Karlsruhe Institute of Technology

Multivariate estimation in nonparametric models: Stochastic neural networks and Lévy processes

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Maximilian F. Steffen, M.Sc.

Tag der mündlichen Prüfung:	14. Februar 2024
Referent:	Prof. Dr. Mathias Trabs
Korreferenten:	Prof. Dr. Denis Belomestny
	Prof. Dr. Vicky Fasen-Hartmann

Acknowledgments

First, I would like to thank Mathias Trabs for his excellent supervision during my doctoral studies. You always took time to have discussions with me which helped to clarify my thoughts and spark new ideas while also giving me the space to work independently. Seeing your intuition for complex problems has improved my skill set immensely. Further, I am grateful for the generous offer to join you in moving to KIT as this has opened up new connections and perspectives. In view of your capabilities, I particularly appreciate your down-to-earthness.

The interdisciplinary collaboration with Sebastian Bieringer and Gregor Kasieczka has been a prolific learning experience for me. Your approach to research is refreshing and inspiring. Sebastian, I have learned a great deal from your aptitude in getting algorithms to not just converge in theory.

I thank Denis Belomestny and Vicky Fasen-Hartmann for acting as referees for my thesis. Despite only having met Denis Belomestny briefly at a conference, you took interest in my research. Vicky, I am also thankful for the questions you posed after various talks I gave and for your contribution to the positive atmosphere at the institute.

For about the first half of my doctoral studies, I was affiliated to Universität Hamburg, before moving to Karlsruhe Institute of Technology. I can happily say that I have been part of a fantastic team at both universities. I would like to thank my colleagues and former colleagues for creating a sense of unity and for their helpfulness.

Financial support by the Deutsche Forschungsgemeinschaft (DFG) through project TR 1349/3-1 is gratefully acknowledged. The empirical studies in Section 3.3 were enabled by the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. Further, support by the state of Baden-Württemberg through bwHPC is gratefully acknowledged. In particular, computational resources for the simulation examples in Section 5.2 were provided.

Finally, I thank my family and my friends for their support and encouragement throughout this journey.

Prior publications

Substantial parts of this thesis are based on preprints which are available on arXiv and which have been submitted to peer reviewed journals.

Section 2.2 is based on

- Steffen, M.F. (2023). PAC-Bayes bounds for high-dimensional multi-index models with unknown active dimension. *arXiv preprint arXiv:2303.13474*.

Chapter 3 and Section 4.1 are based on

- Bieringer, S., Kasieczka, G., Steffen, M.F. & Trabs, M. (2023). Statistical guarantees for stochastic Metropolis-Hastings. *arXiv preprint arXiv:2310.09335*.

All coauthors have contributed equally to this publication. SB and GK have focused on the numerical analysis, while MS and MT have focused on the theoretical results. The presentation of these results in the thesis at hand has been agreed to by all coauthors.

Sections 2.1 and 4.2 are based on

- Steffen, M.F. & Trabs, M. (2023). A PAC-Bayes oracle inequality for sparse neural networks. *arXiv preprint arXiv:2204.12392*.

All coauthors have contributed equally to this publication. The presentation of these results in the thesis at hand has been agreed to by MT.

Chapter 5 is based on

- Steffen, M.F. (2023). Estimating a multivariate Lévy density based on discrete observations. *arXiv preprint arXiv:2305.14315*.

Some of the publications share similar preliminaries, which have been merged for coherency. With this in mind, it should be noted that direct quotes from the publications above appear throughout this thesis.

Contents

Notation

$\lvert\cdot\rvert_q$	ℓ^q-norm of vectors for $q \in [1,\infty]$
$\lvert\cdot\rvert$	Euclidean norm $\lvert\cdot\rvert = \lvert\cdot\rvert_2$
$\lVert\cdot\rVert_{L^p(U)}$	L^p-norm of functions restricted to a set $U \subseteq \mathbb{R}^d$ for $p \in [1,\infty]$
$\lVert\cdot\rVert_{L^p}$	L^p-norm $\lVert\cdot\rVert_{L^p} = \lVert\cdot\rVert_{L^p(\mathbb{R}^d)}$
$\lVert\cdot\rVert_\infty$	supremum norm of functions, i.e. $\lVert\cdot\rVert_\infty = \lVert\cdot\rVert_{L^\infty}$
$\langle\cdot,\cdot\rangle$	standard vector product, i.e. $\langle x,y\rangle = \sum_{k=1}^d x_k y_k$ for $x,y \in \mathbb{R}^d$ and $\langle f,g\rangle = \int fg\,\mathrm{d}x$ for square integrable functions $f,g\colon \mathbb{R}^d \to \mathbb{R}$
x^β	defined by $x^\beta := \prod_{k=1}^d x_k^{\beta_k}$ for $x \in \mathbb{R}^d$ and a multi-index $\beta \in \mathbb{N}_0^d$
$\lvert x\rvert^\beta$	defined by $\lvert x\rvert^\beta := \prod_{k=1}^d \lvert x_k\rvert^{\beta_k}$ for $x \in \mathbb{R}^d$ and a multi-index $\beta \in \mathbb{N}_0^d$
$g^{(\beta)}$	mixed partial derivative of a sufficiently differentiable function $g\colon \mathbb{R}^d \to \mathbb{R}$ in the direction of a multi-index $\beta \in \mathbb{N}_0^d$
$\lceil a\rceil$	smallest integer greater or equal to a
$\lfloor a\rfloor$	largest integer less or equal to a
$\lfloor a\rfloor_<$	largest integer *strictly* less than a
$a \vee b$	maximum of two real numbers a,b
$a \wedge b$	minimum of two real numbers a,b
$\propto$	proportional to
$\lesssim$	less or equal up to a multiplicative constant independent of the parameters involved
$\gtrsim$	greater or equal up to a multiplicative constant independent of the parameters involved
$\overset{d}{=}$	equal in distribution
$\ll$	absolute continuity
span	linear span
KL	Kullback-Leibler divergence, see (2.5)
δ_x	Dirac measure in x
$\mathbb{\lambda}, \mathbb{\lambda}^d$	Lebesgue measure on $\mathbb{R}$ and $\mathbb{R}^d$, respectively
A^{c}	complement of a set A
$\mathbb{1}_A$	indicator function of a set A, defined by $\mathbb{1}_A(x) = 1$ if $x \in A$ and $\mathbb{1}_A(x) = 0$ if $x \notin A$

∇f	gradient of a function f
Δf	Laplacian of a function f
$\mathscr{F}g$	Fourier transform of $g \in L^1(\mathbb{R}^d)$ defined by $\mathscr{F}g(u) = \int \mathrm{e}^{\mathrm{i}\langle u,x\rangle} g(x)\,\mathrm{d}x$
$\mathscr{F}\mu$	Fourier transform of a measure μ defined by $\mathscr{F}\mu(u) = \int \mathrm{e}^{\mathrm{i}\langle u,x\rangle}\,\mu(\mathrm{d}x)$
$\operatorname{supp} f$	support of a function f
$A^\top$	transpose of a matrix A
$\operatorname{tr}(A)$	trace of a matrix A
E_d	$d \times d$ identity matrix
$\mathcal{O}, o$	Landau notation
$\mathcal{O}_\mathbb{P}, o_\mathbb{P}$	stochastic Landau notation
$C^\infty(\mathbb{R}^d)$	class of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ which are differentiable arbitrarily often
$\mathcal{C}_d^s(U)$	class of s-Hölder regular functions $f\colon U \to \mathbb{R}$, for an open subset $U \subseteq \mathbb{R}^d$, with finite Hölder norm $\|f\|_{\mathcal{C}_d^s(U)} \coloneqq \sum_{\lvert\beta\rvert_1 \leqslant \lfloor s\rfloor_<} \sup_{x\in U} \lvert f^{(\beta)}(x)\rvert + \max_{\lvert\beta\rvert_1 = \lfloor s\rfloor_<} \sup_{x,y\in U, x\neq y} \frac{\lvert f^{(\beta)}(x) - f^{(\beta)}(y)\rvert}{\lvert x-y\rvert^{s-\lfloor s\rfloor_<}}$
$\mathcal{C}_d^s(U,C), \mathcal{C}^s(U,C)$	ball of s-Hölder regular functions $f\colon U \to \mathbb{R}$, for an open subset $U \subseteq \mathbb{R}^d$, with Hölder norm $\|f\|_{\mathcal{C}^s(U)} \leqslant C$
$B^\alpha_{q,d^*}(\xi)$	Besov-ellipsoid, see (2.11)
$\mathcal{H}(q, \mathbf{d}, \mathbf{t}, \beta, C_0)$	class of hierarchical functions, see (3.16)

1 Introduction

Essential features in modern data science, especially in machine learning and high-dimensional statistics, are large sample sizes and large parameter space dimensions. In recent years, the flexibility and empirical capabilities of complex models have attracted practitioners, but the training of such models often comes at immense computational costs due to large samples and the large number of parameters. Moreover, a lot of training methods, while strong in practice, cannot statistically guarantee their performance in terms of risk bounds. As a consequence, the design of cutting edge methods is characterized by a tension between numerically feasible and efficient algorithms, and approaches which also satisfy theoretically justified statistical properties.

This thesis contributes solutions to two quite disjoint problems showcasing the wide spectrum of fields where nonparametric statistics can provide answers to the challenges presented by modern applications while also admitting statistical guarantees. First, we consider the classical nonparametric problem of estimating a regression function based on independent noisy observations. Second, there are multivariate time-dependent random phenomena which can be modeled using stochastic processes. Lévy processes serve as a cornerstone for models of such phenomena whenever jumps are involved, but the calibration of the jump behavior in particular presents an ambitious nonparametric problem.

A key tool for analyzing the regression problem will be the so-called Gibbs posterior, which allows for oracle inequalities in rather general settings. As a first step, we construct an estimator for high-dimensional multi-index models with unknown active dimension. Motivated by the rising relevance of uncertainty quantification in deep learning, we then turn our attention to stochastic neural networks, where we can access the Gibbs posterior through Markov chain Monte Carlo (MCMC) methods. To reduce the computational costs associated with such methods for large samples, a naive approach of incorporating stochastic (mini-)batches proves futile as the resulting algorithm yields less accurate estimates. However, it turns out that this drawback can be fixed with a simple correction term allowing for a fully scalable algorithm. Based on this scalable method and using our insights from multi-index models, we provide theoretical extensions to choose the network architecture in a data-driven way and to deal with high-dimensional data. Regarding the calibration of multivariate Lévy processes, we provide a nonparametric estimator

for the jump density which is applicable to fairly general classes of Lévy processes and robust across sampling frequencies.

Related literature

We first provide an overview of the literature related to this thesis.

Stochastic neural networks

For an introduction to neural networks, see e.g. Goodfellow et al. (2016) and Schmidhuber (2015). While early theoretical foundations for neural nets are summarized by Anthony & Bartlett (1999), the excellent approximation properties of deep neural nets, especially with the ReLU activation function, have been discovered in recent years, see e.g. Yarotsky (2017) and the review paper DeVore et al. (2021). In addition to these approximation properties, an explanation of the empirical capabilities of neural networks has recently been given by Schmidt-Hieber (2020) as well as Bauer & Kohler (2019): While classical regression methods suffer from the curse of dimensionality, deep neural network estimators can profit from a hierarchical structure of the regression function and a possibly much smaller intrinsic dimension.

In addition to theoretical guarantees, uncertainty quantification is an important and challenging problem for neural networks. A widely used gateway to uncertainty quantification is the Bayesian approach. It led to the introduction of stochastic neural networks, where a distribution over the network weights is learned, see Graves (2011) and Blundell et al. (2015) and numerous subsequent articles. These Bayes-type methods present two main challenges: the derivation of statistical guarantees and the development of algorithms to sample from the posterior.

Theoretical guarantees through PAC-Bayes bounds

Bayesian methods enjoy high popularity for quantifying uncertainties in complex models. The general approach is to impose a *prior distribution* on the parameter class and use the Bayes formula to compute the *posterior distribution* of the parameters given the data. Since this computation requires the unknown distribution of the data given the parameters, restrictive model assumptions are needed. To circumvent this, the *probably approximately correct* (PAC-)Bayes approach replaces the posterior distribution with an approximation, the *Gibbs posterior*. An estimator drawn from the Gibbs posterior can then be analyzed from a frequentist perspective. In particular, the literature aims for PAC-Bayes bounds, see the review papers by Guedj (2019) and

Alquier (2021). Such bounds are divided into empirical PAC-Bayes bounds, see e.g. McAllester (1999a,b), and oracle PAC-Bayes bounds, see e.g. Catoni (2004, 2007). Both types of bounds control the error of an estimator based on the Gibbs posterior with a high probability and can be derived in quite general settings. Empirical PAC-Bayes bounds are in terms of the empirical risk, whereas oracle PAC-Bayes bounds are in terms of the *oracle*, that is, the theoretical risk minimizer over a class of parameters.

In a regression setting, PAC-Bayes bounds have been studied for instance by Audibert (2004, 2009); Audibert & Catoni (2011) and the references therein. In a high-dimensional regression, sparsity can be used to reduce the effective dimension of the model. In the context of oracle PAC-Bayes bounds this has been done through additive models by Guedj & Alquier (2013), as well as through single-index models by Alquier & Biau (2013). Similar ideas have been used by Castillo et al. (2015) in a Bayesian sparse linear regression. Empirical PAC-Bayes bounds are investigated intensively for (deep) neural nets, see Dziugaite & Roy (2017); Pérez-Ortiz et al. (2021) and further references in Alquier (2021, Section 3.3).

The analysis of the Bayesian procedure from a frequentist point of view embeds into the non-parametric Bayesian inference, see Ghosal & van der Vaart (2017). Coverage of credible sets has been studied, for instance, by Szabó et al. (2015) and Rousseau & Szabó (2020) and based on the Bernstein-von Mises theorem by Castillo & Nickl (2014) among others.

While contraction rates for Bayesian neural networks have been studied by Polson & Ročková (2018) and Chérief-Abdellatif (2020), the theoretical properties of credible sets are not well understood so far. Franssen & Szabó (2022) have studied an empirical Bayesian approach, where only the last layer of the network is Bayesian while the remainder of the network remains fixed.

Sampling via MCMC methods

In order to apply Bayesian estimators in practice, we need to sample from the posterior distribution. The classical approach are MCMC methods. For large parameter spaces, gradient-based Monte Carlo methods are particularly useful, with e.g. Langevin dynamics serving as a prototypical example. Advanced methods such as Metropolis adjusted Langevin (MALA), see e.g. Besag (1994); Roberts & Tweedie (1996a), and Hamiltonian Monte Carlo, see e.g. Duane et al. (1987); Neal (2011), equip the Markov chain with a Metropolis-Hastings (MH) step to accept or reject the proposed next state of the chain. From the practical point of view, the MH step improves robustness with respect to the choice of the tuning parameters and, in theory MH speeds up the convergence of the Markov chain.

If the sample size is large, the computational cost of gradient-based MCMC methods can be reduced by replacing the gradient of the full loss over all observations by a stochastic gradient. This is a standard approach in empirical risk minimization and has been applied successfully to Langevin dynamics as well, see Alexos et al. (2022); Li et al. (2016); Patterson & Teh (2013); Welling & Teh (2011). In this case, the MH steps remain as a computational bottleneck: Since the target distribution depends on the full data set, we have to compute the loss on the full sample to calculate the acceptance probabilities. Among the approaches to circumvent this problem, see Bardenet et al. (2017) for a review, a *stochastic MH* step is presumably the most natural one. There, the full loss in the acceptance probability is replaced by a (mini-)batch approximation which reduces the computational cost considerably, see Wu et al. (2022).

Tailoring Markov chains to the needs of current neural network applications is a field of ongoing investigation. Different efforts were made to improve efficiency by mixing, that is transitioning between modes of the posterior landscape. Zhang et al. (2020) employ a scheduled step-size to help the algorithm move between different modes of the posterior, while contour stochastic gradient MCMC, see Deng et al. (2020b, 2022), use a piece-wise continuous function to flatten the posterior landscape which is itself determined through MCMC sampling or from parallel chains. Parallel chains of different temperature are employed by Deng et al. (2020a) at the cost of memory space during computation. Only limited research on scaling MCMC for large data has been done. Most recently, Cobb & Jalaian (2021) introduced a splitting scheme for Hamiltonian Monte Carlo maintaining the full Hamiltonian.

In view of possibly better scaling properties, variational Bayes methods have been studied intensively in recent years. Variational Bayes methods also do not sample from the posterior distribution itself, but approximate the posterior within a parametric distribution class which can be easily sampled from, see Blei et al. (2017) for a review. The theoretical understanding of variational Bayes methods is a current research topic, see Zhang & Zhou (2020); Zhang & Gao (2020); Ray & Szabó (2022), and references therein.

Statistics for Lévy processes

Lévy processes are a staple to model continuous-time phenomena involving jumps, for instance in physics and finance, see Woyczyński (2001) and Cont & Tankov (2004), respectively. Naturally, many of these applications call for multivariate processes which significantly complicates the estimation problem compared to the one-dimensional case. Matters worsen as practitioners often only have time-discrete data at their disposal which obstructs the identification of the jumps and hence the calibration of such models. Statistical results in this setting are typically limited to the one-dimensional case or omit the estimation of the jump distribution, despite the

practical relevance.

On a theoretical level, the distribution of the Lévy process is uniquely determined by its characteristic triplet, that is, the volatility matrix, the drift and the Lévy measure. The former two characterize the Gaussian component of the process, while the latter characterizes the jump distribution. From a statistical point of view, the estimation of the Lévy measure is most challenging as we are faced with a nonparametric problem.

The literature commonly distinguishes between the following two observation regimes for a Lévy process observed at equidistant time points $0 < \delta, 2\delta, \ldots, n\delta =: T$: Under the low-frequency regime, δ is fixed as $n \to \infty$, whereas $\delta \searrow 0$ under the high-frequency regime. Fig. 1.1 illustrates the estimation problem.

Motivated by the clearer separation between the jumps themselves and the Gaussian component as $\delta \searrow 0$ under the high-frequency regime, threshold-based estimators have been applied extensively. Beyond the overview given by Aït-Sahalia & Jacod (2012), Duval & Mariucci (2021) apply such an approach to the estimation of the Lévy measure, Gegler & Stadtmüller (2010) study the estimation of the entire Lévy triplet, and Mies (2020) estimates the Blumenthal-Getoor index. An alternative approach under the high-frequency regime is the use of sieve estimators, see e.g. Figueroa-López (2011). However, these references are restricted to the one-dimensional case and multidimensional extensions seem intricate due to the multitude of directions in which the process can jump. A notable exception is the work by Bücher & Vetter (2013), who estimate the tail-integrals of a multivariate Lévy process.

Under the low-frequency regime, we cannot identify the intermittent jumps even in the absence of a Gaussian component resulting in an ill-posed inverse problem, see Neumann & Reiß (2009). A popular way out is the spectral method, see Belomestny & Reiß (2015), which leverages the relationship of the Lévy triplet with the characteristic function of the process at any time point. Turning the observations of the process into increments, this characteristic function is estimated and then used to draw inference on parts of the Lévy triplet. The method was first considered by Belomestny & Reiß (2006) in the context of exponential Lévy models and has since been applied extensively, see Belomestny (2010), Gugushvili (2012), Nickl & Reiß (2012), Reiß (2013), Trabs (2015).

The estimation of the volatility matrix itself has previously been studied by Papagiannouli (2020, high-frequency) and Belomestny & Trabs (2018, low-frequency). A related issue is the estimation of the covariance matrix in deconvolution problems, see Belomestny et al. (2019).

An effect emerging for multivariate processes is the possibility of different dependence structures between the components which can be in disagreement with the existence of a Lévy density in

the form of a Lebesgue density on the whole state space. Nonparametric statistical results in such settings are even rarer. Belomestny (2011) estimates the Lévy density for a time changed Lévy process with independent components. An alternative approach is to characterize the dependence structure with Lévy copulas, see Cont & Tankov (2004) for an introduction and the aforementioned Bücher & Vetter (2013) for the estimation of the Lévy copula under a high-frequency regime.

Additionally, there is an active field of research on Bayesian inference for stochastic processes in general and Lévy processes in particular, see Belomestny et al. (2022) for an overview.

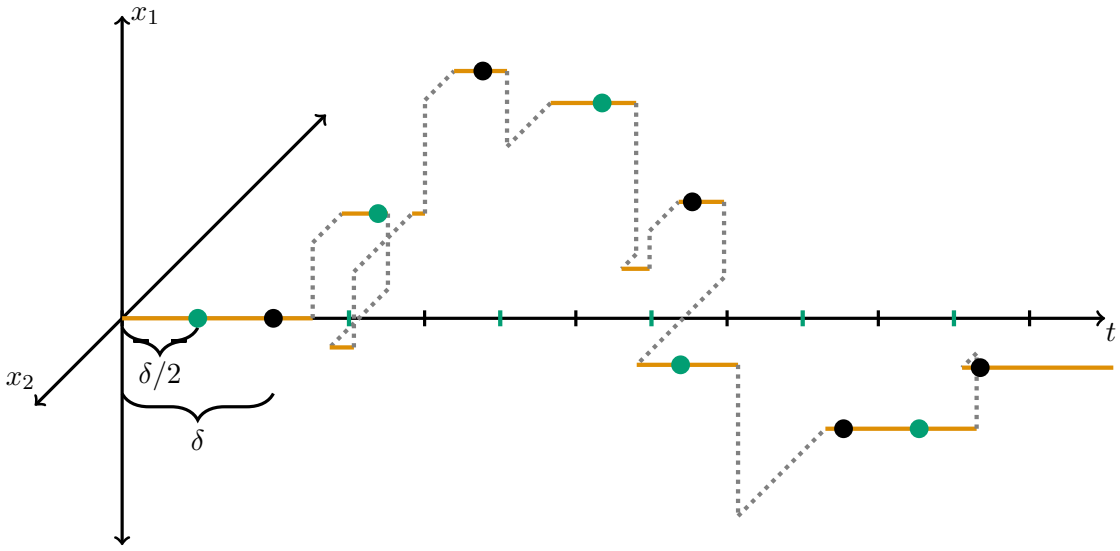

Figure 1.1: Illustration of a trajectory of a two-dimensional pure jump Lévy process (in orange). The jumps are indicated by the dashed gray lines, but we only observe the black dots. For a narrower time grid, i.e., with $\delta/2$, fewer jumps are missed as we now additionally observe the green dots.

Own contributions

In the following, the main contributions of this thesis are outlined in the context of the related literature. A central model under investigation in this thesis is the nonparametric regression model

$$Y = f(\mathbf{X}) + \varepsilon$$

with a random pair $(\mathbf{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$, an observation error ε satisfying $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$ almost surely (a.s.) and an unknown regression function $f\colon \mathbb{R}^p \to \mathbb{R}$. The aim is to estimate the regression function based on a training sample $\mathcal{D}_n := (\mathbf{X}_i, Y_i)_{i=1,\dots,n}$ given by n i.i.d. copies of

$(\mathbf{X}, Y)$. To model time-dependent random phenomena, we consider Lévy processes. The aim is to estimate the multivariate Lévy density ν of a Lévy process $L = (L_t)_{t \geqslant 0}$ based on observations $L_\delta, L_{2\delta}, \ldots, L_{n\delta}$ on a time grid with time difference $\delta > 0$.

High-dimensional multi-index models with unknown active dimension

A popular approach to reduce the effective dimension of nonparametric regression models is to impose a multi-index structure on the regression function (Li, 1991). While we do not assume that the observations exactly follow a multi-index model, our method builds upon an approximation of the regression function of the form

$$f(\mathbf{x}) \approx g^*(W^*\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^p, \tag{1.1}$$

for some *active dimension* d^* substantially smaller than p, a sparse *dimension reduction matrix* $W^* \in \mathbb{R}^{d^* \times p}$, and a (measurable) *link function* $g^* \colon \mathbb{R}^{d^*} \to \mathbb{R}$. Following the aforementioned Li (1991), the estimation of the space spanned by the rows of W^* has been studied extensively in the literature, see e.g. Hristache et al. (2001), Xia (2007), and Dalalyan et al. (2008), but under the assumption of a known active dimension d^*. While some research has been done on the estimation of d^* itself, see Xia et al. (2002) and Zhu et al. (2006), the estimation of the overall model has relied on estimating W^* and g^* separately to then analyze the propagation error, see Klock et al. (2021). If the dimension p is large, the estimation problem suffers from the well-known curse of dimensionality. This is of particular importance in numerous recent applications where p may exceed the sample size n. The existing analysis of high-dimensional multi-index models in this setting is rather limited. A notable exception is Yang et al. (2017) who recover the dimension reduction matrix under the assumption that the distribution of the covariates is known.

