Mathematical Methods of Operations Research*, 94(1):35–69.

Bäuerle, N. and Jaśkiewicz, A. (2024). Markov decision processes with risk-sensitive criteria: an overview. *Mathematical Methods of Operations Research*, 99(1):141–178.

Bäuerle, N. and Jaśkiewicz, A. (2025). Time-consistency in the mean-variance problem: A new perspective. *IEEE Transactions on Automatic Control*, 70(1):251–262.

Bäuerle, N., Jaśkiewicz, A., and Nowak, A. S. (2025). Mean–variance optimization in discrete-time decision processes with general utility function. *Automatica*, 174:112142.

Bäuerle, N. and Ott, J. (2011). Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74:361–379.

Bäuerle, N. and Rieder, U. (2011). *Markov Decision Processes with Applications to Finance.* Springer Science & Business Media.

Bäuerle, N. and Rieder, U. (2014). More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120.

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR.

Bellemare, M. G., Dabney, W., and Rowland, M. (2023). *Distributional Reinforcement Learning*. MIT Press. http://www.distributional-rl.org.

Borkar, V. S. (2002). Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2):294–311.

Borkar, V. S. (2010). Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, volume 5.

Brayman, S., Potts, N., Brayman, K., and Komissarov, Y. (2023). Profile to portfolio: Where is the missing link? *Financial Services Review*, 31(4):246–265.

Capponi, A., Olafsson, S., and Zariphopoulou, T. (2022). Personalized robo-advising: Enhancing investment through client interaction. *Management Science*, 68(4):2485–2512.

Carmona, R., Delarue, F., et al. (2018). *Probabilistic theory of mean field games with applications I-II*, volume 3. Springer.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2015). Risk-sensitive and robust decision-making: a CVvaR optimization approach. *Advances in Neural Information Processing Systems*, 28.

Chung, K.-J. and Sobel, M. J. (1987). Discounted mdp's: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62.

Cox, A. M., Hobson, D., and OBłÓJ, J. (2014). Utility theory front to back—infering utility from agents'choices. *International Journal of Theoretical and Applied Finance*, 17(03):1450018.

Cui, X., Li, X., and Li, D. (2014). Unified framework of mean-field formulations for optimal multi-period mean-variance portfolio selection. *IEEE Transactions on Automatic Control*, 59(7):1833–1844.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. (2018a). Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. (2018b). Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Derman, C. and Strauch, R. E. (1966). A note on memoryless rules for controlling sequential control processes. *The Annals of Mathematical Statistics*, 37:276–278.

Filar, J. A., Krass, D., and Ross, K. W. (1995). Percentile performance criteria for limiting average Markov decision processes. *IEEE Transactions on Automatic Control*, 40(1):2–10.

Gerstenberg, J., Neininger, R., and Spiegel, D. (2023). On solutions of the distributional Bellman equation. *Electronic Research Archive*, pages 4459–4483.

Gilbert, H., Weng, P., and Xu, Y. (2017). Optimizing quantiles in preference-based Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Gut, A. and Stadtmüller, U. (2021). Variations of the elephant random walk. *Journal of Applied Probability*, 58(3):805–829.

He, X. D. and Zhou, X. Y. (2011). Portfolio choice via quantiles. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 21(2):203–231.

Hernández-Lerma, O. and Lasserre, J. B. (1996). *Discrete-Time Markov Control Processes, Basic Optimality Criteria*. Springer Science & Business Media.

Howard, R. A. and Matheson, J. E. (1972). Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369.

Jaquette, S. C. (1976). A utility criterion for Markov decision processes. *Management Science*, 23(1):43–49.

Kallenberg, L. (2002). Finite state and action MDPs. *Handbook of Markov decision processes*, pages 21–87.

Kantorovich, L. (2006). On a problem of Monge. *Journal of Mathematical Sciences*, 133(4).

Li, X., Zhong, H., and Brandeau, M. L. (2022). Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447.

Lyle, C., Bellemare, M. G., and Castro, P. S. (2019). A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511.

Mannor, S. and Tsitsiklis, J. (2011). Mean-variance optimization in Markov decision processes. *arXiv preprint arXiv:1104.5601*.

Marthe, A., Garivier, A., and Vernade, C. (2024). Beyond average return in Markov decision processes. *Advances in Neural Information Processing Systems*, 36.

Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–290.

Moghimi, M. and Ku, H. (2025). Beyond cvar: Leveraging static spectral risk measures for enhanced decision-making in distributional reinforcement learning. *arXiv preprint arXiv:2501.02087*.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704.

Moulos, V. (2021). Bicausal optimal transport for Markov chains via dynamic programming. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1688–1693. IEEE.

Pires, B. Á., Rowland, M., Borsa, D., Guo, Z. D., Khetarpal, K., Barreto, A., Abel, D., Munos, R., and Dabney, W. (2025). Optimizing return distributions with distributional dynamic programming. *arXiv preprint arXiv:2501.13028*.

Piunovskiy, A. B. (1997). *Optimal Control of Random Sequences in Problems with Constraints*. Springer Science & Business Media.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rachev, S. T. and Rüschendorf, L. (2006a). *Mass Transportation Problems: Volume I: Theory*. Springer Science & Business Media.

Rachev, S. T. and Rüschendorf, L. (2006b). *Mass Transportation Problems: Volume II: Applications*. Springer Science & Business Media.

Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M. G., and Dabney, W. (2019). Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR.

Ruszczyński, A. (2010). Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125:235–261.

Shen, Y., Stannat, W., and Obermayer, K. (2013). Risk-sensitive Markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672.

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328.

Terpin, A., Lanzetti, N., and Dörfler, F. (2024). Dynamic programming in probability spaces via optimal transport. *SIAM Journal on Control and Optimization*, 62(2):1183–1206.

Villani, C. (2008). *Optimal Transport: Old and New*, volume 338. Springer.

Wu, C. and Lin, Y. (1999). Minimizing risk models in Markov decision processes with policies depending on target values. *Journal of Mathematical Analysis and Applications*, 231(1):47–67.