As our first application of the estimation approach via the Gibbs posterior and to demonstrate its potential, we consider multi-index models. Such an estimator has been applied successfully to the single-index model (i.e. $d^* = 1$) without miss-specification by Alquier & Biau (2013). We generalize the estimation method for single-index models to the more flexible class of multi-index models. In particular, we aim for a method which adapts to the unknown active dimension d^*, the sparsity of W^* and the regularity of g^* to achieve a good approximation of the form (1.1) based on the given data.

Our method allows for a fully data driven complete calibration of the high-dimensional multi-index model with unknown active dimension. The estimator achieves the minimax-optimal rate of convergence (up to a logarithmic factor) for such estimation problems and no additional price

is paid for the unknown active dimension.

Statistical guarantees for stochastic Metropolis-Hastings

Approaching the same nonparametric regression with neural networks introduces a new focus on algorithmic aspects. We demonstrate that an MCMC based method with a stochastic MH step is computationally feasible for large samples and we can prove an optimal bound for the prediction risk as well as uncertainty statements for the underlying posterior distribution. Fig. 1.2 illustrates the quantification of uncertainty with stochastic neural networks.

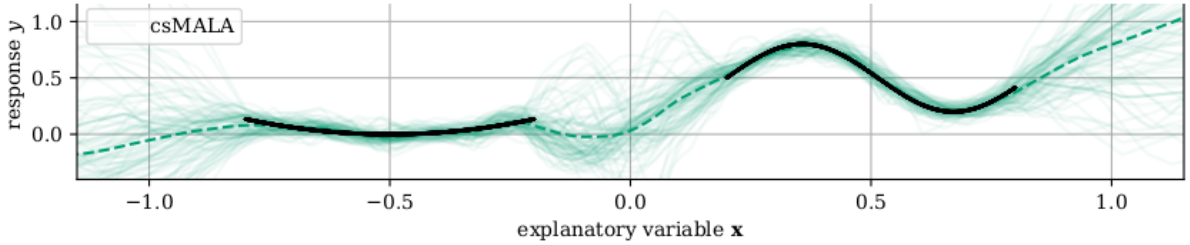

Figure 1.2: Illustration of uncertainty quantification in nonparametric regression via stochastic neural networks. Noisy data (black dots) of true regression function is only available in the intervals $[-0.8, -0.2]$ and $[0.2, 0.8]$. The thin green lines show samples drawn from the surrogate posterior of our corrected stochastic MALA (csMALA) method, which incorporates a corrected stochastic MH step. The dashed green line shows the corresponding posterior mean. Note how the spread of the green lines increases in the areas where no data is available resulting in higher uncertainty.

Bardenet et al. (2017, Section 6.1) have argued heuristically that the naive stochastic MH step reduces the effective sample size, which determines, for instance, contraction rates of the posterior distribution, to the size of the batch. To rigorously understand the statistical consequences of a stochastic MH step, we apply the pseudo-marginal MH perspective by Andrieu & Roberts (2009) and Maclaurin & Adams (2014). It turns out that a Markov chain with a stochastic MH step does not converge to the original target posterior distribution, but a different distribution, which we call *surrogate posterior* and whose statistical performance is indeed determined by the batch size only. However, we show that there is a simple correction term in the risk such that the resulting stochastic MH chain converges to a surrogate posterior which achieves the full statistical power in terms of optimal contraction rates.

We investigate the distance of the surrogate posteriors associated with the stochastic MH algorithm and the corrected stochastic MH algorithm to the original posterior distribution in terms of the Kullback-Leibler divergence. While these approximation results could be used to analyze

the surrogate posteriors based on properties of the original posterior as done for variational Bayes methods, see Ray & Szabó (2022), we instead investigate the surrogate posteriors directly which allows for sharp results.

We prove PAC-Bayes oracle inequalities for the surrogate posteriors of the stochastic MH method and its corrected modification in the context of deep neural networks. Based on that, we can conclude contraction rates as well as rates of convergence for the surrogate posterior mean. Applied to Hölder regular hierarchical regression functions, the contraction rate of the corrected stochastic MH procedure coincides with the minimax-optimal rate by Schmidt-Hieber (2020) (up to a logarithmic factor). The network estimator achieves this upper bound without prior knowledge of the hierarchical structure and the regularity of the regression function. While the aforementioned paper has analyzed sparse deep neural networks with ReLU activation function, similar results for fully connected networks are given by Kohler & Langer (2021), and we exploit their main approximation theorem. Moreover, we investigate size and coverage of credible balls from the surrogate posterior.

A simulation study demonstrates the merit of the correction term for sampling from a 10401 dimensional parameter space for a low-dimensional regression task. The samples from the surrogate posterior of our corrected stochastic MH algorithm, as well as their mean, show a significant improvement in terms of the empirical prediction risk and the size of credible balls over those taken from the surrogate posterior of the naive stochastic MH algorithm. The correction term cancels the bias on the size of accepted batches introduced by the stochastic setting. The Python code of the implementation is available, see Bieringer et al. (2023).

Stochastic neural networks with mixing priors

An advantage of estimators based on the Gibbs posterior is that modifications to the prior can allow the estimator to adapt to various structural properties of the model as we have seen in our analysis of multi-index model as well as in the literature, see the aforementioned Alquier & Biau (2013); Guedj & Alquier (2013). We apply this approach to the stochastic MH leading to the following two contributions.

The first concerns the choice of the network architecture. We demonstrate that a mixing prior over network architectures of varying size allows the stochastic MH to choose the optimal architecture in a fully data driven way. We suffer no additional loss in the rate of convergence. In particular, we still achieve the minimax-optimal rate of convergence (up to a logarithmic factor) over the class of hierarchical functions.

The second extension leads to an algorithm which can handle high-dimensional data by incorporating sparsity. The training of sparse neural networks with a Bayesian approach to estimate a Hölder regular regression function has been studied by the aforementioned Polson & Ročková (2018). We prove a PAC-Bayes oracle inequality which does not depend on the total number of weights of the neural network, but only the number of nonzero weights. This allows us to employ the approximation properties of sparse neural networks demonstrated by Schmidt-Hieber (2020). Again, we attain the minimax-optimal rate of convergence (up to a logarithmic factor). As an immediate consequence, this stochastic neural network offers an alternative to our approach for estimating high-dimensional multi-index models, which solidifies the connection between these models, hierarchical functions, and neural networks.

Estimating a multivariate Lévy density

As a contribution to statistics for Lévy processes, we provide an estimator for the multivariate Lévy density based on discrete observations of the Lévy process. More specifically, we adapt the spectral method to the multivariate setting by constructing a nonparametric estimator for the Lévy density ν, assuming that it exists. Our estimator requires no knowledge of the volatility and drift parameters and works uniformly over fully nonparametric classes of Lévy processes under mild assumptions. In particular, Lévy processes with infinite jump activity are allowed. The uniform rates we achieve naturally extend those from the one-dimensional case and optimality in our setting is discussed. Our estimation method is robust across sampling frequencies.

When estimating the Lévy density close to the origin, we enhance our method with an estimator for the volatility. However, even the proved minimax-optimal rates of convergence in the literature are too slow as to not affect our overall rates under the low-frequency regime. It is sufficient to estimate the trace of the volatility matrix and we show that this can be done with a much faster rate. With this enhancement, there is no additional loss for the unknown volatility in the rate for the estimation of the Lévy density.

Regarding the various possible dependence structures of multivariate Lévy processes, we propose a quantification of the estimation error when integrating against regular test functions without modifications to our method.

A key contribution in our proofs is to develop a generalization of a uniform risk bound for the empirical characteristic function to the multivariate case. We illustrate our estimation method with three simulation examples.

Additionally, we provide an outlook on how the spectral method could be extended to the estimation of the jump density of a high-dimensional Lévy process by incorporating our results

regarding nonparametric regression. For instance, Xu & Darve (2020) have proposed an estimation scheme based on neural networks in the one-dimensional case. While their empirical results look promising, they do not provide a theoretical analysis of their method.

Organization of the thesis

The dissertation is structured as follows: In Chapter 2, we introduce a general approach for deriving PAC-Bayes oracle inequalities in nonparametric regression models and apply the approach to high-dimensional multi-index models. Building on this methodology, in Chapter 3, we first focus on algorithmic aspects which provide computationally feasible access to the Gibbs posterior. Then, we apply these algorithmic concepts to the training of neural networks leading to statistical guarantees in the form of oracle inequalities and credible sets. Extensions of the algorithm are constructed in Chapter 4. There, we incorporate mixing priors to choose the network architecture in a data driven way and handle high-dimensional data. In Chapter 5, we focus on the second main problem of estimating the multivariate Lévy density based on discrete observations of a Lévy process. For readability, the proofs have been postponed to the end of their respective chapter. In Chapter 6, we provide an outlook on how our results could be combined to construct a nonparametric estimator for the Lévy density of a high-dimensional Lévy process.

2 A PAC-Bayes oracle inequality for high-dimensional multi-index models

In this chapter, we outline a general approach for deriving PAC-Bayes oracle inequalities using high-dimensional multi-index models as a guiding example. First, we introduce the estimation principle along with some notation. The strategy for deriving general PAC-Bayes oracle inequalities is presented in Section 2.1. In Section 2.2, we demonstrate how this strategy can be applied to high-dimensional multi-index models with unknown active dimension. In particular, this will allow us to circumvent the curse of dimensionality in a high-dimensional regression setting.

The general approach is well established in the literature, see the review paper Alquier (2021), and is introduced in the following for a self-contained presentation.

The aim is to estimate a regression function $f\colon \mathbb{R}^p \to \mathbb{R}$, $p \in \mathbb{N}$ based on a training sample $\mathcal{D}_n \coloneqq (\mathbf{X}_i, Y_i)_{i=1,\dots,n} \subset \mathbb{R}^p \times \mathbb{R}$ given by $n \in \mathbb{N}$ i.i.d. copies of generic random variables $(\mathbf{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with

$$Y = f(\mathbf{X}) + \varepsilon \tag{2.1}$$

and observation noise ε satisfying $\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0$ almost surely (a.s.). Equivalently, $f(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ a.s. For any estimator $\widehat{f}$, the prediction risk and its empirical counterpart are given by

$$R(\widehat{f}) \coloneqq \mathbb{E}_{(\mathbf{X},Y)}\big[\big(Y - \widehat{f}(\mathbf{X})\big)^2\big] \qquad \text{and} \qquad R_n(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \widehat{f}(\mathbf{X}_i)\big)^2,$$

respectively, where $\mathbb{E}$ denotes the expectation under $\mathbb{P}$ and $\mathbb{E}_Z$ is the expectation only with respect to a random variable Z. The accuracy of the estimation procedure will be quantified in terms of the excess risk

$$\mathcal{E}(\widehat{f}) \coloneqq R(\widehat{f}) - R(f) = \mathbb{E}_{\mathbf{X}}\big[\big(\widehat{f}(\mathbf{X}) - f(\mathbf{X})\big)^2\big] = \|\widehat{f} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})},$$

where $\mathbb{P}^{\mathbf{X}}$ denotes the distribution of $\mathbf{X}$.

We consider a parametric class of potential estimators $\mathcal{F} = \{f_\vartheta : \vartheta \in \Theta\}$ for some parameter space Θ with a potentially large parameter dimension $P \in \mathbb{N}$. For $f_\vartheta \in \mathcal{F}$ we abbreviate $R(\vartheta) = R(f_\vartheta)$ and

$$R_n(\vartheta) = R_n(f_\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\vartheta) \qquad \text{with} \qquad \ell_i(\vartheta) = \big(Y_i - f_\vartheta(\mathbf{X}_i)\big)^2.$$

Fixing a prior probability measure Π on Θ, the corresponding *Gibbs posterior* $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ is defined as the solution to the minimization problem

$$\inf_\nu \Big(\int R_n(\vartheta)\, \nu(\mathrm{d}\vartheta) + \frac{1}{\lambda}\, \mathrm{KL}(\nu \mid \Pi) \Big),$$

where the infimum is taken over all probability distributions ν on Θ. Hence, $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ will concentrate at parameters ϑ with a small empirical risk $R_n(\vartheta)$, but it takes into account a regularization term determined by the Kullback-Leibler divergence (denoted by KL, see (2.5) for a definition) to the prior distribution Π and weighted via the *inverse temperature parameter* $\lambda > 0$. This optimization problem has a unique solution given by

$$\Pi_\lambda(\mathrm{d}\vartheta \mid \mathcal{D}_n) \propto \exp\big(-\lambda R_n(\vartheta)\big)\Pi(\mathrm{d}\vartheta), \tag{2.2}$$

see Lemma 2.1 below. While (2.2) coincides with the classical Bayesian posterior distribution if $Y_i = f_\vartheta(\mathbf{X}_i) + \varepsilon_i$ with i.i.d. $\varepsilon_i \sim \mathcal{N}(0, n/(2\lambda))$, the so-called tempered likelihood $\exp(-\lambda R_n(\vartheta))$, see e.g. Guedj (2019), serves as a proxy for the unknown distribution of the observations given ϑ. As we will see, the method is indeed applicable under quite general assumptions on the regression model.

Based on the Gibbs posterior distribution the regression function can be estimated via a random draw from the posterior

$$\widehat{f}_\lambda \coloneqq f_{\widehat{\vartheta}_\lambda} \qquad \text{for} \qquad \widehat{\vartheta}_\lambda \mid \mathcal{D}_n \sim \Pi_\lambda(\cdot \mid \mathcal{D}_n), \tag{2.3}$$

or via the posterior mean

$$\bar{f}_\lambda \coloneqq \mathbb{E}\big[f_{\widehat{\vartheta}_\lambda} \,\big|\, \mathcal{D}_n\big] = \int f_\vartheta\, \Pi_\lambda(\mathrm{d}\vartheta \mid \mathcal{D}_n). \tag{2.4}$$

Another popular approach is to use the maximum a posteriori (MAP) estimator, but we focus on the previous two estimators.

As a benchmark for our estimators, we define the *oracle choice* for ϑ as

$$\vartheta^* \in \operatorname*{arg\,min}_{\vartheta\in\Theta} R(\vartheta) = \operatorname*{arg\,min}_{\vartheta\in\Theta} \mathcal{E}(\vartheta),$$

assuming that the minimization problem admits a solution, which will be the case in all settings that we consider. The minimization problem may have multiple solutions, in which case we simply choose one of them. This causes no issues as we are only interested in the corresponding risk $R(\vartheta^*) = R(f_{\vartheta^*})$. The oracle ϑ^* is not available to the practitioner, as it depends on the unknown distribution of $(\mathbf{X}, Y)$, but we will show that the performance of our estimators is almost as good as that of the oracle.

2.1 A general PAC-Bayes bound

Let μ, ν be probability measures on a measurable space $(E, \mathscr{A})$. The *Kullback-Leibler divergence* of μ with respect to ν is defined via

$$\mathrm{KL}(\mu \mid \nu) \coloneqq \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu, & \text{if } \mu \ll \nu, \\ \infty, & \text{otherwise.} \end{cases} \tag{2.5}$$

The following classical lemma is a key ingredient for PAC-Bayes bounds, cf. Catoni (2004, p. 159) or Alquier (2021). We include the short proof for the sake of completeness.

Lemma 2.1. *Let $h\colon E \to \mathbb{R}$ be a measurable function such that $\int \exp\circ h \,\mathrm{d}\mu < \infty$. With the convention $\infty - \infty = -\infty$ it then holds that*

$$\log\Big(\int \mathrm{e}^h \,\mathrm{d}\mu\Big) = -\inf_{\nu\ll\mu}\Big(\mathrm{KL}(\nu \mid \mu) - \int h \,\mathrm{d}\nu\Big), \tag{2.6}$$

where the infimum is taken over all probability measures $\nu \ll \mu$ on $(E, \mathscr{A})$. If additionally, h is bounded from above on the support of μ, then the infimum in (2.6) is attained for $\nu = \varrho$ with the Gibbs distribution ϱ, i.e. $\frac{\mathrm{d}\varrho}{\mathrm{d}\mu} \propto \mathrm{e}^h$.

Proof. For $D \coloneqq \int \mathrm{e}^h \,\mathrm{d}\mu$, we have $\mathrm{d}\varrho = D^{-1}\mathrm{e}^h \mathrm{d}\mu$ and obtain for all $\nu \ll \mu$:

$$\begin{aligned} 0 \leqslant \mathrm{KL}(\nu \mid \varrho) = \int \log\left(\tfrac{\mathrm{d}\nu}{\mathrm{d}\varrho}\right) \mathrm{d}\nu &= \int \log\left(\tfrac{\mathrm{d}\nu}{\mathrm{e}^h \mathrm{d}\mu / D}\right) \mathrm{d}\nu \\ &= \mathrm{KL}(\nu \mid \mu) - \int h \,\mathrm{d}\nu + \log\Big(\int \mathrm{e}^h \,\mathrm{d}\mu\Big). \end{aligned}$$

□

With this lemma at hand, we can derive a general PAC-Bayes bound. The basic proof strategy is standard in the PAC-Bayes literature, see e.g. Alquier & Biau (2013). However, the variant we present here allows for a more transparent view of an underlying concentration-type condition, instead of direct assumptions on the regression model.

Proposition 2.2 (PAC-Bayes bound)**.** *Set* $\mathcal{E}_n(\vartheta) \coloneqq R_n(\vartheta) - R_n(f)$*. Assume that*

$$\max\left\{\mathbb{E}\big[\exp\big(\lambda\big(\mathcal{E}(\vartheta) - \mathcal{E}_n(\vartheta)\big)\big)\big], \mathbb{E}\big[\exp\big(\lambda\big(\mathcal{E}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big]\right\} \leqslant \exp\big(C_{n,\lambda}\lambda\mathcal{E}(\vartheta)\big) \tag{2.7}$$

for some $\lambda > 0$ *and a constant* $C_{n,\lambda} \in (0, 1/2]$*. Then, we have for any* $\mathcal{D}_n$*-dependent (in a measurable way) probability measure* $\varrho \ll \Pi$ *that*

$$\mathcal{E}(\widehat{\vartheta}_\lambda) \leqslant 3\int \mathcal{E}\,\mathrm{d}\varrho + \frac{4}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(2/\delta)\big)$$

with probability of at least $1 - \delta$*.*

Remark 2.3. Here and in the following, the $1-\delta$ probability in takes into account the randomness of the data and of the estimate. For smaller $C_{n,\lambda}$, the 3 in front of the integrated excess risk can be improved to $1 + \tau$ for any $\tau > 0$ at the cost of a larger multiplicative constant in front of the remaining terms.

The proof uses neither the independence of the data $\mathcal{D}_n$, nor the explicit form of the risk R and the empirical risk R_n. In particular, the same result holds for a sample with dependent data points and generic R and R_n, where R_n is bounded from below. Therefore, Proposition 2.2 can be used as a general tool for proving PAC-Bayes oracle inequalities by verifying (2.7) and then choosing ϱ to balance $\int \mathcal{E}\,\mathrm{d}\varrho$ and $\mathrm{KL}(\varrho \mid \Pi)$. The trade-off here is that the integrated excess risk is small if ϱ has most of its mass around the oracle ϑ^*, whereas the Kullback-Leibler term is small if ρ is similar to the prior Π.

To verify the concentration inequality (2.7), we need some assumption on the dependence structure of the data $\mathcal{D}_n$. In particular, it holds in our setting of an i.i.d. sample under rather mild conditions on the model and the parameter class:

Assumption 2.A.

(a) ***Bounded regression function:*** *For some constant* $C \geqslant 1$*, we have* $\|f\|_\infty \leqslant C$*.*

(b) ***Bounded estimators:*** *For some constant* $\widetilde{C} \geqslant 1$*, we have* $\|f_\vartheta\|_\infty \leqslant \widetilde{C}$ *for all* $\vartheta \in \Theta$*.*

(c) ***Conditional sub-Gaussianity of observation noise:*** *There are constants* $\sigma, \Gamma > 0$

such that

$$\mathbb{E}[|\varepsilon|^k \mid \mathbf{X}] \leqslant \frac{k!}{2}\sigma^2\Gamma^{k-2} \ \text{a.s.}, \qquad \forall k \geqslant 2.$$

Lemma 2.4. *Grant Assumption 2.A and set* $V \coloneqq 8(C + \widetilde{C})(\Gamma \vee (C + \widetilde{C}))$ *and* $C_{n,\lambda} \coloneqq \frac{\lambda}{n}\frac{2((C+\widetilde{C})^2+4\sigma^2)}{1-V\lambda/n}$. *Then, we have for all* $\lambda \in [0, n/V)$ *that*

$$\max\big\{\mathbb{E}\big[\exp\big(\lambda\big(\mathcal{E}(\vartheta) - \mathcal{E}_n(\vartheta)\big)\big)\big], \mathbb{E}\big[\exp\big(\lambda\big(\mathcal{E}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big]\big\} \leqslant \exp\big(C_{n,\lambda}\lambda\mathcal{E}(\vartheta)\big).$$

2.2 Application to high-dimensional multi-index models

In this section, we demonstrate how the methodology related to Proposition 2.2 can be applied to high-dimensional multi-index models. In particular, we extend the analysis for single-index models by Alquier & Biau (2013). The results in this section are based on Steffen (2023b).

While we do not assume that the observations exactly follow a multi-index model, our method builds upon an approximation of the regression function of the form

$$f(\mathbf{x}) \approx g^*(W^*\mathbf{x}), \qquad \forall \mathbf{x} \in \mathbb{R}^p,$$

for some *active dimension* d^* substantially smaller than p, a sparse *dimension reduction matrix* $W^* \in \mathbb{R}^{d^*\times p}$ and a (measurable) *link function* $g^*\colon \mathbb{R}^{d^*} \to \mathbb{R}$. As a standard assumption in the theory of multi-index models, we suppose that the dimension reduction matrix is (semi-)orthogonal, i.e., $W^*(W^*)^\top = E_{d^*}$ is the identity matrix, see Xia (2008, Proposition 1.1). Indeed, this allows for the interpretation of $W^*\mathbf{X}$ as a rotation of the covariates, projected onto the first d^* coordinates followed by another rotation.

With a prior Π for the parameters (W, g), the Π-density of Gibbs posterior distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ in the multi-index setting reads as (up to normalization)

$$\frac{\mathrm{d}\Pi_\lambda(W, g \mid \mathcal{D}_n)}{\mathrm{d}\Pi} \propto \exp\big(-\lambda R_n(W, g)\big)$$

with a tuning parameter $\lambda > 0$ and empirical prediction risk

$$R_n(W^*, g^*) = \frac{1}{n}\sum_{i=1}^n \big(Y_i - g^*(W^*\mathbf{X}_i)\big)^2.$$

For clarity in this first application, we focus on the estimator obtained from drawing

$$\widehat{f}_\lambda = \widehat{g}_\lambda(\widehat{W}_\lambda\cdot) \qquad \text{with} \qquad (\widehat{W}_\lambda, \widehat{g}_\lambda) \mid \mathcal{D}_n \sim \Pi_\lambda(\cdot \mid \mathcal{D}_n). \tag{2.8}$$

In Chapter 3, we will demonstrate that the extension of our results to the posterior mean (2.4) is feasible.

We will choose a sieve prior that prefers models with a low active dimension, sparse dimension reduction matrices and regular link functions. Let Π be supported on $\bigcup_{d=1}^{p} \mathcal{S}_d \times \mathcal{G}_d$ for some classes $\mathcal{S}_d$ and $\mathcal{G}_d$ for W and g, respectively. For $\mathcal{S}_d$ we will study a class of sparse matrices while $\mathcal{G}_d$ will be given by finite wavelet approximations. The prior is uniform for a given sparsity and a wavelet projection level. The posterior weights each tuple of parameters (W, g) based on its empirical performance (with respect to the empirical loss function) on the data, where the tuning parameter λ determines the impact of $R_n(W, g)$ in comparison to the prior beliefs.

We will prove an oracle inequality verifying our estimator is not worse than the optimal choices for $W \in \mathcal{S}_d$ and $g \in \mathcal{G}_d$ for any d. In particular, the overall quality of the method depends on the approximation properties of the spaces $\mathcal{S}_d$ and $\mathcal{G}_d$.

2.2.1 Construction of the prior

To construct the prior, we will introduce for any dimension $d = 1, \dots, p$ classes $\mathcal{S}_d$ and $\mathcal{G}_d$ together with priors μ_d and ν_d for the dimension reduction matrix W and the link function g, respectively. Based on that we can then define the prior Π on $\bigcup_{d=1}^{p} \mathcal{S}_d \times \mathcal{G}_d$.

We start with a fixed active dimension $d \in \{1, \dots, p\}$. While it is common in the literature to assume that the dimension reduction matrix W^* is (semi-)orthogonal, we will not impose this restriction for the estimation method. Instead, we only require that the candidate matrices have ℓ^2-standardized rows, i.e. for $W = (w_1, \dots, w_d)^\top \in \mathbb{R}^{d \times p}$ with row vectors $w_i = (w_{i,1}, \dots, w_{i,p}) \in \mathbb{R}^p$ we impose $|w_i| = 1$. To encode sparsity, let

$$\mathcal{I}_d \coloneqq \left\{ I \,\middle|\, \emptyset \neq I \coloneqq I_1 \times \cdots \times I_d,\, I_1, \dots, I_d \subseteq \{1, \dots, p\} \right\}$$

contain all potential sets of *active coordinates,* that is I_i describes the active coordinates in the i-th argument of the link function. For $I = I_1 \times \cdots \times I_d \in \mathcal{I}_d$ the number of active coordinates is $\|I\| \coloneqq \sum_{i=1}^{d} |I_i|$, where $|I_i|$ denotes the cardinality of I_i. Note that $\emptyset \neq I = I_1 \times \cdots \times I_d$ already implies $I_1, \dots, I_d \neq \emptyset$. The parameter set $\mathcal{S}_d(I)$ of sparse dimension reduction matrices is given by

$$\begin{aligned} \mathcal{S}_d(I) &\coloneqq \left\{ W = (w_1, \dots, w_d)^\top \in \mathbb{R}^{d \times p} \mid w_i \in \mathcal{S}(I_i),\, i = 1, \dots, d \right\}, \qquad \text{where} \\ \mathcal{S}(I_i) &\coloneqq \left\{ w_i = (w_{i,1}, \dots, w_{i,p}) \in \mathbb{R}^p \mid |w_i| = 1, \forall j \notin I_i : w_{i,j} = 0 \right\}. \end{aligned}$$

Finally, we define $\mathcal{S}_d = \bigcup_{I\in\mathcal{I}_d} \mathcal{S}_d(I)$. Note that $\mathcal{S}_d(I) \supseteq \widetilde{S}_d(I)$ for

$$\widetilde{\mathcal{S}}_d(I) \coloneqq \{W = (w_1, \dots, w_d)^\top \in \mathbb{R}^{d\times p} \mid |w_1| = \cdots = |w_d| = 1, w_{i,j} \neq 0 \text{ iff } j \in I_i,\ j = 1, \dots, p\}.$$

In $\widetilde{S}_d(I)$ the index set I exactly describes the sparsity of W. However, we consider the prior on the compact set $\mathcal{S}_d(I)$ to ensure the existence of solutions to minimization problems over $\mathcal{S}_d(I)$ and thus the existence of an oracle dimension reduction matrix.

To construct a prior measure μ_d on $\mathcal{S}_d$, we use the uniform distribution on the set of dimension reduction matrices with a given active dimension d and with sparsity $i = \|I\|$. These uniform distributions are then weighted geometrically such that sparse dimension reduction matrices are preferred by the prior. Denoting the uniform distribution on $\mathcal{S}_d(I)$ by $\mu_{d,I}$, the prior measure on $\mathcal{S}_d$ is thus given by the mixture

$$\mu_d \coloneqq \sum_{i=d}^{dp} 2^{-i+d-1} \frac{1}{|\mathcal{I}_{d,i}|} \sum_{I\in\mathcal{I}_{d,i}} \mu_{d,I} \Big/ (1 - 2^{(1-p)d-1}) \qquad \text{where} \qquad \mathcal{I}_{d,i} \coloneqq \{I \in \mathcal{I}_d \mid \|I\| = i\}.$$

Here and in the following two analogous constructions, the basis 2 of the geometric weights can be replaced by an arbitrary fixed $a > 1$. The theoretical results remain unchanged up to constants.

To define a class $\mathcal{G}_d$ and a prior ν_d for the link function, we will use a multivariate tensor product wavelet basis on $\mathbb{R}^d$, see e.g. Daubechies (1992); Giné & Nickl (2016). Let φ and ψ be a continuously differentiable scaling and wavelet function on $\mathbb{R}$, respectively, and write $\psi_0 \coloneqq \varphi$, $\psi_1 \coloneqq \psi$. We will use compactly supported regular Daubechies wavelets. For $M \in \mathbb{N}_0, N \in \mathbb{N}$ we define the index set

$$\begin{aligned}\mathcal{Z}^d_{M,N} \coloneqq &\{l = (0, l_2, 0) \mid l_2 \in \mathbb{Z}^d,\ |l_2|_\infty \leqslant N\} \\ &\cup \big\{l = (l_1, l_2, l_3) \in \mathbb{N}_0 \times \mathbb{Z}^d \times \{0,1\}^d \mid l_1 \leqslant M, |l_2|_\infty \leqslant 2^{l_1} N, l_3 \neq 0\big\},\end{aligned}$$

where l_1 is the approximation level, l_2 is a shift parameter and l_3 is due to the tensor structure. The system $(\Psi_l)_{l\in\mathcal{Z}^d_{\infty,\infty}}$ with

$$\Psi_l(\mathbf{x}) \coloneqq 2^{l_1 d/2} \prod_{i=1}^d \psi_{l_{3,i}}(2^{l_1} x_i - l_{2,i}), \qquad \mathbf{x} \in \mathbb{R}^d, l = (l_1, l_2, l_3) \in \mathcal{Z}^d_{\infty,\infty},$$

is an orthonormal basis of $L^2(\mathbb{R}^d)$. In particular, each $g \in L^2(\mathbb{R}^d)$ admits a wavelet series representation $g = \sum_{l\in\mathcal{Z}^d_{\infty,\infty}} \langle g, \Psi_l\rangle \Psi_l$. Throughout, we fix a sufficiently large constant $N \in \mathbb{N}$

and abbreviate $\mathcal{Z}_M^d \coloneqq \mathcal{Z}_{M,N}^d$. For $\xi > 0$ we define the compact wavelet coefficient ball

$$\mathcal{B}_{d,M}(\xi) \coloneqq \big\{\beta \in \mathbb{R}^{\mathcal{Z}_M^d} \,\big|\, \|\beta\|_{\mathcal{B}} \leqslant \xi\big\}, \qquad \text{where}$$
$$\|\beta\|_{\mathcal{B}} \coloneqq L^d \sum_{l \in \mathcal{Z}_M^d} 2^{l_1(d/2+1)} |\beta_l|, \qquad \text{with} \qquad L \coloneqq \|\psi\|_\infty \vee \|\varphi\|_\infty \vee \|\psi'\|_\infty \vee \|\varphi'\|_\infty \vee 1,$$

which determines the finite dimensional approximation space

$$\mathcal{G}_{d,M}(\xi) \coloneqq \big\{g = \Phi_{d,M}(\beta) \,\big|\, \beta \in \mathcal{B}_{d,M}(\xi)\big\} \qquad \text{via} \qquad \Phi_{d,M}(\beta) \coloneqq \sum_{l \in \mathcal{Z}_M^d} \beta_l \Psi_l, \beta \in \mathbb{R}^{\mathcal{Z}_M^d}.$$

For any $g = \Phi_{d,M}(\beta)$ we write $\|g\|_{\mathcal{B}} \coloneqq \|\beta\|_{\mathcal{B}}$ which corresponds to the Besov norm with regularity $1 + d$ and integrability parameter 1 on $\operatorname{span}\{\Psi_l : l \in \mathcal{Z}_M^d\}$. In particular, we have for any $g \in \mathcal{G}_{d,M}(\xi)$

$$\|g\|_\infty \leqslant \|g\|_{\mathcal{B}} \leqslant \xi \qquad \text{and} \qquad \|(\nabla g)_i\|_\infty \leqslant \|g\|_{\mathcal{B}} \leqslant \xi, \qquad \forall i \in \{1, \dots, d\}. \tag{2.9}$$

For $C > 0$ we set $\mathcal{G}_d \coloneqq \bigcup_{M=0}^n \mathcal{G}_{d,M}(C+1)$.

The prior ν_d on $\mathcal{G}_d$ is defined as a random coefficient prior with uniformly distributed coefficients on $\mathcal{G}_{d,M}(C+1)$ and geometrically decreasing weights in the approximation level M. To this end, let $\widetilde{\nu}_{d,M}$ be the uniform distribution on $\mathcal{B}_{d,M}(C+1)$ and let $\nu_{d,M} \coloneqq \widetilde{\nu}_{d,M}(\Phi_{d,M}^{-1}(\cdot))$ denote the push-forward measure of $\widetilde{\nu}_{d,M}$ under $\Phi_{d,M}$. Then, we set

$$\nu_d \coloneqq \sum_{M=0}^n 2^{-M} \nu_{d,M} \Big/ (2 - 2^{-n}).$$

We can now define the prior for a fixed active dimension d as the product measure $\pi_d \coloneqq \mu_d \otimes \nu_d$. Finally, we mix over all possible active dimensions to account for the fact that d^* is unknown. Encoding a preference for simple models, i.e. small active dimensions, via weights 2^{-d}, the final prior on $\bigcup_{d=1}^p \mathcal{S}_d \times \mathcal{G}_d$ is given by

$$\Pi = \sum_{d=1}^p 2^{-d} \pi_d \Big/ (1 - 2^{-p}).$$

Note that the structure of the prior ensures that drawing from Π will yield a link function and a dimension reduction matrix with matching active dimension.

2.2.2 Oracle inequalities

For an active dimension $d \in \{1, \dots, p\}$, an active index set $I \in \mathcal{I}_d$ of the dimension reduction matrix and an approximation level $M \in \{0, \dots, n\}$ of the link function, we define the *oracle choice* on $\mathcal{S}_d(I) \times \mathcal{G}_{d,M}(C)$ as

$$(W^*_{d,I}, g^*_{d,M}) \coloneqq \operatorname*{arg\,min}_{(W,g)\in\mathcal{S}_d(I)\times\mathcal{G}_{d,M}(C)} R(W, g). \tag{2.10}$$

Note that the minimization in g is over $\mathcal{G}_{d,M}(C)$, whereas the prior is defined on $\mathcal{G}_{d,M}(C+1)$ which ensures that a small neighborhood of $g^*_{d,M}$ is contained in the support of the prior. A solution to the minimization problem in (2.10) always exists since we have equivalently

$$(W^*_{d,I}, \beta^*_{d,M}) = \operatorname*{arg\,min}_{(W,\beta)\in\mathcal{S}_d(I)\times\mathcal{B}_{d,M}(C)} \mathbb{E}\big[\big(Y - \Phi_{d,M}(\beta)(W\mathbf{X})\big)^2\big]$$

with compact $\mathcal{S}_d(I) \times \mathcal{B}_{d,M}(C)$ and continuous $(W, \beta) \mapsto \mathbb{E}\big[\big(Y - \Phi_{d,M}(\beta)(W\mathbf{X})\big)^2\big]$. Our main result in this section gives a theoretical guarantee that the estimator $(\widehat{W}_\lambda, \widehat{g}_\lambda)$ from (2.8) is almost as good as the best oracle $(W^*_{d,I}, g^*_{d,M})$ for all possible active dimensions in terms of the excess risk. To this end, we need some mild assumptions on the regression model.

Assumption 2.B.

(a) ***Bounded regression function:*** *For some constant $C \geqslant 1$ we have $\|f\|_\infty \leqslant C$.*

(b) ***Bounded inputs:*** *For some constant $K \geqslant 1$ we have $|\mathbf{X}|_\infty \leqslant K$ a.s.*

(c) ***Conditional sub-Gaussianity of observation noise:*** *There are constants $\sigma, \Gamma > 0$ such that*

$$\mathbb{E}[|\varepsilon|^k|\mathbf{X}] \leqslant \frac{k!}{2}\sigma^2\Gamma^{k-2} \text{ a.s.}, \qquad \forall k \geqslant 2.$$

We obtain the following non-asymptotic oracle inequality. It generalizes Alquier & Biau (2013, Theorem 2) not only with respect to the multi-index approach with unknown active dimension, but also with respect to some technical but practically relevant aspects such as the ℓ^2-normalization of W and the wavelet basis.

Theorem 2.5 (PAC-Bayes oracle inequality)**.** *Under Assumption 2.B there are constants $Q_0, Q_1 > 0$ depending only on $C, \Gamma, \sigma > 0$ such that for $\lambda = n/Q_0$ and sufficiently large n we have for all $\delta \in (0, 1)$ with a probability of at least $1 - \delta$ that*

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant \min_{d,I,M} \Big(3\mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \frac{Q_1}{n}\big(\|I\| \log(p \vee n) + 16^d N^d 2^{dM} \log(n) + \log(2/\delta)\big)\Big),$$

where the minimum is taken over all triplets (d, I, M) *with* $d \in \{1, \dots, p\}$, $I \in \mathcal{I}_d$ *and* $M \in \{0, \dots, n\}$.

Remark 2.6. An explicit admissible choice for λ is $\lambda = n/((2C+1)(\Gamma \vee (2C+1)) + 4((2C+1)^2 + 4\sigma^2))$. The dependence of Q_1 on C, Γ, σ is at most quadratic and $n \geqslant n_0 = 5 \vee (C+1) \vee K$ is sufficiently large.

The right-hand side of the oracle inequality can be interpreted similarly to the classical bias-variance decomposition in non-parametric statistics. The first term

$$\mathcal{E}(W^*_{d,I}, g^*_{d,M}) = \mathbb{E}[(g^*_{d,M}(W^*_{d,I}\mathbf{X}) - f(\mathbf{X}))^2]$$

quantifies the approximation error while second term is an upper bound for the stochastic error. In particular, we recover $\|I\| \log(p \vee n)/n$ (or $\|I\| \log(p)/n$ if $p \geqslant n$) as the typical error term for estimating sparse matrices with sparsity $\|I\|$, see van de Geer et al. (2011), while $16^d N^d 2^{dM} \log(n)/n$ is due to the estimation of $\mathcal{O}\big(16^d N^d 2^{dM}\big)$ many wavelet coefficients each with (squared) accuracy $\log(n)/n$ paying a logarithmic price for adaptivity.

The minimum over all (d, I, M) in the upper bound shows that the estimator adapts to the active dimension, the sparsity of the dimension reduction matrix and the regularity of the link function. One can show the same result in a multi-index model with a known active dimension d^* by using π_{d^*} as a prior instead of Π. The only difference (up to a different constant Q_1) in the result is that the minimum in the upper bound is only taken over all pairs $(I, M) \in \mathcal{I}_{d^*} \times \{0, \dots, n\}$. Consequently, no additional price is paid for not knowing the true active dimension of the model.

In the well-specified setting and under assumptions on the distribution of $W^*\mathbf{X}$ as well as a Besov-type regularity assumption on the link function, we derive explicit convergence rates from Theorem 2.5.

Assumption 2.C.

(a) ***Multi-index model:*** *There exist* $d^* \in \{1, \dots, p\}$, $W^* \in \mathcal{S}^{d^*}$ *and* $g^* \colon \mathbb{R}^{d^*} \to \mathbb{R}$ *such that* $f = g^*(W^*\cdot)$.

(b) ***Bounded dimension reduced inputs:*** *For* $B_1 \geqslant 1$, *we have* $|W^*\mathbf{X}|_\infty \leqslant B_1$.

(c) ***Lebesgue density of dimension reduced inputs:*** $W^*\mathbf{X}$ *has a Lebesgue density on* $\mathbb{R}^{d^*}$ *bounded by a constant* $B_2 \geqslant 1$.

For the true dimension reduction matrix W^* we write $\|W^*\|_0 \coloneqq \|I^*\|$ for the minimal (with

respect to $\|\cdot\|$) $I^* \in \mathcal{I}_{d^*}$ such that $W^* \in \mathcal{S}_{d^*}(I^*)$. The regularity of g^* will be measured in terms of its Besov norm. We recover Sobolev balls for $q = 2$, cf. Giné & Nickl (2016, (4.164)).

Definition 2.7. The Besov ellipsoid in $\mathbb{R}^{d^*}$ with regularity $\alpha > 0$ and integrability parameter $q \in [0, \infty)$ is given by

$$B^{\alpha}_{q,d^*}(\xi) \coloneqq \Big\{ g \in L^2(\mathbb{R}^{d^*}) \,\Big|\, \sum_{l \in \mathcal{Z}^{d^*}_{\infty,\infty}} 2^{q l_1 \alpha} |\langle g, \Psi_l \rangle|^q \leqslant \xi^q \Big\} \tag{2.11}$$

for a radius $\xi > 0$.

Corollary 2.8 (Convergence rate)**.** *Let the assumptions of Theorem 2.5 be fulfilled in addition to Assumption 2.C. Take $\lambda = n/Q_0$ with Q_0 from Theorem 2.5. Suppose that $g^* \in B^{\alpha}_{2,d^*}(\xi)$ with $\xi = C(L^{d^*} 2 N^{d^*/2} 16^{d^*/2})^{-1}$ for some $\alpha > 2 + d^*$. Then, for sufficiently large n and with a probability of at least $1 - \delta$, we have*

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant Q_2 \Big(\frac{\log n}{n} \Big)^{\frac{2\alpha}{2\alpha + d^*}} + Q_2 \Big(\frac{\|W^*\|_0 \log(p \vee n)}{n} + \frac{\log(2/\delta)}{n} \Big),$$

where Q_2 is a constant only depending on $C, \Gamma, \sigma, N, B_1, B_2$ and d^*.

Remark 2.9. If W^* is sparse (i.e. $\|W^*\|_0$ is small), then the dominating term in the upper bound of the excess risk of the PAC-Bayesian estimator is of order

$$\Big(\frac{\log n}{n} \Big)^{\frac{2\alpha}{2\alpha + d^*}},$$

which is the usual minimax-optimal rate (up to a logarithmic factor) for such estimation problems, see e.g. Tsybakov (2009). Note that if d^* is substantially smaller than p, then we have successfully circumvented the curse of dimensionality, since the dimension which appears in the rate is now only d^*. As an alternative to the wavelet construction, one can use the multivariate trigonometric system on $[-1, 1]^{d^*}$, assume $\mathbf{X} \in [-1, 1]^p$ and ℓ^1-standardized rows of W^* (which ensures $W^* \mathbf{X} \in [-1, 1]^{d^*}$) leading to a more direct generalization of Alquier & Biau (2013). However, the orthogonality assumption on W^* seems more natural and is in line with the literature.

2.3 Proofs

We begin with the proofs of the general PAC-Bayes bound and the concentration inequality. Then, we apply them to prove the results from Section 2.2.

2.3.1 Proof of Proposition 2.2

The concentration inequality from the assumption gives

$$\mathbb{E}\big[\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(\vartheta)-\lambda\mathcal{E}_n(\vartheta)-\log(\delta^{-1})\big)\big]\vee\mathbb{E}\big[\exp\big(\lambda\mathcal{E}_n(\vartheta)-\lambda(1+C_{n,\lambda})\mathcal{E}(\vartheta)-\log(\delta^{-1})\big)\big]\leqslant\delta.$$

Integrating in ϑ with respect to the prior probability measure Π and applying Fubini's theorem, we conclude

$$\begin{aligned}&\mathbb{E}\Big[\int\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(\vartheta)-\lambda\mathcal{E}_n(\vartheta)-\log(\delta^{-1})\big)\,\mathrm{d}\Pi(\vartheta)\Big]\leqslant\delta\qquad\text{and}\\&\mathbb{E}\Big[\int\exp\big(\lambda\mathcal{E}_n(\vartheta)-\lambda(1+C_{n,\lambda})\mathcal{E}(\vartheta)-\log(\delta^{-1})\big)\,\mathrm{d}\Pi(\vartheta)\Big]\leqslant\delta.\end{aligned}\tag{2.12}$$

For the posterior distribution $\Pi_\lambda(\cdot\mid\mathcal{D}_n)\ll\Pi$ with corresponding Radon-Nikodym density

$$\frac{\mathrm{d}\Pi_\lambda(\vartheta\mid\mathcal{D}_n)}{\mathrm{d}\Pi}=D_\lambda^{-1}\exp(-\lambda R_n(\vartheta)),\qquad D_\lambda\coloneqq\int\exp(-\lambda R_n(\vartheta))\,\mathrm{d}\Pi(\vartheta)$$

with respect to Π, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_n,\widehat{\vartheta}\sim\Pi_\lambda(\cdot|\mathcal{D}_n)}\big[\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(\widehat{\vartheta})-\lambda\mathcal{E}_n(\widehat{\vartheta})-\log(\delta^{-1})+\lambda R_n(\widehat{\vartheta})+\log D_\lambda\big)\big]\\&\quad=\mathbb{E}_{\mathcal{D}_n,\widehat{\vartheta}\sim\Pi_\lambda(\cdot|\mathcal{D}_n)}\Big[\exp\Big(\lambda(1-C_{n,\lambda})\mathcal{E}(\widehat{\vartheta})-\lambda\mathcal{E}_n(\widehat{\vartheta})-\log(\delta^{-1})-\log\Big(\frac{\mathrm{d}\Pi_\lambda(\widehat{\vartheta}\mid\mathcal{D}_n)}{\mathrm{d}\Pi}\Big)\Big)\Big]\\&\quad=\mathbb{E}_{\mathcal{D}_n}\Big[\int\exp\big(\lambda(1-C_{n,\lambda})\mathcal{E}(\vartheta)-\lambda\mathcal{E}_n(\vartheta)-\log(\delta^{-1})\big)\,\mathrm{d}\Pi(\vartheta)\Big]\leqslant\delta.\end{aligned}$$

Since $\mathbb{1}_{[0,\infty)}(x)\leqslant\mathrm{e}^{\lambda x}$ for all $x\in\mathbb{R}$, we deduce for $\widehat{\vartheta}\sim\Pi_\lambda(\cdot\mid\mathcal{D}_n)$ with a probability not larger than δ that

$$\big(1-C_{n,\lambda}\big)\mathcal{E}(\widehat{\vartheta})-\mathcal{E}_n(\widehat{\vartheta})+R_n(\widehat{\vartheta})-\lambda^{-1}\big(\log(\delta^{-1})-\log D_\lambda\big)\geqslant0.$$

As $1-C_{n,\lambda}>0$, we thus have with a probability of at least $1-\delta$ that

$$\mathcal{E}(\widehat{\vartheta})\leqslant(1-C_{n,\lambda})^{-1}\big(-R_n(f)+\lambda^{-1}\big(\log(\delta^{-1})-\log D_\lambda\big)\big).$$

Lemma 2.1 yields

$$-\log D_\lambda=-\log\Big(\int\exp(-\lambda R_n(\vartheta))\,\mathrm{d}\Pi(\vartheta)\Big)=\inf_{\varrho\ll\Pi}\Big(\mathrm{KL}(\varrho\mid\Pi)+\int\lambda R_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big).$$

Therefore, for any $\varrho \ll \Pi$, it holds with probability of at least $1-\delta$ that

$$\mathcal{E}(\widehat{\vartheta}) \leqslant (1-C_{n,\lambda})^{-1}\Big(\int \mathcal{E}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta) + \lambda^{-1}\big(\log(\delta^{-1}) + \mathrm{KL}(\varrho \mid \Pi)\big)\Big).$$

In order to reduce the integral $\int \mathcal{E}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)$ to $\int \mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta)$, we use Jensen's inequality and (2.12) to obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n}\Big[\exp\Big(\int \lambda\mathcal{E}_n(\vartheta) - \lambda(1+C_{n,\lambda})\mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta) - \mathrm{KL}(\varrho \mid \Pi) - \log(\delta^{-1})\Big)\Big] \\
&\quad= \mathbb{E}_{\mathcal{D}_n}\Big[\exp\Big(\int \lambda\mathcal{E}_n(\vartheta) - \lambda(1+C_{n,\lambda})\mathcal{E}(\vartheta) - \log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(\vartheta)\Big) - \log(\delta^{-1})\,\mathrm{d}\varrho(\vartheta)\Big)\Big] \\
&\quad\leqslant \mathbb{E}_{\mathcal{D}_n,\vartheta\sim\varrho}\Big[\exp\Big(\lambda\mathcal{E}_n(\vartheta) - \lambda(1+C_{n,\lambda})\mathcal{E}(\vartheta) - \log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(\vartheta)\Big) - \log(\delta^{-1})\Big)\Big] \\
&\quad= \mathbb{E}_{\mathcal{D}_n}\Big[\int \exp\big(\lambda\mathcal{E}_n(\vartheta) - \lambda(1+C_{n,\lambda})\mathcal{E}(\vartheta) - \log(\delta^{-1})\big)\,\mathrm{d}\Pi(\vartheta)\Big] \leqslant \delta.
\end{aligned}$$

Using $\mathbb{1}_{[0,\infty)}(x) \leqslant \mathrm{e}^{\lambda x}$ again, we deduce that with probability of at least $1-\delta$

$$\int \mathcal{E}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta) \leqslant (1+C_{n,\lambda})\int \mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta) + \lambda^{-1}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(\delta^{-1})\big).$$

Therefore, we conclude that with probability of at least $1-2\delta$

$$\mathcal{E}(\widehat{\vartheta}) \leqslant (1-C_{n,\lambda})^{-1}\Big((1+C_{n,\lambda})\int \mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta) + \frac{2}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(\delta^{-1})\big)\Big),$$

which yields the claimed bound since $C_{n,\lambda} \leqslant 1/2$. □

2.3.2 Proof of Lemma 2.4

We can rewrite $\mathcal{E}_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n Z_i$ with centered and independent random variables

$$Z_i \coloneqq (Y_i - f_\vartheta(\mathbf{X}_i))^2 - (Y_i - f(\mathbf{X}_i))^2 = -\big(2\varepsilon_i + f(\mathbf{X}_i) - f_\vartheta(\mathbf{X}_i)\big)\big(f_\vartheta(\mathbf{X}_i) - f(\mathbf{X}_i)\big).$$

Since f and f_ϑ are bounded by C and $\widetilde{C}$, respectively, and ε_i is sub-Gaussian we have

$$\mathbb{E}\big[Z_i^2\big] = \mathbb{E}\big[\big(2\varepsilon_i + f(\mathbf{X}_i) - f_\vartheta(\mathbf{X}_i)\big)^2\big(f_\vartheta(\mathbf{X}_i) - f(\mathbf{X}_i)\big)^2\big] \leqslant 2(4\sigma^2 + (C+\widetilde{C})^2)\mathcal{E}(\vartheta) =: U$$

and for $k \geqslant 3$

$$\begin{aligned}
\mathbb{E}[(Z_i)_+^k] &\leqslant \mathbb{E}\big[|2\varepsilon_i + f(\mathbf{X}_i) - f_\vartheta(\mathbf{X}_i)|^k |f_\vartheta(\mathbf{X}_i) - f(\mathbf{X}_i)|^{k-2}(f_\vartheta(\mathbf{X}_i) - f(\mathbf{X}_i))^2\big] \\
&\leqslant (C+\widetilde{C})^{k-2}\mathbb{E}\big[|2\varepsilon_i + f(\mathbf{X}_i) - f_\vartheta(\mathbf{X}_i)|^k (f_\vartheta(\mathbf{X}) - f(\mathbf{X}))^2\big]
\end{aligned}$$

$$\begin{aligned}
&\leqslant (C+\widetilde{C})^{k-2}2^{k-1}\big(k!2^{k-1}\sigma^2\Gamma^{k-2}+(C+\widetilde{C})^k\big)\mathcal{E}(\vartheta)\\
&\leqslant (C+\widetilde{C})^{k-2}k!8^{k-2}\big(\Gamma^{k-2}\vee(C+\widetilde{C})^{k-2}\big)U\\
&= k!UV^{k-2}.
\end{aligned}$$

In view of $\mathbb{E}[\mathcal{E}_n(\vartheta)] = \mathcal{E}(\vartheta)$, a variant of Bernstein's inequality, see Massart (2007, inequality (2.21)), yields for $\lambda \in (0, n/V)$ that

$$\mathbb{E}\big[\exp\big(\lambda\big(\mathcal{E}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\big)\big] \leqslant \exp\Big(\frac{U\lambda^2}{n(1-V\lambda/n)}\Big) = \exp\big(C_{n,\lambda}\lambda\mathcal{E}(\vartheta)\big). \qquad \square$$

2.3.3 Proof of Theorem 2.5

We extend the proof strategy by Alquier & Biau (2013) to the multi-index setting with unknown active dimension.

Assumption 2.B together with (2.9) ensures that we can apply Lemma 2.4 with $V = 8(2C+1)(\Gamma\vee(2C+1))$ to obtain for $\lambda = n/(V+4((2C+1)^2+4\sigma^2))$ and $\varrho \ll \Pi$ that

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant 3\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) + \frac{Q_3}{n}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(2/\delta)\big) \tag{2.13}$$

with a probability of at least $1-\delta$, where the constant Q_3 only depends on C, Γ and σ.

To choose ϱ, we fix some triplet (d, I, M) with $d \in \{1,\dots,p\}$, $I = I_1 \times \cdots \times I_d \in \mathcal{I}_d$, $M \in \{0,\dots,n\}$ as well as $\eta, \gamma \in (0,1]$ and set

$$\varrho := \varrho_{d,I,M,\eta,\gamma} := \varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma}, \tag{2.14}$$

where $\varrho^1_{d,I,\eta}$ and $\varrho^2_{d,M,\gamma}$ are the uniform distribution with respect to $\mu_{d,I}$ and $\nu_{d,M}$ on a ball of radius η and γ around the oracle $W^*_{d,I} = (w^*_{d,I,1},\dots,w^*_{d,I,d})^\top \in \mathbb{R}^{d\times p}$ and $g^*_{d,M}$, respectively. Specifically, we set

$$\begin{aligned}
&\frac{\mathrm{d}\varrho^1_{d,I,\eta}}{\mathrm{d}\mu_{d,I}}(W) := \prod_{i=1}^d \frac{\mathrm{d}\varrho^{1,i}_{d,I,\eta}}{\mathrm{d}\mu_{I_i}}(w_i),\ \forall W = (w_1,\dots,w_d)^\top, \quad \text{where} \\
&\frac{\mathrm{d}\varrho^{1,i}_{d,I,\eta}}{\mathrm{d}\mu_{I_i}}(w_i) \propto \mathbb{1}_{\{|w_i - w^*_{d,I,i}|\leqslant\eta\}} \quad \text{and} \quad \frac{\mathrm{d}\varrho^2_{d,M,\gamma}}{\mathrm{d}\nu_{d,M}}(g) \propto \mathbb{1}_{\{\|g-g^*_{d,M}\|_{\mathcal{B}}\leqslant\gamma\}},
\end{aligned} \tag{2.15}$$

where μ_{I_i} denotes the uniform distribution on $\mathcal{S}(I_i)$. To complete the proof, we need to bound the terms on the right hand side of (2.13) for this choice of ϱ.

First, we deal with the Kullback-Leibler divergence term using the following two lemmas:

Lemma 2.10. *For $\varrho = \varrho_{d,I,M,\eta,\gamma}$ from (2.14) and with $\pi_{d,I,M} = \mu_{d,I} \otimes \nu_{d,M}$, we have*

$$\mathrm{KL}(\varrho \mid \Pi) \leqslant \|I\| \log(ep) + (\|I\| + M + 2)\log(2) + \mathrm{KL}(\varrho \mid \pi_{d,I,M}) =: T_1 + \mathrm{KL}(\varrho \mid \pi_{d,I,M}).$$

Lemma 2.11. *For $\varrho = \varrho_{d,I,M,\eta,\gamma}$ from (2.14) and with $\pi_{d,I,M} = \mu_{d,I} \otimes \nu_{d,M}$, we have*

$$\mathrm{KL}(\varrho \mid \pi_{d,I,M}) \leqslant \|I\| \log(5/\eta) + 16^d N^d 2^{dM} \log((C+1)/\gamma) =: T_2.$$

Thus,

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant 3 \int \mathcal{E}(W, g) \,\mathrm{d}\varrho(W, g) + \frac{Q_3}{n}\big(T_1 + T_2 + \log(2/\delta)\big) \tag{2.16}$$

with a probability of at least $1 - \delta$.

Second, we control the integral term in (2.16) by splitting it into

$$\begin{aligned}
\int \mathcal{E}(W, g) \,\mathrm{d}\varrho(W, g) = \mathcal{E}(W^*_{d,I}, g^*_{d,M}) &\\
&+ \int \mathbb{E}[(g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X}))^2] \,\mathrm{d}\varrho(W, g) \\
&+ \int \mathbb{E}[(g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X}))^2] \,\mathrm{d}\varrho(W, g) \\
&+ \int \mathbb{E}[2(Y - g^*_{d,M}(W^*_{d,I}\mathbf{X}))(g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X}))] \,\mathrm{d}\varrho(W, g) \\
&+ \int \mathbb{E}[2(Y - g^*_{d,M}(W^*_{d,I}\mathbf{X}))(g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X}))] \,\mathrm{d}\varrho(W, g) \\
&+ \int \mathbb{E}[2(g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X}))(g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X}))] \,\mathrm{d}\varrho(W, g) \\
=: \mathcal{E}(W^*_{d,I}, g^*_{d,M}) + U_1 + U_2 + U_3 + U_4 + U_5 &
\end{aligned} \tag{2.17}$$

and treating the terms $U_1, \ldots, U_5$ sequentially.

Similarly to (2.9), $g = \Phi_{d,M}(\beta) \in \mathcal{G}_{d,M}(C+1)$ with $\|\beta - \beta^*_{d,M}\|_{\mathcal{B}} \leqslant \gamma$ implies

$$\|g - g^*_{d,M}\|_\infty \leqslant \|g - g^*_{d,M}\|_{\mathcal{B}} = \|\beta - \beta^*_{d,M}\|_{\mathcal{B}} \leqslant \gamma.$$

As a consequence, we obtain

$$U_1 = \int \mathbb{E}[(g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X}))^2] \,\mathrm{d}\varrho^2_{d,M,\gamma}(g) \leqslant \int \sup_{\mathbf{x} \in \mathbb{R}^d} (g^*_{d,M}(\mathbf{x}) - g(\mathbf{x}))^2 \,\mathrm{d}\varrho^2_{d,M,\gamma}(g) \leqslant \gamma^2. \tag{2.18}$$

Any $g \in \mathcal{G}_{d,M}(C+1)$ is differentiable as a linear combination of only finitely many basis elements.

Therefore, applying the fundamental theorem of calculus to the mapping

$$h\colon [-1,1] \to \mathbb{R},\ s \mapsto g((W + s(W^*_{d,I} - W))\mathbf{X}(\omega))$$

with a fixed $\omega \in \Omega$ (which we omit from here on) yields

$$g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X}) = \int_0^1 h'(s)\,\mathrm{d}s = \Big\langle (W^*_{d,I} - W)\mathbf{X}, \int_0^1 \nabla g((W + s(W^*_{d,I} - W))\mathbf{X})\,\mathrm{d}s \Big\rangle$$

and combined with (2.9) we obtain

$$\begin{aligned}
|g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X})| &= \Big|\Big\langle (W^*_{d,I} - W)\mathbf{X}, \int_0^1 \nabla g((W + s(W^*_{d,I} - W))\mathbf{X})\,\mathrm{d}s \Big\rangle\Big| \\
&\leqslant \big|(W^*_{d,I} - W)\mathbf{X}\big| \int_0^1 \nabla g((W + s(W^*_{d,I} - W))\mathbf{X})\,\mathrm{d}s\Big| \\
&\leqslant |\mathbf{X}|\Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\Big)^{1/2}\sqrt{d}C \\
&\leqslant pK\Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\Big)^{1/2}\sqrt{d}C \ \ \mathbb{P}\text{-a.s.}
\end{aligned} \tag{2.19}$$

Using the above, we deduce

$$\begin{aligned}
U_2 &= \int \mathbb{E}[(g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X}))^2]\,\mathrm{d}\varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma}(W,g) \\
&\leqslant d(pKC)^2 \int \cdots \int \sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\,\mathrm{d}\varrho^{1,1}_{d,I,\eta}(w_1)\ldots\mathrm{d}\varrho^{1,d}_{d,I,\eta}(w_d) \\
&\leqslant (dpKC\eta)^2.
\end{aligned} \tag{2.20}$$

By construction, $\varrho^2_{d,M,\gamma}$ is centered around $g^*_{d,M}$ and thus

$$\int g(\mathbf{x})\,\mathrm{d}\varrho^2_{d,M,\gamma}(g) = g^*_{d,M}(\mathbf{x}),\ \forall \mathbf{x} \in \mathbb{R}^p. \tag{2.21}$$

In particular, we have

$$U_3 = 0. \tag{2.22}$$

Using Fubini's theorem together with (2.21), and applying the Cauchy-Schwarz inequality, we obtain

$$|U_4| = 2\Big|\mathbb{E}\Big[(Y - g^*_{d,M}(W^*_{d,I}\mathbf{X}))\int\int g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X})\,\mathrm{d}\varrho^2_{d,M,\gamma}(g)\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big]\Big|$$

$$= 2\Big|\mathbb{E}\Big[(Y - g^*_{d,M}(W^*_{d,I}\mathbf{X})) \int g(W^*_{d,I}\mathbf{X}) - g^*_{d,M}(W\mathbf{X})\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big]\Big|$$
$$\leqslant 2\Big(R(W^*_{d,I}, g^*_{d,M})\mathbb{E}\Big[\Big(\int g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g^*_{d,M}(W\mathbf{X})\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big)^2\Big]\Big)^{1/2}. \tag{2.23}$$

Repeating the argument from treating U_2, but now with $g = g^*_{d,M}$, we have

$$\mathbb{E}\Big[\Big(\int g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g^*_{d,M}(W\mathbf{X})\,\mathrm{d}\varrho^1_{d,I,\eta}(W)\Big)^2\Big] \leqslant (dpKC\eta)^2. \tag{2.24}$$

Clearly, $g \equiv 0 \in \mathcal{G}_{d,M}(C)$ and thus we get by definition of $(W^*_{d,I}, g^*_{d,M})$ that

$$R(W^*_{d,I}, g^*_{d,M}) \leqslant R(W^*_{d,I}, g) = \mathbb{E}[Y^2] = \mathbb{E}[f(\mathbf{X})^2] + 2\mathbb{E}\big[f(\mathbf{X})\mathbb{E}[\varepsilon|\mathbf{X}]\big] + \mathbb{E}[\varepsilon^2] \leqslant C^2 + \sigma^2. \tag{2.25}$$

Plugging (2.24) and (2.25) into (2.23), we have

$$|U_4| \leqslant 2dpKC\eta\sqrt{C^2+\sigma^2}. \tag{2.26}$$

Finally, applying (2.19) again yields

$$|U_5| \leqslant 2\int \mathbb{E}[|g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X})||g(W^*_{d,I}\mathbf{X}) - g(W\mathbf{X})|]\,\mathrm{d}\varrho(W,g)$$
$$\leqslant 2\sqrt{d}pKC\mathbb{E}\Big[\int |g^*_{d,M}(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X})|\Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\Big)^{1/2}\mathrm{d}\varrho((w_1,\dots,w_d)^\top, g)\Big]$$
$$= 2\sqrt{d}pKC\int |g^*_{d,M}(W^*_{d,I}\mathbf{X}(\omega)) - g(W^*_{d,I}\mathbf{X}(\omega))|$$
$$\cdot\Big(\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\Big)^{1/2}\mathrm{d}\mathbb{P}\otimes\varrho(\omega,(w_1,\dots,w_d)^\top,g)$$
$$\leqslant 2\sqrt{d}pKC\Big(\int \big(g^*_{d,M}(W^*_{d,I}\mathbf{X}(\omega)) - g(W^*_{d,I}\mathbf{X}(\omega))\big)^2\mathrm{d}\mathbb{P}\otimes\varrho(\omega,(w_1,\dots,w_d)^\top,g)\Big)^{1/2}$$
$$\cdot\Big(\int\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\,\mathrm{d}\mathbb{P}\otimes\varrho(\omega,(w_1,\dots,w_d)^\top,g)\Big)^{1/2} \tag{2.27}$$
$$\leqslant 2\sqrt{d}pKC\Big(\int \mathbb{E}[(g(W^*_{d,I}\mathbf{X}) - g(W^*_{d,I}\mathbf{X}))^2]\,\mathrm{d}\varrho^2_{d,M,\gamma}(g)\Big)^{1/2}$$
$$\cdot\Big(\int\sum_{i=1}^d |w^*_{d,I,i} - w_i|^2\,\mathrm{d}\varrho^1_{d,I,\eta}((w_1,\dots,w_d)^\top)\Big)^{1/2}$$
$$\leqslant 2dpKC\eta\gamma, \tag{2.28}$$

where (2.27) follows from the Cauchy-Schwarz inequality for integration with respect to the product measure $\mathbb{P}\otimes\varrho^1_{d,I,\eta}\otimes\varrho^2_{d,M,\gamma}$.

Choosing $\eta = (2dpKC\sqrt{C^2+\sigma^2}n)^{-1}, \gamma = n^{-1}$ when summarizing (2.17), (2.18), (2.20), (2.22), (2.26) and (2.28), we have

$$\begin{aligned}\int \mathcal{E}(W,g)\,\mathrm{d}\varrho(W,g) &\leqslant \mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \gamma^2 + (dpKC\eta)^2 + 2dpKC\eta\sqrt{C^2+\sigma^2} + 2dpKC\eta\gamma \\ &\leqslant \mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \frac{4}{n}.\end{aligned}$$

With these choices for η and γ, we get for sufficiently large n that

$$\begin{aligned}T_1 + T_2 &= \|I\|\log(\mathrm{e}p) + (\|I\| + M + 2)\log(2) + \|I\|\log(5dpn) + 16^d N^d 2^{dM}\log((C+1)n) \\ &\leqslant Q_4\big(\|I\|\log(p\vee n) + 16^d N^d 2^{dM}\log(n)\big)\end{aligned}$$

with a constant Q_4 independent of the parameters involved.

Summarizing the above, we arrive at

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant 3\mathcal{E}(W^*_{d,I}, g^*_{d,M}) + \frac{Q_5}{n}\big(\|I\|\log(p\vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big) \tag{2.29}$$

with a probability of at least $1-\delta$, where Q_5 is a constant only depending on C, Γ and σ. Note that the upper bound in (2.29) is deterministic. Choosing a triplet (d, I, M) such that this upper bound is minimized (which is always possible, since there are only finitely many choices for (d, I, M)) completes the proof of Theorem 2.5. □

2.3.4 Proof of Corollary 2.8

Plugging in d^* and I^* in the minimum in Theorem 2.5, we obtain that

$$\begin{aligned}\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) &\leqslant \min_{d,I,M}\Big(3\mathcal{E}(W^*_{d,I}, f^*_{d,M}) + \frac{Q_1}{n}\big(\|I\|\log(p\vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big)\Big) \\ &\leqslant \min_{\substack{0\leqslant M\leqslant n,\\ g\in\mathcal{G}_{d^*,M}(C)}}\Big(3\mathcal{E}(W^*, g) + \frac{Q_1}{n}\big(\|W^*\|_0\log(p\vee n) + 16^d N^d 2^{dM}\log(n) + \log(2/\delta)\big)\Big)\end{aligned} \tag{2.30}$$

with a probability of at least $1-\delta$. The rest of the proof consists of choosing M to balance the terms on the right hand side of (2.30) by using an approximation of g^*, namely

$$g_M = \sum_{l\in\mathcal{Z}_M^{d^*}} \langle g^*, \Psi_l\rangle \Psi_l$$

and then determining the projection level M. To do this, we have to verify that g_M is a valid choice for g in the sense that $g_M \in \mathcal{G}_{d^*,M}(C)$. Indeed, the Cauchy-Schwarz inequality ensures that

$$\begin{aligned} L^{d^*} \sum_{l \in \mathcal{Z}_M^{d^*}} 2^{l_1(d/2+1)} |\langle g^*, \Psi_l \rangle| &\leqslant L^{d^*} \Big(\sum_{l \in \mathcal{Z}_M^{d^*}} 2^{2l_1(1-\alpha+d^*/2)} \Big)^{\frac{1}{2}} \Big(\sum_{l \in \mathcal{Z}_M^{d^*}} 2^{2l_1\alpha} |\langle g^*, \Psi_l \rangle|^2 \Big)^{1/2} \\ &\leqslant L^{d^*} 2 N^{d^*/2} 16^{d^*/2} \Big(\sum_{l \in \mathcal{Z}_{\infty,\infty}^{d^*}} 2^{2l_1\alpha} |\langle g^*, \Psi_l \rangle|^2 \Big)^{1/2} \\ &\leqslant C. \end{aligned}$$

Using Assumption 2.C, we see that g_M admits an excess risk of

$$\begin{aligned} \mathcal{E}(W^*, g_M) &= \mathbb{E}[(g_M(W^*\mathbf{X}) - g^*(W^*\mathbf{X}))^2] \\ &= \int_{[-B_1,B_1]^{d^*}} \varrho(x)(g_M(x) - g^*(x))^2 \mathbb{\lambda}^{d^*}(\mathrm{d}x) \\ &\leqslant B_2 \int_{[-B_1,B_1]^{d^*}} (g_M(x) - g^*(x))^2 \mathbb{\lambda}^{d^*}(\mathrm{d}x) \\ &\leqslant B_2 \sum_{l \in \mathcal{Z}_{\infty,N}^{d^*} \setminus \mathcal{Z}_M^{d^*}} |\langle g^*, \Psi_l \rangle|^2 \\ &\leqslant B_2 2^{-2\alpha M} \sum_{l \in \mathcal{Z}_{\infty,N}^{d^*} \setminus \mathcal{Z}_M^{d^*}} 2^{2l_1\alpha} |\langle g^*, \Psi_l \rangle|^2 \\ &\leqslant B_2 2^{-2\alpha M} (2 N^{d^*/2} 16^{d^*/2})^{-1} C L^{-d^*}. \end{aligned} \tag{2.31}$$

Applying (2.31) to (2.30), we see for sufficiently large n and with a probability of at least $1 - \delta$ that

$$\mathcal{E}(\widehat{W}_\lambda, \widehat{g}_\lambda) \leqslant Q_6 \min_{0 \leqslant M \leqslant n} \Big(2^{-2\alpha M} + 2^{d^* M} \frac{\log n}{n} + \frac{\|W^*\|_0 \log(p \vee n)}{n} + \frac{\log(2/\delta)}{n} \Big)$$

with a constant Q_6 only depending on $C, \Gamma, \sigma, N, B_1, B_2$ and d^*. To balance the order of the terms depending on M, we choose

$$M = \Big\lceil \log\Big(\frac{n}{\log n}\Big) \Big/ \big((2\alpha + d^*)\log(2)\big) \Big\rceil,$$

which completes the proof. □

2.3.5 Proofs of auxiliary lemmas

2.3.5.1 Proof of Lemma 2.10

We employ another auxiliary lemma:

Lemma 2.12. *It holds that*

$$\begin{aligned}\mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \Pi) =& \log(G(d,I,M)) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M}), \qquad \textit{where}\\ G(d,I,M) :=& (1-2^{-p})(1-2^{(1-p)d-1})(2-2^{-n})2^{\|I\|+M+1}|\mathcal{I}_{d,\|I\|}|.\end{aligned}$$

Now, we can combine $|\mathcal{I}_{d,\|I\|}| \leqslant \binom{dp}{\|I\|}$ with the basic inequality $\binom{dp}{\|I\|} \leqslant \left(\frac{dp\mathrm{e}}{\|I\|}\right)^{\|I\|}$ and $\|I\| \geqslant d$ to obtain

$$\begin{aligned}\mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \Pi) &= \log(G(d,I,M)) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M})\\ &\leqslant \log\left(2^{\|I\|+M+2}\tbinom{dp}{\|I\|}\right) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M})\\ &\leqslant \|I\|\log(\mathrm{e}p) + (\|I\|+M+2)\log(2) + \mathrm{KL}(\varrho_{d,I,M,\eta,\gamma} \mid \pi_{d,I,M}). \qquad \square\end{aligned}$$

2.3.5.2 Proof of Lemma 2.11

We split the proof into two further auxiliary lemmas:

Lemma 2.13. *For $\varrho^1_{d,I,\eta}$ from (2.15), we have*

$$\mathrm{KL}(\varrho^1_{d,I,\eta} \mid \mu_{d,I}) \leqslant \|I\|\log(5/\eta).$$

Lemma 2.14. *For $\varrho^2_{d,M,\gamma}$ from (2.15), we have*

$$\mathrm{KL}(\varrho^2_{d,M,\gamma} \mid \nu_{d,M}) \leqslant 16^d N^d 2^{dM}\log((C+1)/\gamma).$$

The assertion then follows directly via

$$\begin{aligned}\mathrm{KL}(\varrho \mid \pi_{d,I,M}) &= \mathrm{KL}(\varrho^1_{d,I,\eta} \otimes \varrho^2_{d,M,\gamma} \mid \mu_{d,I} \otimes \nu_{d,M})\\ &= \mathrm{KL}(\varrho^1_{d,I,\eta} \mid \mu_{d,I}) + \mathrm{KL}(\varrho^2_{d,M,\gamma} \mid \nu_{d,M})\\ &\leqslant \|I\|\log(5/\eta) + 16^d N^d 2^{dM}\log((C+1)/\gamma). \qquad \square\end{aligned}$$

2.3.5.3 Proof of Lemma 2.12

To simplify the notation we write $\varrho = \varrho_{d,I,M,\eta,\gamma}$ and $\pi_{d,I,M} = \mu_{d,I} \otimes \nu_{d,M}$. We will show that

$$\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi} = G(d,I,M)\frac{\mathrm{d}\varrho}{\mathrm{d}\pi_{d,I,M}}, \tag{2.32}$$

from which we can deduce

$$\begin{aligned}\mathrm{KL}(\varrho \mid \Pi) = \int \log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\Big)\,\mathrm{d}\varrho &= \log(G(d,I,M)) + \int \log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\pi_{d,I,M}}\Big)\,\mathrm{d}\varrho \\ &= \log(G(d,I,M)) + \mathrm{KL}(\varrho \mid \pi_{d,I,M}).\end{aligned}$$

For (2.32), we need to show that

$$\varrho(A) = \int_A G(d,I,M)^{-1}\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\,\mathrm{d}\pi_{d,I,M} \tag{2.33}$$

holds for all Borel-measurable sets $A = A_1 \times A_2 \subseteq \cup_{d=1}^{p} \mathcal{S}_d \times \mathcal{G}_d$. Observe that for

$$\begin{aligned}\mathcal{S}_{d,\Leftrightarrow}(J) &\coloneqq \{W = (w_1,\dots,w_d)^\top \in \mathcal{S}_d \mid (w_{i,j} \neq 0 \Leftrightarrow j \in J_i)\,\forall i \in \{1,\dots,d\}, j \in \{1,\dots,p\}\} \\ \mathcal{G}_{d,\widetilde{M},\neq}(C+1) &\coloneqq \big\{g = \Phi_{d,\widetilde{M}}(\beta) \in \mathcal{G}_{d,\widetilde{M}}(C+1) \,\big|\, \exists l \in \mathcal{Z}_{\widetilde{M}}^{d} : |l_2|_\infty = 2^{l_1}\widetilde{M},\, \beta_l \neq 0\big\}\end{aligned}$$

with $J = J_1 \times \cdots \times J_d \in \mathcal{I}_d$ and $\widetilde{M} \in \{0,\dots,n\}$, we have

$$\mu_{d,J}(\mathcal{S}_{d,\Leftrightarrow}(J)) = \nu_{d,\widetilde{M}}(\mathcal{G}_{d,\widetilde{M},\neq}(C+1)) = 1. \tag{2.34}$$

In particular, (2.34) holds for $J = I$ and $\widetilde{M} = M$. Since also $\varrho(\mathcal{S}_{d,\Leftrightarrow}(I) \times \mathcal{G}_{d,M,\neq}(C+1)) = 1$, no generality is lost in additionally assuming that

$$A_1 \subseteq \mathcal{S}_{d,\Leftrightarrow}(I) \qquad \text{and} \qquad A_2 \subseteq \mathcal{G}_{d,M,\neq}(C+1).$$

Note that

$$\mathcal{S}_{d,\Leftrightarrow}(J) \cap \mathcal{S}_{d,\Leftrightarrow}(I) = \emptyset\,\forall J \neq I \quad \text{and} \quad \mathcal{G}_{d,\widetilde{M},\neq}(C+1) \cap \mathcal{G}_{d,M,\neq}(C+1) = \emptyset\,\forall \widetilde{M} \neq M. \tag{2.35}$$

Combining (2.34) with (2.35), we find

$$\int_{A_1} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W,g)\,\mathrm{d}\mu_{d,J}(W) = 0$$

for any $g \in A_2$ and $J \neq I$. Similarly, we have for any $W \in A_1$ and $\widetilde{M} \neq M$ that

$$\int_{A_2} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W, g)\, \mathrm{d}\nu_{d,\widetilde{M}}(g) = 0.$$

Therefore, repeated application of Fubini's theorem yields

$$\begin{aligned}
\varrho(A) = \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\, \mathrm{d}\Pi &= \sum_{c=1}^{p} 2^{-c} \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\, \mathrm{d}\pi_c \Big/ (1 - 2^{-p}) \\
&= \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\, \mathrm{d}\mu_d \otimes \nu_d \Big/ \big(2^d(1 - 2^{-p})\big) \\
&= \int_{A_1 \times A_2} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W, g)\, \mathrm{d}\mu_d \otimes \nu_d(W, g) \Big/ \big(2^d(1 - 2^{-p})\big) \\
&= G(d, I, M)^{-1} \int_{A_1 \times A_2} \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(W, g)\, \mathrm{d}\mu_{d,I} \otimes \nu_{d,M}(W, g) \\
&= G(d, I, M)^{-1} \int_A \frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}\, \mathrm{d}\pi_{d,I,M}.
\end{aligned}$$

Thus, we have shown (2.33). □

2.3.5.4 Proof of Lemma 2.13

We will show that

$$\mathrm{KL}(\varrho_{d,I,\eta}^{1,i} \mid \mu_{I_i}) \leqslant |I_i| \log(5/\eta), \qquad \forall i = 1, \dots, d. \tag{2.36}$$

The assertion follows immediately via

$$\mathrm{KL}(\varrho \mid \mu_{d,I}) = \mathrm{KL}\Big(\bigotimes_{i=1}^{d} \varrho_{d,I,\eta}^{1,i} \,\Big|\, \bigotimes_{i=1}^{d} \mu_{I_i}\Big) = \sum_{i=1}^{d} \mathrm{KL}(\varrho_{d,I,\eta}^{1,i} \mid \mu_{I_i}) \leqslant \|I\| \log(5/\eta),$$

where the first equality holds barring a slight breach of conventions for product measures. To show (2.36), we fix $i \in \{1, \dots, d\}$ and for simplicity of the notation, we set $\varrho \coloneqq \varrho_{d,I,\eta}^{1,i}$ and $J = I_i$. Plugging the μ_J-density of ϱ into the definition of the Kullback-Leibler divergence, we easily obtain

$$\mathrm{KL}(\varrho \mid \mu_J) = -\log\Big(\int \mathbb{1}_{\{|w - w^*_{d,I,i}| \leqslant \eta\}} \mu_J(\mathrm{d}w)\Big) = -\log(\widetilde{\mu}(\{w \in \mathbb{R}^{|J|} \mid |w - \widetilde{w}^*| \leqslant \eta\})), \tag{2.37}$$

where $\widetilde{w}^*$ is the projection of $w^*_{d,I,i}$ onto the coordinates whose indices are elements of J and $\widetilde{\mu}_J$ denotes the uniform distribution on the unit sphere in $\mathbb{R}^{|J|}$.

We want to show a lower bound for $\widetilde{\mu}(\{w \in \mathbb{R}^{|J|} \mid |w - \widetilde{w}^*| \leqslant \eta\})$, which is the proportion of the

surface of the unit sphere in $\mathbb{R}^{|J|}$ covered by the η-ball around $\widetilde{w}^*$ to the surface of the entire unit sphere. By rotational symmetry of the uniform distribution on this sphere, no generality is lost by assuming $\widetilde{w}^* = (1, 0, \dots, 0)^\top \in \mathbb{R}^{|J|}$.

For any $\widetilde{w} = (\widetilde{w}_1, \dots, w_{|J|})^\top \neq 0$ with $|\widetilde{w}| \leqslant 1$, the η-ball around $\widetilde{w}^*$ covers the same part of the surface of the unit sphere as the $\widetilde{\eta}$-ball around $\widetilde{w}$, where $\widetilde{\eta} = \sqrt{\eta^2|\widetilde{w}| - 2|\widetilde{w}| + |\widetilde{w}|^2 + 1}$. Using this dependence between $\widetilde{w}, \eta$ and $\widetilde{\eta}$, it is easily checked that $|\widetilde{w}| = \sqrt{1 - \widetilde{\eta}^2}$ is solved for $0 < \widetilde{\eta} = \sqrt{\eta^2 - \eta^4/4} < 1$. Henceforth, we fix this $\widetilde{\eta}$. Suppose that $N_{\widetilde{\eta}}$ is the smallest number of $\widetilde{\eta}$-balls with centers in the unit ball that suffice to cover the entire unit ball. Denote their centers by $t_1, \dots, t_{N_{\widetilde{\eta}}}$. In particular, these balls cover the entire unit sphere and therefore, at least one of them covers at least $N_{\widetilde{\eta}}^{-1}$ of the surface of the unit sphere. This can only be the case for $t_j \neq 0$, because otherwise $\widetilde{\eta} < 1$ implies $\{w \in \mathbb{R}^{|J|} \mid |w - t_j| \leqslant \widetilde{\eta}\} \cap \{w \in \mathbb{R}^{|J|} \mid |w| = 1\} = \emptyset$. If we now change the length of such a $w := t_j \neq 0$ (without changing its orientation and without changing $\widetilde{\eta}$), the coverage of the unit sphere provided by the corresponding $\widetilde{\eta}$-ball also changes. In particular, we will show that if $|w| \neq \sqrt{1 - \widetilde{\eta}^2}$, then decreasing (or increasing) $|w|$ towards $\sqrt{1 - \widetilde{\eta}^2}$, enlarges the coverage of the corresponding ball on the unit sphere. Thus, the proportional coverage of the $\widetilde{\eta}$-ball around $|w|^{-1}\sqrt{1 - \widetilde{\eta}^2}w$ (which has, as we showed above, the same proportional coverage of the unit sphere as the η-ball around $\widetilde{w}^*$ that we are actually trying to control) is bounded from below by the proportional coverage of the $\widetilde{\eta}$-ball around $\widetilde{w}$, which in turn is bounded from below by $N_{\widetilde{\eta}}^{-1}$. Using the fact that $N_{\widetilde{\eta}} \leqslant (3/\widetilde{\eta})^{|J|}$ combined with $\widetilde{\eta} \geqslant \eta/\sqrt{2}$, we have

$$N_{\widetilde{\eta}}^{-1} \geqslant (3/\widetilde{\eta})^{-|J|} \geqslant (3\sqrt{2}/\eta)^{-|J|} \geqslant (5/\eta)^{-|J|}$$

and therefore (2.36) follows from (2.37) with

$$\mathrm{KL}(\varrho \mid \mu_J) = -\log(\widetilde{\mu}(\{w \in \mathbb{R}^{|J|} \mid |w - \widetilde{w}^*| \leqslant \eta\})) \leqslant -\log(N_{\widetilde{\eta}}^{-1}) \leqslant |J| \log(5/\eta).$$

It remains to show that changing the length of $w \neq 0$ towards $\sqrt{1 - \widetilde{\eta}^2}$ increases the proportional coverage of the corresponding $\widetilde{\eta}$-ball. By rotational symmetry, we can assume $w = (w_1, 0, \dots, 0)^\top \in \mathbb{R}^{|J|}$ with some $0 < w_1 \leqslant 1$ at no loss of generality. Now, it is sufficient to show that

$$\{y \in \mathbb{R}^{|J|} \mid |y| = 1,\, |y - w| \leqslant \widetilde{\eta}\} \subseteq \{y \in \mathbb{R}^{|J|} \mid |y| = 1,\, |y - \widetilde{w}| \leqslant \widetilde{\eta}\},$$

where $\widetilde{w} = (\sqrt{1 - \widetilde{\eta}^2}, 0, \dots, 0)^\top \in \mathbb{R}^{|J|}$. In this setting, and as $\eta, \widetilde{\eta} > 0$, the relationship

$$\widetilde{\eta}^2 = \eta^2|\widetilde{w}| - 2|\widetilde{w}| + |\widetilde{w}|^2 + 1$$

is equivalent to

$$\eta^2 = (2\widetilde{w}_1 - \widetilde{w}_1^2 - 1 + \widetilde{\eta}^2)/\widetilde{w}_1.$$

Using elementary calculus together with the fact that $\widetilde{\eta} < 1$, it is straightforward to see

$$(2w_1 - w_1^2 - 1 + \widetilde{\eta}^2)/w_1 \leqslant 2(1 - \sqrt{1 - \widetilde{\eta}^2}).$$

In combination with the relationship between η and $\widetilde{\eta}$, we obtain

$$\begin{aligned}
\{y \in \mathbb{R}^{|J|} \mid |y - w| \leqslant \widetilde{\eta},\, |y| = 1\} &= \{y \in \mathbb{R}^{|J|} \mid |y| = 1,\, |y - \widetilde{w}^*|^2 \leqslant (2w_1 - w_1^2 - 1 + \widetilde{\eta}^2)/w_1\} \\
&\subseteq \{y \in \mathbb{R}^{|J|} \mid |y| = 1,\, |y - \widetilde{w}^*|^2 \leqslant 2(1 - \sqrt{1 - \widetilde{\eta}^2})\} \\
&= \{y \in \mathbb{R}^{|J|} \mid |y| = 1,\, |y - \widetilde{w}^*|^2 \leqslant \eta^2\} \\
&= \{y \in \mathbb{R}^{|J|} \mid |y| = 1,\, |y - \widetilde{w}| \leqslant \widetilde{\eta}\}. \qquad \square
\end{aligned}$$

2.3.5.5 Proof of Lemma 2.14

To simplify the notation, we write $\widetilde{g}^* = g^*_{d,M}$ and $\widetilde{\beta}^* = \beta^*_{d,M}$. We will show that

$$\int \mathbb{1}_{\{\|g - \widetilde{g}^*\|_{\mathcal{B}} \leqslant \gamma\}} \mathrm{d}\nu_{d,M}(g) = ((C+1)/\gamma)^{-|\mathcal{Z}^d_M|}. \tag{2.38}$$

The assertion follows directly with

$$\begin{aligned}
\mathrm{KL}(\varrho^2_{d,M,\gamma} \mid \nu_{d,M}) = -\log\Big(\int \mathbb{1}_{\{\|g - g^*\|_{\mathcal{B}} \leqslant \gamma\}} \mathrm{d}\nu_{d,M}(g)\Big) &= |\mathcal{Z}^d_M| \log((C+1)/\gamma) \\
&\leqslant 16^d N^d 2^{dM} \log((C+1)/\gamma).
\end{aligned}$$

We now verify (2.38) using the definition of $\nu_{d,M}$. If we let $\lambda^{\mathcal{Z}^d_M}$ denote the Lebesgue measure on $\mathbb{R}^{\mathcal{Z}^d_M}$, it holds that

$$\begin{aligned}
\int \mathbb{1}_{\{\|g - \widetilde{g}^*\|_{\mathcal{B}} \leqslant \gamma\}} \,\mathrm{d}\nu_{d,M}(g) &= \int \mathbb{1}_{\{\|\Phi_{d,M}(\beta) - \widetilde{g}^*\|_{\mathcal{B}} \leqslant \gamma\}} \,\mathrm{d}\widetilde{\nu}_{d,M}(\beta) \\
&= \int \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant \gamma\}} \,\mathrm{d}\widetilde{\nu}_{d,M}(\beta) \\
&= \frac{\int \mathbb{1}_{\{\|\beta\|_{\mathcal{B}} \leqslant C+1\}} \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant \gamma\}} \,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}{\int \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant C+1\}} \,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)} \\
&= \frac{\int \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant \gamma\}} \,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}{\int \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant C+1\}} \,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)} \\
&= \Big(\frac{\gamma}{C+1}\Big)^{|\mathcal{Z}^d_M|} \frac{\int \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant 1\}} \,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)}{\int \mathbb{1}_{\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant 1\}} \,\mathrm{d}\lambda^{\mathcal{Z}^d_M}(\beta)} \\
&= \Big(\frac{C+1}{\gamma}\Big)^{-|\mathcal{Z}^d_M|},
\end{aligned}$$

where we have used that

$$\|\beta\|_{\mathcal{B}} \leqslant \|\widetilde{\beta}^*\|_{\mathcal{B}} + \|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant C + \gamma \leqslant C + 1$$

on $\{\|\beta - \widetilde{\beta}^*\|_{\mathcal{B}} \leqslant \gamma\}$ in the fourth equality. This implies that if the second indicator in the integral in the numerator is 1, so is the first. □

3 Statistical guarantees for stochastic Metropolis-Hastings

In the previous chapter we demonstrated statistical properties of an estimator based on the Gibbs posterior in multi-index models. In this chapter we extend the methodology to the more flexible class of hierarchical functions by harnessing the approximation properties of neural networks.

The large number of parameters involved in the training of such networks necessitates a new focus on algorithmic aspects. In particular, we develop an algorithm which allows us to access the Gibbs posterior in a way that scales with the sample size. In Section 3.1, we use a fairly general setting to explain this algorithm, which is then applied to neural networks in Section 3.2. We illustrate our method with a numerical example in Section 3.3. The results are based on Bieringer et al. (2023).

For a streamlined presentation of the results and proofs, we *reset* the constants $Q_0, Q_1, \dots$ from the previous chapter.

3.1 Stochastic Metropolis-adjusted Langevin algorithm

Throughout, we consider the regression setting from the introduction of Chapter 2. As the parameter set, we fix $\Theta = [B, B]^P$ for some $B \geqslant 1$ and a potentially large number of parameters $P \in \mathbb{N}$. We choose a uniform prior $\Pi = \mathcal{U}(\Theta)$. In particular the prior and the posterior distribution have Lebesgue densities which we denote by the same symbols as the distributions themselves. For a more concise presentation, we postpone the specification of the class $\mathcal{F} = \{f_\vartheta \mid \vartheta \in \Theta\}$ to Section 3.2.

3.1.1 Metropolis-adjusted Langevin algorithm

To apply the estimators $\widehat{f}_\lambda$ and $\bar{f}_\lambda$ from (2.3) and (2.4) in practice, we need to sample from the Gibbs posterior

$$\Pi_\lambda(\vartheta \mid \mathcal{D}_n) \propto \exp\big(-\lambda R_n(\vartheta)\big)\Pi(\vartheta)$$

from (2.2). The MCMC approach is to construct a Markov chain $(\vartheta^{(k)})_{k\in\mathbb{N}_0}$ with stationary distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$, see Robert & Casella (2004). In particular, the *Langevin* MCMC sampler is given by

$$\vartheta^{(k+1)} = \vartheta^{(k)} - \gamma\nabla_\vartheta R_n(\vartheta^{(k)}) + sW_k, \tag{3.1}$$

where $\nabla_\vartheta R_n(\vartheta)$ denotes the gradient of $R_n(\vartheta)$ with respect to ϑ. Moreover, $\gamma > 0$ is the learning rate and $sW_k \sim \mathcal{N}(0, s^2 E_P)$ is i.i.d. white noise with noise level $s > 0$. This approach can also be interpreted as a noisy version of the gradient descent method commonly used to train neural networks. In practice this approach requires careful tuning of the procedural parameters and Langevin-MCMC suffers from relatively slow polynomial convergence rates of the distribution of $\vartheta^{(k)}$ to the target distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$, see Nickl & Wang (2022); Cheng & Bartlett (2018). Only in special cases, the convergence rates are faster, see e.g. Freund et al. (2022) for an overview and Dalalyan & Riou-Durand (2020) for the case of log-concave densities. This convergence rate can be considerably improved by adding an MH step resulting in the *Metropolis-adjusted Langevin algorithm* (MALA), see Roberts & Tweedie (1996a).

Applying the generic MH algorithm to $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ and taking into account that the prior Π is uniform, we obtain the following iterative method: Starting with some initial choice $\vartheta^{(0)} \in \mathbb{R}^P$, we successively generate $\vartheta^{(k+1)}$ given $\vartheta^{(k)}$, $k \in \mathbb{N}_0$, by

$$\vartheta^{(k+1)} = \begin{cases} \vartheta' & \text{with probability } \alpha(\vartheta' \mid \vartheta^{(k)}) \\ \vartheta^{(k)} & \text{with probability } 1 - \alpha(\vartheta' \mid \vartheta^{(k)}) \end{cases},$$

where ϑ' is a random variable drawn from some conditional proposal density $q(\cdot \mid \vartheta^{(k)})$ and the *acceptance probability* is chosen as

$$\alpha(\vartheta' \mid \vartheta) = \exp\big(-\lambda R_n(\vartheta') + \lambda R_n(\vartheta)\big)\mathbb{1}_{[-B,B]^P}(\vartheta')\frac{q(\vartheta \mid \vartheta')}{q(\vartheta' \mid \vartheta)} \wedge 1. \tag{3.2}$$

In view of (3.1) the probability density q of the proposal distribution is given by

$$q(\vartheta' \mid \vartheta) = \frac{1}{(2\pi s^2)^{P/2}} \exp\Big(-\frac{1}{2s^2}\big|\vartheta' - \vartheta + \gamma\nabla_\vartheta R_n(\vartheta)\big|^2\Big). \tag{3.3}$$

The standard deviation s should not be too large as otherwise the acceptance probability might

be too small. As a result the proposal would rarely be accepted, the chain might not be sufficiently randomized and the convergence to the invariant target distribution would be too slow in practice. On the other hand, s should not be smaller than the shift $\gamma \nabla_\vartheta R_n(\vartheta)$ in the mean, since otherwise $q(\vartheta \mid \vartheta')$ might be too small. The MH step ensures that $(\vartheta^{(k)})_{k \in \mathbb{N}_0}$ is a Markov chain with invariant distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ (under rather mild conditions on q). The convergence to the invariant distribution follows from Roberts & Tweedie (1996b, Theorem 2.2) with geometric rate.

To calculate the estimators $\widehat{f}_\lambda$ and $\bar{f}_\lambda$, one chooses a *burn-in* time $b \in \mathbb{N}$ to let the distribution of the Markov chain stabilize at its invariant distribution and then sets

$$\widehat{f}_\lambda = f_{\vartheta^{(b)}} \qquad \text{and} \qquad \bar{f}_\lambda = \frac{1}{N} \sum_{k=1}^{N} f_{\vartheta^{(b+ck)}}.$$

A sufficiently large *gap length* $c \in \mathbb{N}$ ensures the necessary variability and reduced dependence between $\vartheta^{(b+ck)}$ and $\vartheta^{(b+c(k+1))}$, whereas $N \in \mathbb{N}$ has to be large enough for a good approximation of the expectation by the empirical mean.

3.1.2 Stochastic MALA

The gradient has to be calculated only once in each MALA iteration. Hence, using the full gradient $\nabla_\vartheta R_n(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\vartheta \ell_i(\vartheta)$, the additional computational price of MALA compared to training a standard neural network by empirical risk minimization only comes from a larger number of necessary iterations due to the rejection with probability $1 - \alpha(\vartheta' \mid \vartheta^{(k)})$. For large data sets however the standard training of a neural network would rely on a stochastic gradient method, where the gradient $\frac{1}{m} \sum_{i \in \mathcal{B}} \nabla_\vartheta \ell_i(\vartheta)$ is only calculated on (mini-)batches $\mathcal{B} \subset \{1, \dots, n\}$ of size $m < n$. While we could replace $\nabla_\vartheta R_n(\vartheta)$ in (3.3) by a stochastic approximation without any additional obstacle, the MH step still requires the calculation of the loss $\ell_i(\vartheta')$ for all $1 \leqslant i \leqslant n$ in (3.2).

To avoid a full evaluation of the empirical risk $R_n(\vartheta)$, a natural approach is to replace the empirical risks in $\alpha(\vartheta' \mid \vartheta)$ by a batch-wise approximation, too. To study the consequences of this approximation we follow a pseudo-marginal MH approach, see Andrieu & Roberts (2009); Maclaurin & Adams (2014); Bardenet et al. (2017); Wu et al. (2022).

We augment our target distribution by a set of auxiliary random variables $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(\rho)$ with some $\rho \in (0, 1]$ and aim for a reduction of the empirical risk $R_n(\vartheta)$ to the stochastic

approximation

$$R_n(\vartheta, Z) \coloneqq \frac{1}{n\rho} \sum_{i=1}^{n} Z_i \ell_i(\vartheta)$$

in the algorithm. Hence, we define the joint target distribution by

$$\begin{aligned}\bar{\Pi}_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n) &\propto \prod_{i=1}^{n} \rho^{z_i}(1-\rho)^{1-z_i} \exp\big(-\lambda R_n(\vartheta, z)\big)\Pi(\vartheta) \\ &\propto \exp\Big(-\lambda R_n(\vartheta, z) + \log\Big(\frac{\rho}{1-\rho}\Big)\sum_{i=1}^{n} z_i\Big)\Pi(\vartheta), \qquad z \in \{0,1\}^n.\end{aligned}$$

The marginal distribution in ϑ is then given by

$$\bar{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) = \sum_{z\in\{0,1\}^n} \bar{\Pi}_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n) \propto \prod_{i=1}^{n} \Big(\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\Big)\Pi(\vartheta). \tag{3.4}$$

As proposal for the MH algorithm we use

$$\begin{aligned}\bar{q}(\vartheta', z' \mid \vartheta, z) &= q_{\mathrm{s}}(\vartheta' \mid \vartheta, z) \prod_{i=1}^{n} \rho^{z_i'}(1-\rho)^{1-z_i'} \qquad \text{with} \\ q_{\mathrm{s}}(\vartheta' \mid \vartheta, z) &= \frac{1}{(2\pi s^2)^{P/2}} \exp\Big(-\frac{1}{2s^2}\big|\vartheta' - \vartheta + \gamma\nabla_\vartheta R_n(\vartheta, z)\big|^2\Big).\end{aligned} \tag{3.5}$$

Hence, the proposed $Z' = z'$ is indeed a vector of independent $\mathrm{Ber}(\rho)$-random variables and $q_{\mathrm{s}}(\vartheta' \mid \vartheta, z)$ is the stochastic analogue to q from (3.3) with a stochastic gradient. The resulting acceptance probabilities are given by

$$\begin{aligned}\alpha(\vartheta', z' \mid \vartheta, z) &= \frac{\bar{q}(\vartheta, z \mid \vartheta', z')\bar{\Pi}_{\lambda,\rho}(\vartheta', z' \mid \mathcal{D}_n)}{\bar{q}(\vartheta', z' \mid \vartheta, z)\bar{\Pi}_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n)} \wedge 1 \\ &= \frac{q_{\mathrm{s}}(\vartheta \mid \vartheta', z')}{q_{\mathrm{s}}(\vartheta' \mid \vartheta, z)} \mathbb{1}_{[-B,B]^P}(\vartheta') \mathrm{e}^{-\lambda R_n(\vartheta', z') + \lambda R_n(\vartheta, z)} \wedge 1.\end{aligned}$$

We observe that $\alpha(\vartheta', z' \mid \vartheta, z)$ corresponds to a stochastic MH step where we have to evaluate the loss $\ell_i(\vartheta')$ for the new proposal ϑ' only if $z_i' = Z_i' \sim \mathrm{Ber}(\rho)$ is one, i.e. with probability ρ. Calculating $\alpha(\vartheta', z' \mid \vartheta, z)$ thus requires only few evaluations of $\ell_i(\vartheta)$ for small values of ρ. The expected number of data points on which the gradient and the loss have to be evaluated is $n\rho$ and corresponds to a batch size of $m = n\rho$.

Generalizing (2.3), we define the stochastic MH estimator

$$\widehat{f}_{\lambda,\rho} \coloneqq f_{\widehat{\vartheta}_{\lambda,\rho}} \qquad \text{for} \qquad \widehat{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n \sim \bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n). \tag{3.6}$$

For $\rho = 1$ we recover the standard MALA.

As discussed by Bardenet et al. (2017), the previous derivation reveals that the stochastic MH step leads to a different invariant distribution of the Markov chain, namely (3.4) instead of the Gibbs posterior from (2.2). Writing

$$\bar{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) \propto \exp\big(-\lambda \bar{R}_{n,\rho}(\vartheta)\big)\Pi(\mathrm{d}\vartheta) \quad \text{with} \quad \bar{R}_{n,\rho}(\vartheta) \coloneqq -\frac{1}{\lambda}\sum_{i=1}^{n} \log\big(\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\big), \tag{3.7}$$

we observe that $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ is itself a Gibbs posterior distribution, the *surrogate posterior*, corresponding to the modified risk $\bar{R}_{n,\rho}(\vartheta)$. Note that $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ coincides with $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ for $\rho = 1$ and thus $\widehat{f}_\lambda = \widehat{f}_{\lambda,1}$ and $\bar{f}_\lambda = \bar{f}_{\lambda,1}$ in distribution. Whether $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ also behaves as our original target distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ for $\rho < 1$ depends on the choice of λ and ρ:

Lemma 3.1. *If f and all f_ϑ are bounded by some constant $C > 0$, then we have*

$$\frac{1}{n\rho}\,\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \Pi_\lambda(\cdot \mid \mathcal{D}_n)\right) \leqslant \left(\frac{\lambda}{n\rho}\right)^2 \left(64C^4 + \frac{4}{n}\sum_{i=1}^{n} \varepsilon_i^4\right).$$

For $\rho < 1$ and the probability distribution $\varpi_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) :\propto \exp\big(\rho \sum_{i=1}^n \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}\big)\Pi(\mathrm{d}\vartheta)$ we moreover have

$$\frac{1}{n\rho}\,\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)\right) \leqslant \frac{\rho}{1-\rho}.$$

On the one hand, if $\frac{\lambda}{n\rho}$ is sufficiently small, then the surrogate posterior $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ is indeed a good approximation for the Gibbs posterior $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$. On the other hand, for $\rho \to 0$ the distribution $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ behaves as the distribution $\varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ with density proportional to $\exp\big(\rho \sum_{i=1}^n \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}\big)\Pi(\mathrm{d}\vartheta)$. For large $\frac{\lambda}{n\rho}$ the terms $\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}$ rapidly decay for all ϑ with $\ell_i(\vartheta) > 0$, i.e. $\varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ emphasizes interpolating parameter choices. For all ϑ where $\frac{\lambda}{n\rho}\ell_i(\vartheta)$ is relatively large the density converges to a constant. Therefore, in the extreme case $\rho \to 0$ and $\frac{\lambda}{n\rho} \to \infty$ the distribution $\varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ and thus $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ converge to the uninformative prior with interpolating spikes at parameters where $\ell_i(\vartheta)$ are zero.

We illustrate Lemma 3.1 in a simple setting where $Y_i = \mathcal{N}(0, 0.5)$ and $f_\vartheta(\mathbf{x}) \equiv \vartheta$ for $\vartheta \in [-1, 1]$. The densities of the measures $\Pi(\cdot \mid \mathcal{D}_n)$, $\bar{\Pi}(\cdot \mid \mathcal{D}_n)$ and $\varpi(\cdot \mid \mathcal{D}_n)$ are shown in Fig. 3.1 for different choices of λ and ρ. Fig. 3.1 confirms the predicted approximation properties: $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ behaves similarly to $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ if λ is not too large (orange lines) or ρ is not too small (left figure). Additionally, we observe that $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ is still informative if λ is in the order $n\rho$ even if it is not close to the Gibbs posterior at all.

The scaling of the Kullback-Leibler distance with $n\rho$ in Lemma 3.1 is quite natural in this setting.

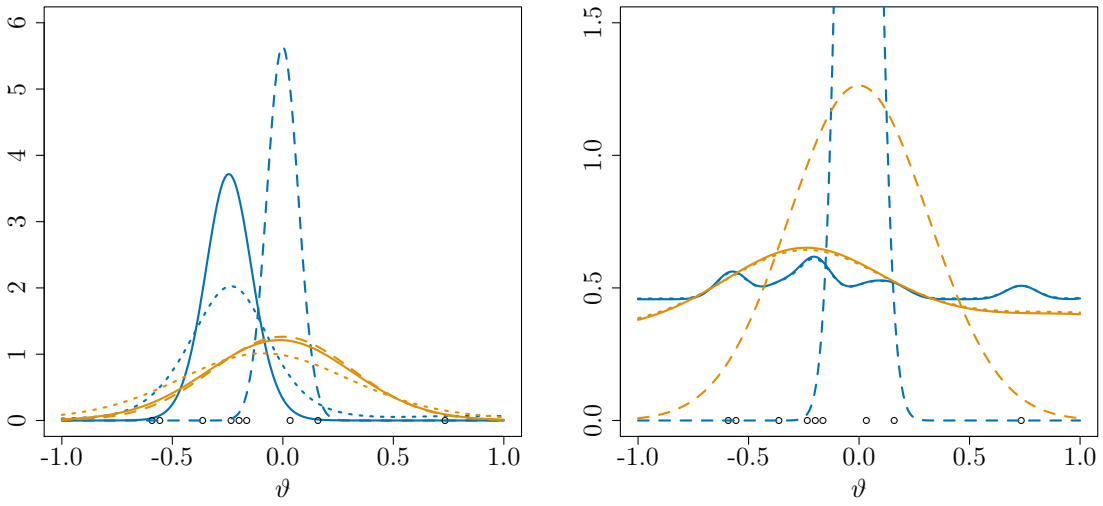

Figure 3.1: *Points*: $Y_1, \dots, Y_n \sim \mathcal{N}(0, 0.5)$ for $n = 10$. *Solid lines*: densities of $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ with $\lambda = 10n$ (blue) and $\lambda = n/2$ (orange) and $\rho = 0.9$ (left) and $\rho = 0.1$ (right). *Dashed lines*: corresponding densities of $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$. Dotted lines: corresponding densities of $\varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$.

In particular, applying an approximation result from the variational Bayes literature by Ray & Szabó (2022, Theorem 5) we obtain for the two reference measures $\mathbb{Q} \in \{\Pi_\lambda(\cdot \mid \mathcal{D}_n), \varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)\}$ and a high probability parameter set Θ_n with $\mathbb{Q}(\Theta_n^\mathsf{c}) \leqslant C\mathrm{e}^{-n\rho}$ for some constant $C > 0$ that

$$\mathbb{E}\big[\bar{\Pi}_{\lambda,\rho}(\Theta_n \mid \mathcal{D}_n)\big] \leqslant \frac{2}{n\rho}\mathbb{E}\big[\,\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \mathbb{Q}\right)\big] + C\mathrm{e}^{-n\rho/2}. \tag{3.8}$$

Hence, for $\frac{\lambda}{n\rho} \to 0$ we could analyze the surrogate posterior via the Gibbs posterior itself at the cost of the approximation error $\frac{1}{n\rho}\,\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \Pi_\lambda(\cdot \mid \mathcal{D}_n)\right)$. Instead of this route, we will directly investigate $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ which especially allows for λ in the order of $n\rho$.

3.1.3 Corrected stochastic MALA

The computational advantage of the stochastic MH algorithm due to the reduction of the information parameter from n to ρn comes at the cost of a slower convergence rate, see Theorem 3.5.

To remedy this loss while retaining scalability, we define another joint target distribution as

$$\widetilde{\Pi}_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n) \propto \prod_{i=1}^{n} \left(\mathrm{e}^{-\frac{\lambda}{n}\ell_i(\vartheta)z_i}(1-\rho)^{1-z_i}\right)\Pi(\vartheta)$$

$$\propto \exp\Big(-\frac{\lambda}{n}\sum_{i=1}^n z_i\ell_i(\vartheta) - \log(1-\rho)\sum_{i=1}^n z_i\Big)\Pi(\vartheta), \qquad z\in\{0,1\}^n,$$

with marginal distribution in ϑ given by

$$\widetilde{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) = \sum_{z\in\{0,1\}^n} \widetilde{\Pi}_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n) \propto \prod_{i=1}^n \Big(\rho\frac{\mathrm{e}^{-\frac{\lambda}{n}\ell_i(\vartheta)}}{\rho} + 1 - \rho\Big)\Pi(\vartheta)$$
$$= \exp\big(-\lambda\widetilde{R}_{n,\rho}(\vartheta)\big)\Pi(\vartheta) \qquad \text{with}$$
$$\widetilde{R}_{n,\rho}(\vartheta) \coloneqq -\frac{1}{\lambda}\sum_{i=1}^n \log\big(\mathrm{e}^{-\frac{\lambda}{n}\ell_i(\vartheta)} + 1 - \rho\big). \tag{3.9}$$

Compared to $\bar{R}_{n,\rho}$ from (3.7) there is no ρ in the first term in the logarithm. In line with (2.3) and (2.4), we obtain the estimators

$$\widetilde{f}_{\lambda,\rho} \coloneqq f_{\widetilde{\vartheta}_{\lambda,\rho}} \qquad \text{for} \qquad \widetilde{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n \sim \widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \tag{3.10}$$

and

$$\bar{f}_{\lambda,\rho} \coloneqq \mathbb{E}\big[f_{\widetilde{\vartheta}_{\lambda,\rho}} \,\big|\, \mathcal{D}_n\big] = \int f_\vartheta \,\widetilde{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n). \tag{3.11}$$

To sample from $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ the MH algorithm with proposal density $q(\vartheta', z' \mid \vartheta, z) = q_{\mathrm{s}}(\vartheta' \mid \vartheta, z)\prod_{i=1}^n \rho^{z_i'}(1-\rho)^{1-z_i'}$ as in (3.5) leads to the acceptance probabilities

$$\alpha(\vartheta', z' \mid \vartheta, z) = \frac{q_{\mathrm{s}}(\vartheta \mid \vartheta', z')}{q_{\mathrm{s}}(\vartheta' \mid \vartheta, z)}\mathbb{1}_{[-B,B]^p}(\vartheta')\exp\Big(-\sum_{i=1}^n z_i'\big(\tfrac{\lambda}{n}\ell_i(\vartheta') + \log\rho\big)$$
$$+ \sum_{i=1}^n z_i\big(\tfrac{\lambda}{n}\ell_i(\vartheta) + \log\rho\big)\Big) \wedge 1.$$

To take the randomized batches into account, we thus introduce a small *correction term* $\frac{\log\rho}{\lambda}|Z| = \mathcal{O}_{\mathbb{P}}(\frac{n}{\lambda}\rho\log\rho)$ in the empirical risks. The resulting surrogate posterior $\widetilde{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n)$ achieves a considerably improved approximation of the Gibbs distribution $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$:

Lemma 3.2. *If f and all f_ϑ are bounded by some constant $C > 0$, then we have*

$$\frac{1}{n}\,\mathrm{KL}\big(\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \mid \Pi_{\lambda/(2-\rho)}(\cdot \mid \mathcal{D}_n)\big) \leqslant \Big(\frac{\lambda}{n}\Big)^2\Big(32C^4 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i^4\Big).$$

Compared to Lemma 3.1, the approximation error of $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ in terms of the Kullback-Leibler distance is now determined by the full sample size n instead of the possibly much smaller batch

size ρn as for the stochastic MH algorithm. The only price to pay is a reduction of the inverse temperature parameter λ by the factor $(2-\rho)^{-1} \in [1/2, 1]$. As already mentioned in (3.8), we can conclude contraction and coverage results for $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ by combining Ray & Szabó (2022, Theorem 5) with Lemma 3.2 if $\lambda/n \to 0$. A direct analysis of $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ will even allow for λ of the order n in our main results and thus lead to results as good as we can hope for the Gibbs measure itself.

The corrected stochastic MALA (csMALA) is summarized in Algorithm 3.1. The implementation omits the restriction of the proposed network weights to $[-B, B]^P$ which is practically negligible for sufficiently large constant B and the correction term $\frac{\log \rho}{\lambda}|Z| = \mathcal{O}_{\mathbb{P}}(\frac{n}{\lambda}\rho \log \rho)$ in the empirical risk is weighted by some tuning parameter $\zeta \geqslant 0$. For $\zeta = 0$ we recover the uncorrected method. In theory we always set $\zeta = 1$, but in practice the flexibility gained from choosing ζ was beneficial.

Algorithm 3.1 csMALA - corrected stochastic MALA

Input: inverse temperature $\lambda > 0$, learning rate $\gamma > 0$, standard deviation $s > 0$, correction parameter $\zeta \geqslant 0$, batch size $m \in \{1, \dots, n\}$, burn-in time $b \in \mathbb{N}$, gap length $c \in \mathbb{N}$, number of draws $N \in \mathbb{N}$.

1. Initialize $\vartheta^{(0)} \in \mathbb{R}^P$ and $Z^{(0)} \sim \mathrm{Ber}(\frac{m}{n})^{\otimes n}$.
2. Calculate $R_n^{(0)} = \frac{1}{n}\sum_{i=1}^n Z_i^{(0)} \ell_i(\vartheta^{(0)}) + \zeta \frac{\log \rho}{\lambda}|Z^{(0)}|$ and $\nabla R_n^{(0)} = \nabla_\vartheta R_n(\vartheta^{(0)}, Z^{(0)})$.
3. For $k = 0, \dots, b + cN$ do:
   (a) Draw $Z' \sim \mathrm{Ber}(\frac{m}{n})^{\otimes n}$.
   (b) Draw $\vartheta' \sim \mathcal{N}(\vartheta^{(k)} - \gamma \nabla R_n^{(k)}, s^2)$ and calculate $R_n' = \frac{1}{n}\sum_{i=1}^n Z_i' \ell_i(\vartheta') + \zeta \frac{\log \rho}{\lambda}|Z'|$ and $\nabla R_n' = \nabla_\vartheta R_n(\vartheta', Z')$.
   (c) Calculate acceptance probability

$$\alpha^{(k+1)} = \exp\Big(\lambda R_n^{(k)} + \frac{1}{2s^2}\big|\vartheta' - \vartheta^{(k)} + \gamma \nabla R_n^{(k)}\big|^2 - \lambda R_n' - \frac{1}{2s^2}\big|\vartheta^{(k)} - \vartheta' + \gamma \nabla R_n'\big|^2\Big).$$

   (d) Draw $u \sim \mathcal{U}([0,1])$.
   If $u \leqslant \alpha^{(k+1)}$, then set $\vartheta^{(k+1)} = \vartheta', R_n^{(k+1)} = R_n', \nabla R_n^{(k+1)} = \nabla R_n'$,
   else set $\vartheta^{(k+1)} = \vartheta^{(k)}, R_n^{(k+1)} = R_n^{(k)}, \nabla R_n^{(k+1)} = \nabla R_n^{(k)}$.

Output: $\widetilde{f}_{\lambda,\rho} = f_{\vartheta^{(b)}}$, $\bar{f}_{\lambda,\rho} = \frac{1}{N}\sum_{k=1}^N f_{\vartheta^{(b+ck)}}$

3.2 Application to stochastic neural networks

In this section, we apply the methodology from the previous section to neural networks and state the resulting statistical guarantees.

It is worth noting that our analysis is independent of the choice of the proposal distribution. We derive oracle inequalities for the estimators $\widehat{f}_{\lambda,\rho}$ (Theorem 3.5) and $\widetilde{f}_{\lambda,\rho}$ (Theorem 3.3) and as a consequence an analogous oracle inequality for $\bar{f}_{\lambda,\rho}$ (Corollary 3.6), which verify that these estimators are not much worse than the optimal choice for ϑ. We also discuss the properties of credible balls.

In the sequel, the class $\mathcal{F}$ is chosen as a class of neural networks. More precisely, we consider *feedforward multilayer perceptrons* with $p \in \mathbb{N}$ inputs, $L \in \mathbb{N}$ hidden layers and constant width $r \in \mathbb{N}$. The latter restriction is purely for notational convenience. The *rectified linear unit* (ReLU) $\phi(x) \coloneqq \max\{x, 0\}, x \in \mathbb{R}$, is used as activation function. We write $\phi_v x \coloneqq \big(\phi(x_i + v_i)\big)_{i=1,\dots,d}$ for vectors $x, v \in \mathbb{R}^d$. With this notation we can represent such neural networks as

$$g_\vartheta(\mathbf{x}) \coloneqq W^{(L+1)}\phi_{v^{(L)}}W^{(L)}\phi_{v^{(L-1)}}\cdots W^{(2)}\phi_{v^{(1)}}W^{(1)}\mathbf{x} + v^{(L+1)}, \qquad \mathbf{x} \in \mathbb{R}^p,$$

where the parameter vector ϑ contains all entries of the weight matrices $W^{(1)} \in \mathbb{R}^{r\times p}$, $W^{(2)}, \dots, W^{(L)} \in \mathbb{R}^{r\times r}$, $W^{(L+1)} \in \mathbb{R}^{1\times r}$ and the shift ('bias') vectors $v^{(1)}, \dots, v^{(L)} \in \mathbb{R}^r$, $v^{(L+1)} \in \mathbb{R}$. The total number of network parameters is $P \coloneqq (p+1)r + (L-1)(r+1)r + r + 1$. A possibly more intuitive layer-wise representation of g_ϑ is given by

$$\begin{aligned} \mathbf{x}^{(0)} &\coloneqq \mathbf{x} \in \mathbb{R}^p, \\ \mathbf{x}^{(l)} &\coloneqq \phi(W^{(l)}\mathbf{x}^{(l-1)} + v^{(l)}),\ l = 1, \dots, L, \\ g_\vartheta(\mathbf{x}) \coloneqq \mathbf{x}^{(L+1)} &\coloneqq W^{(L+1)}\mathbf{x}^{(L)} + v^{(L+1)}, \end{aligned} \tag{3.12}$$

where the activation function is applied coordinate-wise. We denote the class of all such functions g_ϑ by $\mathcal{G}(p, L, r)$. Note that these neural networks can be interpreted as an iterated composition of multi-index models with matrices $W^{(1)}, \dots, W^{(L)}$ and link functions $\phi(\cdot + v^{(1)}), \dots, \phi(\cdot + v^{(L)})$ with multivariate output. The term *dimension reduction matrices* is explicitly avoided here, as $W^{(2)}, \dots, W^{(L)}$ are square matrices and $W^{(1)} \in \mathbb{R}^{p\times r}$ increases the dimension if $r > p$.

For some $C \geqslant 1$, we introduce the class of clipped networks

$$\mathcal{F}(p, L, r, C) \coloneqq \big\{ f_\vartheta = (-C) \vee (g_\vartheta \wedge C) \,\big|\, g_\vartheta \in \mathcal{G}(p, L, r) \big\}.$$

3.2.1 Oracle inequality

As outlined in Chapter 2, we would like to compare the performance of the estimator $\widetilde{f}_{\lambda,\rho}$ from (3.10) to the best possible network f_{ϑ^*} for the *oracle choice*

$$\vartheta^* \in \operatorname*{arg\,min}_{\vartheta\in[-B,B]^P} R(\vartheta) = \operatorname*{arg\,min}_{\vartheta\in[-B,B]^P} \mathcal{E}(\vartheta). \tag{3.13}$$

A solution to the minimization problem in (3.13) always exists since $[-B,B]^P$ is compact and $\vartheta \mapsto R(\vartheta)$ is continuous. Similarly to Chapter 2, we need some mild assumptions on the regression model:

Assumption 3.A.

(a) ***Bounded regression function:*** *For some constant $C \geqslant 1$ we have $\|f\|_\infty \leqslant C$.*

(b) ***Second moment of inputs:*** *For some constant $K \geqslant 1$ we have $\mathbb{E}[|\mathbf{X}|^2] \leqslant pK$.*

(c) ***Conditional sub-Gaussianity of observation noise:*** *There are constants $\sigma, \Gamma > 0$ such that*

$$\mathbb{E}[|\varepsilon|^k \mid \mathbf{X}] \leqslant \frac{k!}{2}\sigma^2\Gamma^{k-2} \text{ a.s.}, \qquad \text{for all } k \geqslant 2.$$

(d) ***Conditional symmetry of observation noise:*** *ε is conditionally on $\mathbf{X}$ symmetric.*

Note that neither the loss function nor the data are assumed to be bounded. We obtain the following non-asymptotic oracle inequality:

Theorem 3.3 (PAC-Bayes oracle inequality for csMALA)**.** *Under Assumption 3.A there are constants $Q_0, Q_1 > 0$ depending only on C, Γ, σ such that for $\lambda = n/Q_0$ and sufficiently large n we have for all $\delta \in (0,1)$ with probability of at least $1-\delta$ that*

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{Q_1}{n}\big(PL\log(n) + \log(2/\delta)\big). \tag{3.14}$$

Remark 3.4. For $\rho = 1$ we do not need the conditional symmetry condition in Assumption 3.A. An explicit admissible choice for λ is $\lambda = n/\big(2^5C(\Gamma \vee (2C)) + 2^7(C^2+\sigma^2) + 2^3(\sigma C + \sigma^2)\big)$. The dependence of Q_1 on C, Γ, σ is at most quadratic and $n \geqslant n_0 = 2 \vee B \vee K \vee L \vee r \vee p$ is sufficiently large.

Theorem 3.3 can be seen as the counterpart to Theorem 2.5 from the multi-index setting. It is in line with classical PAC-Bayes oracle inequalities, see e.g. Guedj & Alquier (2013), Bissiri et al. (2016). In particular, Chérief-Abdellatif (2020) has obtained a similar oracle inequality

for a variational approximation of the Gibbs posterior distribution. A main step in the proof of Theorem 3.3 is to verify the compatibility between the risk $\widetilde{R}_{n,\rho}$ from (3.9) and the empirical risk R_n as established in Proposition 3.11.

We obtain a similar result for $\widehat{f}_{\lambda,\rho}$ from (3.6). Note that here the stochastic error term is of order $\mathcal{O}(\frac{PL}{n\rho})$ instead of $\mathcal{O}(\frac{PL}{n})$ as in Theorem 3.3 (up to logarithms).

Theorem 3.5 (Oracle inequality for sMALA)**.** *Under Assumption 3.A there are constants $Q_0', Q_1' > 0$ depending only on C, Γ, σ such that for $\lambda = n\rho/Q_0'$ and sufficiently large n we have for all $\delta \in (0,1)$ with probability of at least $1-\delta$ that*

$$\mathcal{E}(\widehat{f}_{\lambda,\rho}) \leqslant 4\mathcal{E}(f_{\vartheta^*}) + \frac{Q_1'}{n\rho}\big(PL\log(n) + \log(2/\delta)\big).$$

In view of Theorem 3.5 the following results are also true for the stochastic MH estimator if n is replaced by $n\rho$. However, we focus only on the analysis of $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ for the sake of clarity.

Denoting

$$r_n^2 \coloneqq 12\|f_{\vartheta^*} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} + \frac{Q_1}{n} PL\log(n), \tag{3.15}$$

we can rewrite (3.14) as

$$\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}\big(\{\vartheta : \|f_{\widetilde{\vartheta}_{\lambda,\rho}} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} > r_n^2 + t^2\} \,\big|\, \mathcal{D}_n\big)\big] \leqslant 2\mathrm{e}^{-nt^2/Q_1}, \qquad t > 0,$$

which is a *contraction rate* result in terms of a frequentist analysis of the nonparametric Bayes method. An immediate consequence is an oracle inequality for the posterior mean $\bar{f}_{\lambda,\rho}$ from (3.11).

Corollary 3.6 (Posterior mean)**.** *Under the conditions of Theorem 3.3 we have with probability of at least $1-\delta$ that*

$$\mathcal{E}(\bar{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{Q_2}{n}\big(PL\log(n) + \log(2/\delta)\big)$$

with a constant Q_2 only depending on C, Γ, σ from Assumption 3.A.

Using the approximation properties of neural networks, the oracle inequality yields the optimal rate of convergence (up to a logarithmic factor) over the following class of hierarchical

functions:

$$\mathcal{H}(q,\mathbf{d},\mathbf{t},\beta,C_0) \coloneqq \Big\{ g_q \circ \cdots \circ g_0 \colon [0,1]^p \to \mathbb{R} \,\Big|\, g_i = (g_{ij})_j^\top \colon [a_i,b_i]^{d_i} \to [a_{i+1},b_{i+1}]^{d_{i+1}},$$
$$g_{ij} \text{ depends on at most } t_i \text{ arguments, } g_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([a_i,b_i]^{t_i},C_0), \text{ for some } |a_i|,|b_i| \leqslant C_0 \Big\}, \tag{3.16}$$

where $\mathbf{d} \coloneqq (p,d_1,\dots,d_q,1) \in \mathbb{N}^{q+2}, \mathbf{t} \coloneqq (t_0,\dots,t_q) \in \mathbb{N}^{q+1}, \beta \coloneqq (\beta_0,...,\beta_q) \in (0,\infty)^{q+1}$ and where $\mathcal{C}_{t_i}^{\beta_i}([a_i,b_i]^{t_i},C_0)$ denote classical Hölder balls with Hölder regularity $\beta_i > 0$. For a detailed discussion of $\mathcal{H}(q,\mathbf{d},\mathbf{t},\beta,C_0)$, see Schmidt-Hieber (2020). In particular, this class covers multi-index models with Hölder regular link functions, which will later allow us to bridge the gap between neural networks and our multi-index analysis, see Corollary 4.6.

Theorem 3.3 reveals the following convergence rate which is in line with the upper bounds by Schmidt-Hieber (2020) and Kohler & Langer (2021):

Proposition 3.7 (Rates of convergence)**.** *Let $\mathbf{X} \in [0,1]^p$. In the situation of Theorem 3.3, there exists a network architecture $(L,r) = (C_1 \log n, C_2 (n/(\log n)^3)^{t^*/(4\beta^*+2t^*)})$ with $C_1, C_2 > 0$ only depending on upper bounds for $q, |\mathbf{d}|_\infty, |\beta|_\infty, C_0$ such that the estimators $\widetilde{f}_{\lambda,\rho}$ and $\bar{f}_{\lambda,\rho}$ satisfy for sufficiently large n uniformly over all hierarchical functions $f \in \mathcal{H}(q,\mathbf{d},\mathbf{t},\beta,C_0)$*

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant Q_3 \Big(\frac{(\log n)^3}{n} \Big)^{2\beta^*/(2\beta^*+t^*)} + Q_3 \frac{\log(2/\delta)}{n} \qquad \text{and}$$
$$\mathcal{E}(\bar{f}_{\lambda,\rho}) \leqslant Q_4 \Big(\frac{(\log n)^3}{n} \Big)^{2\beta^*/(2\beta^*+t^*)} + Q_4 \frac{\log(2/\delta)}{n}$$

with probability of at least $1-\delta$, respectively, where β^* and t^* are given by

$$\beta^* \coloneqq \beta_{i^*}^*, \qquad t^* \coloneqq t_{i^*}^* \qquad \text{for} \qquad i^* \in \operatorname*{arg\,min}_{i=0,\dots,q} \frac{2\beta_i^*}{2\beta_i^* + t_i^*} \qquad \text{and} \qquad \beta_i^* \coloneqq \beta_i \prod_{l=i+1}^{q} (\beta_l \wedge 1).$$

The constants Q_3 and Q_4 only depend on upper bounds for $q, \mathbf{d}, \beta$ and C_0 as well as the constants C, Γ, σ from Assumption 3.A.

Remark 3.8. One could aim for a similar result with a deep network architecture where $(L,r) = (C_1'(n/\log n)^{t^*/(4\beta^*+2t^*)}, C_2')$ to achieve the same rate of convergence just with $\log n$ instead of $(\log n)^3$. This would require an approximation result for this network architecture with a quantitative bound on the network weights. Moreover, the network depth $L = \mathcal{O}(\log n)$ is more suitable for deriving our uncertainty quantification results.

It has been proved by Schmidt-Hieber (2020) that this is the minimax-optimal rate of convergence for the nonparametric estimation of f from this class of hierarchical functions up to loga-

rithmic factors. Studying the special case of classical Hölder balls $\mathcal{C}_p^\beta([0,1]^p, C_0)$, a contraction rate of order $n^{-2\beta/(2\beta+p)}$ has been derived by Polson & Ročková (2018) and Chérief-Abdellatif (2020).

3.2.2 Credible sets

In addition to the contraction rates, the Bayesian approach offers a possibility for uncertainty quantification. For this, we assume that the distribution $\mathbb{P}^{\mathbf{X}}$ of $\mathbf{X}$ is known. We define the *credible ball*

$$\widehat{C}(\tau_\alpha) \coloneqq \{h \in L^2 : \|h - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \tau_\alpha\}, \qquad \alpha \in (0,1),$$

with critical values

$$\tau_\alpha \coloneqq \arg\min_{\tau>0} \big\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : \|f_\vartheta - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \tau \mid \mathcal{D}_n) > 1-\alpha\big\}.$$

By construction $\widehat{C}(\tau_\alpha)$ is the smallest L^2-ball around $\bar{f}_{\lambda,\rho}$ which contains $1-\alpha$ mass of the surrogate posterior measure. Despite the posterior belief, it is not necessarily guaranteed that the true regression function is contained in $\widehat{C}(\tau_\alpha)$. More precisely, the posterior distribution might be quite certain, in the sense that the credible ball is quite narrow, but suffers from a significant bias. In general, it might happen that $\mathbb{P}(f \in \widehat{C}(\tau_\alpha)) \to 0$, see e.g. Knapik et al. (2011, Theorem 4.2) in a Gaussian model. To circumvent this, Rousseau & Szabó (2020) have introduced inflated credible balls where the critical value is multiplied with a slowly diverging factor. While they proved that this method works in several classical nonparametric models with a sieve prior, our neural network setting causes an additional problem. In order to prove coverage, we would like to compare norms in the intrinsic parameter space, i.e. the space of the network weights, with the norm of the resulting predicted regression function. While the fluctuation of f_ϑ can be controlled via the fluctuation of ϑ, more precisely we have $\|f_\vartheta - f_{\vartheta'}\|_{L^2(\mathbb{P}^{\mathbf{X}})} = \mathcal{O}(\Delta(L,r) \cdot |\vartheta - \vartheta'|_\infty)$ with $\Delta(L,r) \coloneqq (2rB)^L$, see Lemma 3.14 below, the converse direction does not hold. Even locally around an oracle choice ϑ^* we cannot hope to control $|\vartheta|_\infty$ via $\|f_\vartheta\|_{L^2(\mathbb{P}^{\mathbf{X}})}$ in view of the ambiguous network parameterization. As a consequence, we define another critical value at the level of the parameter space

$$\tau_\alpha^\vartheta \coloneqq \arg\min_{\tau>0} \big\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : |\vartheta|_\infty \leqslant \Delta(L,r)^{-1}\tau \mid \mathcal{D}_n) > 1-\alpha\big\}.$$

Remark 3.9. The factor $\Delta(L,r)$ in the definition of τ_α^ϑ could be improved by a different geometry in the parameter space at the cost of a different approximation theory for the resulting network

classes. For instance, we may assume that all weight matrices are bounded by B in the ℓ^2-operator norm $\|\cdot\|_2$, which is in line with the weight scaling employed in the theory of neural tangent spaces, cf. Jacot et al. (2018). In this case a minor modification of Lemma 3.14 yields $\|f_\vartheta - f_{\vartheta'}\|_{L^2(\mathbb{P}^{\mathbf{x}})} = \mathcal{O}((2B)^L) \cdot \|\vartheta - \vartheta'\|$ where $\|\vartheta\|$ is defined as the maximal $\|\cdot\|_2$-norm of all weight matrices and all $|\cdot|_2$-norms of the biases. The resulting critical value is given by $\arg\min_{\tau>0}\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : \|\vartheta\| \leqslant (2B)^{-L}\tau \mid \mathcal{D}_n) > 1-\alpha\}$ avoiding the undesirable dependence on the network width r.

Both critical values measure the fluctuation of the posterior. The theoretical properties of the credible ball are summarized in the following theorem:

Theorem 3.10 (Credible balls)**.** *Under Assumption 3.A and with constants $Q_0, Q_1, Q_2 > 0$ from above we have for $\lambda = n/(2Q_0)$, r_n^2 from (3.15) and sufficiently large n that*

$$\mathbb{P}\Big(\operatorname{diam}\big(\widehat{C}(\tau_\alpha)\big) \leqslant 4\sqrt{2r_n^2 + \frac{4(Q_1 \vee Q_2)}{n}\log\frac{2}{\alpha}}\Big) \geqslant 1-\alpha.$$

If the depth L and the width r are chosen such that $L\log(n)\mathcal{E}(f_{\vartheta^*}) = \mathcal{O}(PL\log(n)/\lambda)$, then we have for some constant $\xi > \sqrt{L\log n}$ depending on K, p and α that

$$\mathbb{P}\big(f \in \widehat{C}(\xi\tau_\alpha^\vartheta)\big) \geqslant 1-\alpha.$$

Therefore, the order of the diameter of $\widehat{C}(\tau_\alpha)$ is of the best possible size if L and r are chosen as in Proposition 3.7. On the other hand, the larger credible set $\widehat{C}(\xi\tau_\alpha^\vartheta)$ defines an honest confidence set for a fixed class $\mathcal{H}(q, \mathbf{d}, \mathbf{t}, \beta, C_0)$ of the regression function if ξ is chosen sufficiently large depending on the class parameters. That is, $f \in \mathcal{H}(q, \mathbf{d}, \mathbf{t}, \beta, C_0)$ is contained in $\widehat{C}(\xi\tau_\alpha^\vartheta)$ with probability of at least $1-\alpha$. In that sense ξ is a non-asymptotic version of the inflation factor by Rousseau & Szabó (2020). To circumvent the unknown constant ξ, we can conclude from Theorem 3.10 that for any sequence $a_n \uparrow \infty$, e.g. $a_n = \log n$, we have

$$\mathbb{P}\big(f \in \widehat{C}(a_n\tau_\alpha^\vartheta)\big) \geqslant 1-\alpha \qquad \text{for sufficiently large } n.$$

The condition $L\log(n)\mathcal{E}(f_{\vartheta^*}) = \mathcal{O}(PL\log(n)/\lambda)$ for the coverage result means that the rate is dominated by the stochastic error term and can be achieved with a slightly larger network compared to Proposition 3.7. This guarantees that the posterior is not underfitting and that the posterior's bias is covered by its dispersal.

3.3 Numerical examples

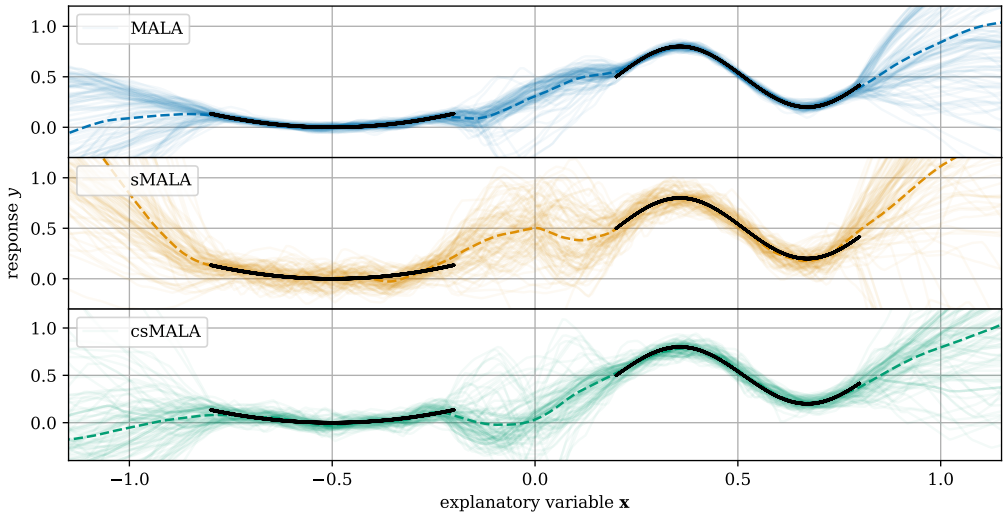

Figure 3.2: 100 samples drawn from the different MALA chains, given a training sample (black markers) of 10000 points. Random variables are drawn for $\rho = 0.1$. The dashed line shows the corresponding posterior mean $\bar{f}_\lambda$.

Section 3.1.3 introduces a correction to the batch-wise approximation of the empirical risk when calculating the MH step. In the following, we show the merit of this correction for learning a regression function using a feed-forward neural network of $L = 2$ layers of $r = 100$ nodes each and ReLU activation. To simplify the description of the experimental setup and to allow for a transparent graphical illustration of the method, see Fig. 3.2, we focus on the case of a one-dimensional regression function. The neural network has a total number of 10401 parameters. The training sample of size 10000 consist of two equally populated intervals $[-0.8, -0.2]$ and $[0.2, 0.8]$ with $\mathbf{X}_i \sim \mathcal{U}([-0.8, -0.2] \cup [0.2, 0.8])$. The true regression function is $f(\mathbf{x}) = 1.5(\mathbf{x} + 0.5)^2 \mathbb{1}_{\{\mathbf{x}<0\}} + (0.3\sin(10\mathbf{x} - 2) + 0.5)\mathbb{1}_{\{\mathbf{x}\geqslant 0\}}$. We generate $Y = f(\mathbf{X}) + \varepsilon$ by adding an observation error $\varepsilon \sim \mathcal{N}(0, 0.02^2)$. In the interval between -0.2 and 0.2 no data is produced in order to illustrate whether the methods recover the resulting large uncertainty due to missing data. For a sufficiently flexible model we expect a large spread between samples from each Markov chain in this region. Fig. 3.2 depicts exactly this behavior, as well as the training sample.

To compare the convergence of MALA, stochastic MALA (sMALA), and our corrected stochastic MALA (csMALA) within reasonable computation time, we initialize the chains with network parameters obtained through optimization of the empirical risk with stochastic gradient descent for 2000 steps. For this pre-training, we use a learning rate of 10^{-3}. The hyperparameters of

	MALA	sMALA	csMALA
λ	n	$n \cdot \rho$	$n \cdot (2 - \rho)$
γ	10^{-4}	10^{-4}	$10^{-4}/\rho$
s		$0.2/\sqrt{P}$	
b		$100000/\rho$	
c		5000	
N		20	

Table 3.1: Parameter choice for the different MALA chains. For $\rho = 0.1$, we chose a burn-in of $b = 50000$ to keep computation costs low.

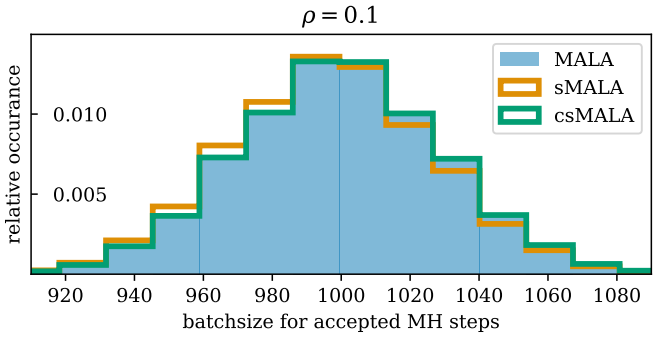

Figure 3.3: Histogram of the summed auxiliary variables, that is the number of training samples contributing to the stochastic risk, for all accepted steps. For MALA the MH acceptance step is calculated on the full sample and the distribution of the samples contribution to the risk gradients is thus unbiased by the batch size.

the subsequent chains are listed in Table 3.1. The inverse temperature is chosen to counteract the different normalization terms of the risk for (s)MALA and csMALA, as well as the reduction of the learning rate by $(2 - \rho)$ through the correction term from Section 3.1.3. The proposal noise level per parameter dimension is normalized with respect to the number of network parameters such that the total length of the noise vector is independent of the parameter space dimension.

To further improve the efficiency of the sampling, we restart Algorithm 3.1 with $\vartheta^{(0)}$ set to the last accepted parameters whenever no proposal has been accepted for 100 steps. Especially for small ρ and large ε, the stochastic MH algorithms exhibit the tendency to get stuck after accepting an outlier batch with low risk.

It is also important to adapt ζ such that $\zeta \frac{\log \rho}{\lambda} \approx \frac{1}{n} \sum_{i=1}^{n} \ell_i(\vartheta^{(k)})$. For ζ lower than this, a bias is introduced towards accepting updates where many points of the data sample contributed to the stochastic risk approximation due to the Bernoulli distributed auxiliary variables. Conversely, for higher values, updates are preferably accepted for low amounts of points in the risk approximation. This bias towards small batches, note the minus sign due to $\log \rho$, can also be observed for the uncorrected sMALA. It arises from the dependence of R_n on the sum of the drawn auxiliary variables Z_i. Fig. 3.3 shows a histogram of this sum for all accepted steps. A clear bias for sMALA towards small batches can be seen. To achieve a good correction, we update ζ every 100 steps to fulfill the preceding correspondence. Over the chain, the correction factor

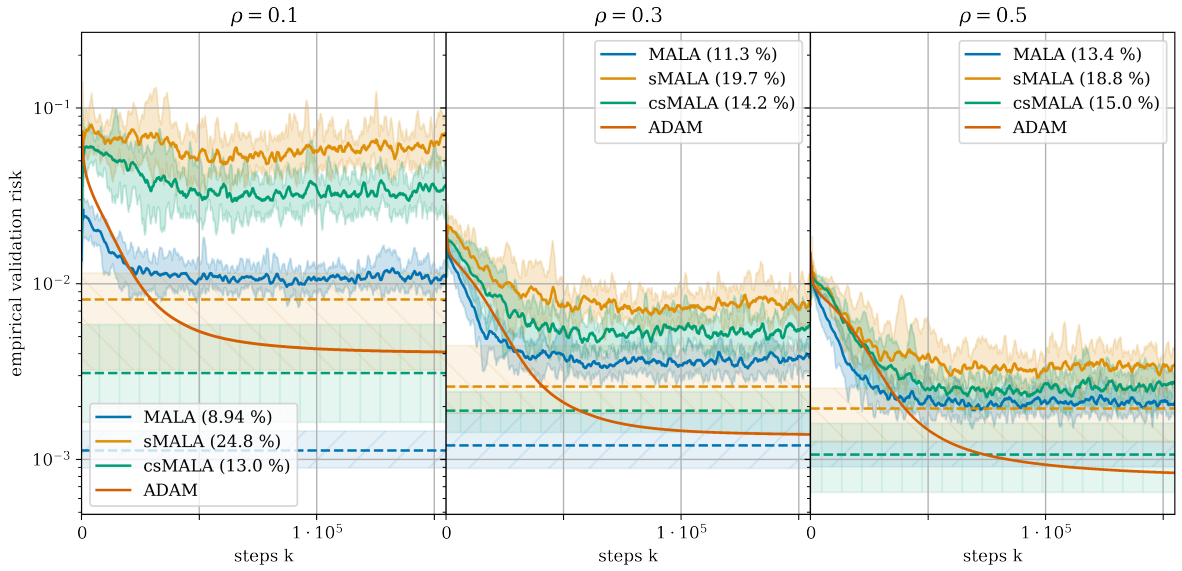

Figure 3.4: Average empirical risk on a validation set of 10000 points during running of the MALA chains. We show different batch probabilities ρ, as well as the values of the posterior mean (dashed lines). Uncertainties correspond to the minimum and maximum values of 10 identical chains. For clarity, as simple moving average over 1501 steps is plotted. In the legend, the average acceptance probability over all 10 chains is given. For easier interpretation of the risk values, we also show the behavior of a gradient-based optimization using `ADAM`.

thus decreases like the empirical risk with ζ close to 0 due to the proportionality to n^{-1}.

We quantify the performance of the estimators gathered from the different chains with an independent validation sample $\mathcal{D}_{n_{\text{val}}}^{\text{val}} := (\mathbf{X}_i^{\text{val}}, Y_i^{\text{val}})_{i=1,\dots,n_{\text{val}}} \subset \mathbb{R}^p \times \mathbb{R}$ of size $n_{\text{val}} = 10000$ drawn from the same intervals as the training sample and calculate the empirical validation risk

$$R_n(\widehat{f}) = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} \left(Y_i^{\text{val}} - \widehat{f}(\mathbf{X}_i^{\text{val}})\right)^2$$

during running of the chain. Fig. 3.4 illustrates the behavior of the empirical validation risk for the different MALA algorithms, as well as for a simple inference fit using `ADAM` (Kingma & Ba, 2014) with a learning rate of 10^{-3}. For a fair comparison, we calculate the gradient updates for all algorithms, including MALA and `ADAM`, from Bernoulli drawn batches, and only calculate the MH step for MALA using the full training sample. We can obverse that the individual samples of MALA outperform those of the sMALA chains, while the samples from the corrected chain achieve substantially better values than those of the uncorrected stochastic algorithm. On a level of individual samples, all chains are outperformed by the gradient-based optimization

using `ADAM`. Investigating the posterior means, MALA outperforms `ADAM` for small values of ρ where our corrected algorithm reaches similar risk values as the gradient-based optimization. For moderate values of ρ the corrected stochastic MALA restores the performance of the full MH step for both, posterior samples and posterior means, at a level similar to `ADAM`. While the acceptance rates of MALA decrease for low ρ and those of sMALA increase, the acceptance rates of the corrected algorithm are stable under variation of the average batch size.

To study the empirical coverage properties, we calculate 10 individual chains per algorithm and ρ, and estimate the credible sets and their average radii. As radius of our credible balls, we approximate the 99.5% quantile $q_{1-\alpha}$ of the mean squared distance to the posterior mean via

$$\tau_{\alpha,n} = q_{1-\alpha}\big((h_1, ..., h_N)\big) \qquad \text{with} \qquad h_k = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} \left| f_{\vartheta^{(b+ck)}}(\mathbf{X}_i^{\text{val}}) - \bar{f}_{\lambda,\rho}(\mathbf{X}_i^{\text{val}}) \right|^2 .$$

To determine the coverage probability, we then calculate the number of chains with a mean squared distance of the posterior mean to the true regression function not exceeding this radius. The results are shown in Table 3.2. While the uncertainty estimates of all algorithms remain conservative, we find that the correction term leads to considerably more precise credible sets.

To illustrate Theorem 3.3 and Theorem 3.5, we also investigate the scaling behavior of the empirical validation risk of the posterior means with the training sample size n while keeping $n\rho$ constant. We expect the risk of MALA to decrease with growing n, while sMALA should not decay due to the constant $n\rho$. The numerical simulation of Fig. 3.5 is in line with our theoretical results. For our corrected algorithm, we recover the scaling behavior of MALA as expected.

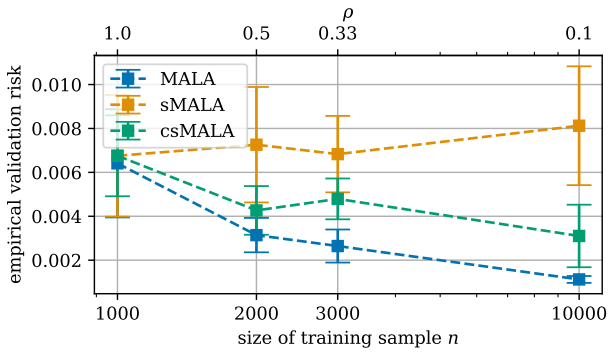

Figure 3.5: Scaling of the empirical risk of the posterior mean $\bar{f}$ on a 10000 point validation set with the size of the training sample. We scale ρ to keep the average batch size $n\rho = 1000$ constant. Error bars report the standard deviation of 10 identical chains.

ρ	MALA	sMALA	csMALA
0.1	1.42 ± 0.16	13.5 ± 1.4	7.72 ± 0.82
0.3	1.10 ± 0.15	3.70 ± 0.51	2.15 ± 0.23
0.5	1.28 ± 0.11	2.76 ± 0.19	1.91 ± 0.36

Table 3.2: Average radii $\tau_\alpha \cdot 10^3$ of credible sets for $\alpha = 0.005$ calculated from 10 Monte Carlo chains. All sets show a coverage probability $\widehat{C}(\tau_\alpha)$ of 100%.

3.4 Proofs

We will start by proving the main theorems. Additional proofs of auxiliary results are postponed to Section 3.4.6 and Section 3.4.7.

3.4.1 Compatibility between $\widetilde{R}_{n,\rho}$ and the excess risk

As outlined in Section 2.1, a key ingredient to PAC-Bayes oracle inequalities is a concentration inequality. The main difference now is that the modified empirical risk $\widetilde{R}_{n,\rho}$ arises from the stochastic MH step, but we would still like to quantify the performance of the estimator in terms of its excess risk $\mathcal{E}(\vartheta) = \mathbb{E}\big[\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\big]$. Therefore, the first step in our analysis is to verify the compatibility of these risks leading to the following concentration inequality, which acts as a counterpart to Lemma 2.4. A concentration inequality for the empirical risk $R_n(\vartheta) - R_n(f)$ follows as the special case where $\rho = 1$.

Proposition 3.11. *Grant Assumption 3.A. Define*

$$\widetilde{\mathcal{E}}_n(\vartheta) := \widetilde{R}_{n,\rho}(\vartheta) - \widetilde{R}_{n,\rho}(f).$$

and set $C_{n,\lambda} := \frac{\lambda}{n}\frac{8(C^2+\sigma^2)}{1-V\lambda/n}$, $V := 16C(\Gamma \vee 2C)$. *Then for all* $\lambda \in [0, n/V) \cap \big[0, \frac{n \log 2}{8(C^2+\sigma^2)}\big]$, $\rho \in (0, 1]$

and $n \in \mathbb{N}$ *we have*

$$\mathbb{E}\big[\exp\big(\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big] \leqslant \exp\big(\big(C_{n,\lambda} + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta)\big) \qquad \textit{and}$$
$$\mathbb{E}\big[\exp\big(-\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big] \leqslant \exp\big(\big(C_{n,\lambda} + \tfrac{3}{4} + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta)\big).$$

Proof. Define $\psi_\rho(x) \coloneqq -\log\big(\mathrm{e}^{-x} + 1 - \rho\big)$ such that

$$\widetilde{\mathcal{E}}_n(\vartheta) = \frac{1}{\lambda}\sum_{i=1}^{n}\big(\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big) - \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(f)\big)\big)$$

with $\ell_i(f) = (Y_i - f(\mathbf{X}_i))^2$.

We have

$$\widetilde{\mathcal{E}}_n(\vartheta) = \frac{1}{n}\sum_{i=1}^{n}\big(\ell_i(\vartheta) - \ell_i(f)\big)\psi_\rho'\big(\xi_i\tfrac{\lambda}{n}\ell_i(\vartheta) + (1-\xi_i)\tfrac{\lambda}{n}\ell_i(f)\big) \tag{3.17}$$

with some random variables $\xi_i \in [0,1]$. Using $\ell_1(\vartheta) - \ell_1(f) = \big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2 + 2\varepsilon_1\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)$, we can decompose the expectation of (3.17) into

$$\begin{aligned}\mathbb{E}\big[\widetilde{\mathcal{E}}_n(\vartheta)\big] &= \mathbb{E}\big[\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\psi_\rho'\big(\xi_1\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1)\tfrac{\lambda}{n}\ell_1(f)\big)\big] \\ &\qquad + 2\mathbb{E}\big[\varepsilon_1\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)\psi_\rho'\big(\xi_1\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1)\tfrac{\lambda}{n}\ell_1(f)\big)\big] \\ &\eqqcolon E_1 + E_2.\end{aligned}$$

We treat both terms separately. We have

$$\begin{aligned}1 \geqslant \psi_\rho'(x) &= (1 + (1-\rho)\mathrm{e}^x)^{-1} \\ &\geqslant \frac{1}{1 + 2(1-\rho)} \geqslant \frac{1}{3} \qquad \text{for } x \in [0, \log 2]\end{aligned}$$

and $\psi_\rho'(x) \in (0,1]$ for all $x \geqslant 0$. In particular, we observe

$$E_1 \leqslant \mathbb{E}\big[\big(f_\vartheta(\mathbf{X}_1) - f(\mathbf{X}_1)\big)^2\big] = \mathcal{E}(\vartheta).$$

If $|\varepsilon_1| \leqslant 2\sigma$, we have $\frac{\lambda}{n}\ell_1(\cdot) \leqslant \frac{\lambda}{n}8(C^2 + \sigma^2) \leqslant \log 2$ for $\frac{\lambda}{n} \leqslant \frac{\log 2}{8(C^2+\sigma^2)}$. Thus,

$$\begin{aligned}E_1 &\geqslant \mathbb{E}\big[\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\psi_\rho'\big(\xi_1\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1)\tfrac{\lambda}{n}\ell_1(f)\big)\mathbb{1}_{\{|\varepsilon_1|\leqslant 2\sigma\}}\big] \\ &\geqslant \frac{1}{3}\mathbb{E}\big[\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\mathbb{P}(|\varepsilon_1| \leqslant 2\sigma \mid \mathbf{X}_1)\big] \\ &= \frac{1}{3}\mathbb{E}\big[\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\big(1 - \mathbb{P}(|\varepsilon_1| > 2\sigma \mid \mathbf{X}_1)\big)\big] \\ &\geqslant \frac{1}{4}\mathbb{E}\big[\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\big],\end{aligned}$$

where we used Chebyshev's inequality in the last step. Hence, $\frac{1}{4}\mathcal{E}(\vartheta) \leqslant E_1 \leqslant \mathcal{E}(\vartheta)$. For E_2 we use $\mathbb{E}[\varepsilon_1\psi_\rho'(\frac{\lambda}{n}\varepsilon_1^2) \mid \mathbf{X}_1] = 0$ by symmetry together with $\ell_1(f) = \varepsilon_1^2$ to obtain for some random $\xi_1' \in [0,1]$ that

$$\begin{aligned}
E_2 &= 2\mathbb{E}\big[\varepsilon_1\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)\big(\psi_\rho'\big(\tfrac{\lambda}{n}\ell_1(f) + \xi_1\tfrac{\lambda}{n}\big(\ell_1(\vartheta) - \ell_1(f)\big)\big) - \psi_\rho'\big(\tfrac{\lambda}{n}\ell_1(f)\big)\big)\big] \\
&= \frac{2\lambda}{n}\mathbb{E}\big[\varepsilon_1\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)\xi_1\big(\ell_1(\vartheta) - \ell_1(f)\big)\psi_\rho''\big(\xi_1'\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1')\tfrac{\lambda}{n}\ell_1(f)\big)\big] \\
&= \frac{\lambda}{n}\mathbb{E}\big[2\xi_1\big(\varepsilon_1\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^3 + 2\varepsilon_1^2\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\big) \\
&\qquad\qquad \cdot \psi_\rho''\big(\xi_1'\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1')\tfrac{\lambda}{n}\ell_1(f)\big)\big].
\end{aligned}$$

Since $\max_{y\geqslant 0}\frac{y}{(1+y)^2} = \frac{1}{4}$, we have

$$|\psi_\rho''(x)| = \frac{(1-\rho)\mathrm{e}^x}{(1+(1-\rho)\mathrm{e}^x)^2} \leqslant \frac{1}{4} \qquad \text{for } x \geqslant 0.$$

Therefore,

$$|E_2| \leqslant \frac{\lambda}{n}\big(\tfrac{1}{2}\mathbb{E}\big[|\varepsilon_1||f_\vartheta(\mathbf{X}_1) - f(\mathbf{X}_1)|^3 + 2\varepsilon_1^2\big(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)\big)^2\big]\big) \leqslant \frac{\lambda}{n}\big(\sigma C + \sigma^2\big)\mathcal{E}(\vartheta).$$

In combination with the bounds for E_1, we obtain

$$\big(\tfrac{1}{4} - \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\mathcal{E}(\vartheta) \leqslant \mathbb{E}\big[\widetilde{\mathcal{E}}_n(\vartheta)\big] \leqslant \big(1 + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\mathcal{E}(\vartheta).$$

Define $Z_i(\vartheta) \coloneqq \frac{n}{\lambda}\big(\psi_\rho\big(\frac{\lambda}{n}\ell_i(\vartheta)\big) - \psi_\rho\big(\frac{\lambda}{n}\ell_i(f)\big)\big)$ such that $\widetilde{\mathcal{E}}_n(\vartheta) = \frac{1}{n}\sum_{i=1}^n Z_i(\vartheta)$. The previous bounds for $\mathbb{E}[\widetilde{\mathcal{E}}_n(\vartheta)]$ yield

$$\begin{aligned}
\mathbb{E}\big[\exp\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \lambda\mathcal{E}(\vartheta)\big)\big] &= \mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n (Z_i(\vartheta) - \mathbb{E}[Z_i(\vartheta)])}\big]\mathrm{e}^{\lambda(\mathbb{E}[\widetilde{\mathcal{E}}_n(\vartheta)] - \mathcal{E}(\vartheta))} \\
&\leqslant \mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n (Z_i(\vartheta) - \mathbb{E}[Z_i(\vartheta)])}\big]\mathrm{e}^{\frac{\lambda^2}{n}(\sigma C + \sigma^2)\mathcal{E}(\vartheta)} \qquad \text{and} \\
\mathbb{E}\big[\exp\big(-\lambda\widetilde{\mathcal{E}}_n(\vartheta) + \lambda\mathcal{E}(\vartheta)\big)\big] &= \mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n (-Z_i(\vartheta) - \mathbb{E}[-Z_i(\vartheta)])}\big]\mathrm{e}^{\lambda(\mathcal{E}(\vartheta) - \mathbb{E}[\widetilde{\mathcal{E}}_n(\vartheta)])} \\
&\leqslant \mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n (-Z_i(\vartheta) - \mathbb{E}[-Z_i(\vartheta)])}\big]\mathrm{e}^{(\frac{3\lambda}{4} + \frac{\lambda^2}{n}(\sigma C + \sigma^2))\mathcal{E}(\vartheta)}.
\end{aligned}$$

To bound the centered exponential moments, we use a variant of Bernstein's inequality, see Massart (2007, inequality (2.21)) similarly to the proof of Lemma 2.4. The second moments are bounded by

$$\begin{aligned}
\mathbb{E}[Z_i^2] &= \mathbb{E}\big[\big(\tfrac{n}{\lambda}\big(\psi_\rho\big(\tfrac{\lambda}{n}\ell_1(\vartheta)\big) - \psi_\rho\big(\tfrac{\lambda}{n}\ell_1(f)\big)\big)\big)^2\big] \\
&= \mathbb{E}\big[\big(\big(\ell_1(\vartheta) - \ell_1(f)\big)\psi_\rho'\big(\xi_1\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1)\tfrac{\lambda}{n}\ell_1(f)\big)\big)^2\big]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\big[\big((f_\vartheta(\mathbf{X}_1) - f(\mathbf{X}_1))^2 + 2\varepsilon_1(f_\vartheta(\mathbf{X}_1) - f(\mathbf{X}_1))\big)^2(\psi_\rho')^2\big(\xi_1\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1)\tfrac{\lambda}{n}\ell_1(f)\big)\big] \\
&\leqslant 2\mathbb{E}\big[\big(f_\vartheta(\mathbf{X}_1) - f(\mathbf{X}_1)\big)^4 + 4\varepsilon_1^2\big(f_\vartheta(\mathbf{X}_1) - f(\mathbf{X}_1)\big)^2\big] \\
&\leqslant 8\big(C^2 + \sigma^2\big)\mathcal{E}(\vartheta) =: U.
\end{aligned}
$$

Moreover, we have for $k \geqslant 3$

$$
\begin{aligned}
\mathbb{E}\big[(Z_i)_+^k\big] &\leqslant \mathbb{E}\big[\big|\ell_1(\vartheta) - \ell_1(f)\big|^k\big|\psi_\rho'\big(\xi_1\tfrac{\lambda}{n}\ell_1(\vartheta) + (1-\xi_1)\tfrac{\lambda}{n}\ell_1(f)\big)\big|^k\big] \\
&\leqslant \mathbb{E}\big[\big|\ell_1(\vartheta) - \ell_1(f)\big|^k\big] \\
&= \mathbb{E}[|f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1) + 2\varepsilon_1|^k|f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1)|^{k-2}(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1))^2] \\
&\leqslant (2C)^{k-2}\mathbb{E}[|f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1) + 2\varepsilon_1|^k(f(\mathbf{X}_1) - f_\vartheta(\mathbf{X}_1))^2] \\
&\leqslant (2C)^{k-2}2^{k-1}((2C)^k + k!2^{k-1}\sigma^2\Gamma^{k-2})\mathcal{E}(\vartheta) \\
&\leqslant (2C)^{k-2}k!8^{k-2}\big((2C)^{k-2} \vee \Gamma^{k-2}\big)U \\
&= k!UV^{k-2}.
\end{aligned}
$$

Hence, the aforementioned variant of Bernstein's inequality yields

$$
\mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n (Z_i(\vartheta) - \mathbb{E}[Z_i(\vartheta)])}\big] \leqslant \exp\Big(\frac{U\lambda^2}{n(1 - V\lambda/n)}\Big) = \exp\big(C_{n,\lambda}\lambda\mathcal{E}(\vartheta)\big)
$$

for $C_{n,\lambda}$ as defined in Proposition 3.11. The same bound remains true if we replace Z_i by $-Z_i$. We conclude

$$
\begin{aligned}
\mathbb{E}\big[\exp\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \lambda\mathcal{E}(\vartheta)\big)\big] &\leqslant \exp\big(\big(C_{n,\lambda} + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta)\big) \qquad \text{and} \\
\mathbb{E}\big[\exp\big(-\lambda\widetilde{\mathcal{E}}_n(\vartheta) + \lambda\mathcal{E}(\vartheta)\big)\big] &\leqslant \exp\big(\big(C_{n,\lambda} + \tfrac{3}{4} + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta)\big). \qquad \square
\end{aligned}
$$

Remark 3.12. Replacing ψ_ρ by $\bar{\psi}_\rho(x) := -\log\big(\rho\mathrm{e}^{-x/\rho} + 1 - \rho\big)$, $x \geqslant 0$, and using

$$
\begin{aligned}
1 \geqslant \bar{\psi}_\rho'(x) &= (\rho + (1-\rho)\mathrm{e}^{x/\rho})^{-1} \\
&\geqslant \frac{1}{\rho + 3(1-\rho)} \geqslant \frac{1}{3} \qquad \text{for } x \in [0, \rho\log 3],
\end{aligned}
$$

we can analogously prove under Assumption 3.A that $\bar{\mathcal{E}}_n(\vartheta) := \bar{R}_{n,\rho}(\vartheta) - \bar{R}_{n,\rho}(f)$ with $\bar{R}_{n,\rho}$ from (3.7) satisfies for all $\lambda \in [0, n/V) \cap \big[0, \frac{n\log 3}{8(C^2+\sigma^2)}\big]$, $\rho \in (0,1]$ and $n \in \mathbb{N}$:

$$
\begin{aligned}
\mathbb{E}\big[\exp\big(\lambda\big(\bar{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big] &\leqslant \exp\big(\big(C_{n,\lambda} + \tfrac{\lambda}{n\rho}4(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta)\big) \qquad \text{and} \\
\mathbb{E}\big[\exp\big(-\lambda\big(\bar{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big] &\leqslant \exp\big(\big(C_{n,\lambda} + \tfrac{1}{4} + \tfrac{\lambda}{n\rho}4(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta)\big).
\end{aligned}
$$

3.4.2 A PAC-Bayes bound for csMALA

In combination with Proposition 3.11, we can verify a PAC-Bayes bound for the excess risk. The basic proof strategy is in line with the PAC-Bayes literature as outlined in Section 2.1.

Proposition 3.13 (PAC-Bayes bound)**.** *Grant Assumption 3.A. For any sample-dependent (in a measurable way) probability measure $\varrho \ll \Pi$, any $\lambda \in (0, n/V)$, and $\rho \in (0,1]$ such that $C_{n,\lambda} + \frac{\lambda}{n}(\sigma C + \sigma^2) \leqslant \frac{1}{8}$, we have*

$$\mathcal{E}(\widetilde{\vartheta}_{\lambda,\rho}) \leqslant 9 \int \mathcal{E} \, \mathrm{d}\varrho + \frac{16}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(2/\delta)\big) \tag{3.18}$$

with probability of at least $1-\delta$.

Proof. Proposition 3.11 yields

$$\begin{aligned}
&\mathbb{E}\big[\exp\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \big(1 + C_{n,\lambda} + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta) - \log(\delta^{-1})\big)\big] \leqslant \delta, \\
&\mathbb{E}\big[\exp\big(\lambda\big(\tfrac{1}{4} - C_{n,\lambda} - \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\mathcal{E}(\vartheta) - \lambda\widetilde{\mathcal{E}}_n(\vartheta) - \log(\delta^{-1})\big)\big] \leqslant \delta.
\end{aligned}$$

Integrating in ϑ with respect to the prior probability measure Π and applying Fubini's theorem, we conclude

$$\begin{aligned}
&\mathbb{E}\Big[\int \exp\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \big(1 + C_{n,\lambda} + \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\lambda\mathcal{E}(\vartheta) - \log(\delta^{-1})\big) \, \mathrm{d}\Pi(\vartheta)\Big] \leqslant \delta \quad \text{and} \quad (3.19) \\
&\mathbb{E}\Big[\int \exp\big(\lambda\big(\tfrac{1}{4} - C_{n,\lambda} - \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\mathcal{E}(\vartheta) - \lambda\widetilde{\mathcal{E}}_n(\vartheta) - \log(\delta^{-1})\big) \, \mathrm{d}\Pi(\vartheta)\Big] \leqslant \delta.
\end{aligned}$$

The Radon-Nikodym derivative of the posterior distribution $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \ll \Pi$ with respect to Π is given by

$$\frac{\mathrm{d}\widetilde{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n)}{\mathrm{d}\Pi} = \widetilde{D}_\lambda^{-1} \exp\Big(- \sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big)\Big)$$

with

$$\widetilde{D}_\lambda \coloneqq \int \mathrm{e}^{-\lambda \widetilde{R}_{n,\rho}(\vartheta)} \, \Pi(\mathrm{d}\vartheta) = \int \exp\Big(- \sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big)\Big) \, \Pi(\mathrm{d}\vartheta). \tag{3.20}$$

We obtain

$$\begin{aligned}
\delta &\geqslant \mathbb{E}_{\mathcal{D}_n}\Big[\int \exp\big(\lambda\big(\tfrac{1}{4} - C_{n,\lambda} - \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\mathcal{E}(\vartheta) - \lambda\widetilde{\mathcal{E}}_n(\vartheta) - \log(\delta^{-1})\big) \, \mathrm{d}\Pi(\vartheta)\Big] \\
&= \mathbb{E}_{\mathcal{D}_n, \widetilde{\vartheta} \sim \widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)}\Big[\exp\Big(\lambda\big(\tfrac{1}{4} - C_{n,\lambda} - \tfrac{\lambda}{n}(\sigma C + \sigma^2)\big)\mathcal{E}(\widetilde{\vartheta}) - \lambda\widetilde{\mathcal{E}}_n(\widetilde{\vartheta}) \\
&\qquad\qquad - \log(\delta^{-1}) - \log\Big(\frac{\mathrm{d}\widetilde{\Pi}_{\lambda,\rho}(\widetilde{\vartheta} \mid \mathcal{D}_n)}{\mathrm{d}\Pi}\Big)\Big)\Big]
\end{aligned}$$

$$= \mathbb{E}_{\mathcal{D}_n,\widetilde{\vartheta}\sim\widetilde{\Pi}_{\lambda,\rho}(\cdot|\mathcal{D}_n)}\Big[\exp\Big(\lambda\big(\tfrac{1}{4} - C_{n,\lambda} - \tfrac{\lambda}{n}(\sigma C+\sigma^2)\big)\mathcal{E}(\widetilde{\vartheta}) - \lambda\widetilde{\mathcal{E}}_n(\widetilde{\vartheta}) \\ - \log(\delta^{-1}) + \sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\widetilde{\vartheta})\big) + \log\widetilde{D}_\lambda\Big)\Big].$$

Since $\mathbb{1}_{[0,\infty)}(x) \leqslant \mathrm{e}^{\lambda x}$ for all $x \in \mathbb{R}$, we deduce with probability not larger than δ that

$$\big(\tfrac{1}{4} - C_{n,\lambda} - \tfrac{\lambda}{n}(\sigma C+\sigma^2)\big)\mathcal{E}(\widetilde{\vartheta}) - \widetilde{\mathcal{E}}_n(\widetilde{\vartheta}) + \frac{1}{\lambda}\sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\widetilde{\vartheta})\big) - \frac{1}{\lambda}\big(\log(\delta^{-1}) - \log\widetilde{D}_\lambda\big) \geqslant 0.$$

Provided $C_{n,\lambda} + \frac{\lambda}{n}(\sigma C+\sigma^2) \leqslant \frac{1}{8}$, we thus have for $\widetilde{\vartheta} \sim \widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ with probability of at least $1-\delta$:

$$\begin{aligned}\mathcal{E}(\widetilde{\vartheta}) &\leqslant 8\Big(\widetilde{\mathcal{E}}_n(\widetilde{\vartheta}) - \frac{1}{\lambda}\sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\widetilde{\vartheta})\big) + \frac{1}{\lambda}\big(\log(\delta^{-1}) - \log\widetilde{D}_\lambda\big)\Big) \\ &\leqslant 8\Big(-\frac{1}{\lambda}\sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(f)\big) + \frac{1}{\lambda}\big(\log(\delta^{-1}) - \log\widetilde{D}_\lambda\big)\Big).\end{aligned}$$

Lemma 2.1 with $h = -\sum_{i=1}^n \psi_\rho(\frac{\lambda}{n}\ell_i(\vartheta))$ yields

$$\log\widetilde{D}_\lambda = -\inf_{\varrho\ll\Pi}\Big(\mathrm{KL}(\varrho \mid \Pi) + \int\sum_{i=1}^n \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big)\,\mathrm{d}\varrho(\vartheta)\Big). \tag{3.21}$$

Therefore, we have with probability of at least $1-\delta$:

$$\begin{aligned}\mathcal{E}(\widetilde{\vartheta}) &\leqslant 8\inf_{\varrho\ll\Pi}\Big(\int\frac{1}{\lambda}\sum_{i=1}^n\big(\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big) - \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(f)\big)\big)\,\mathrm{d}\varrho(\vartheta) + \frac{1}{\lambda}\big(\log(\delta^{-1}) + \mathrm{KL}(\varrho \mid \Pi)\big)\Big) \\ &\leqslant 8\inf_{\varrho\ll\Pi}\Big(\int\widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta) + \frac{1}{\lambda}\big(\log(\delta^{-1}) + \mathrm{KL}(\varrho \mid \Pi)\big)\Big).\end{aligned}$$

In order to reduce the integral $\int\widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)$ to $\int\mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta)$, we use $C_{n,\lambda} + \frac{\lambda}{n}(\sigma C+\sigma^2) \leqslant \frac{1}{8}$, Jensen's inequality, and (3.19) to obtain for any probability measure $\varrho \ll \Pi$ (which may depend on $\mathcal{D}_n$)

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_n}\Big[\exp\Big(\int\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \tfrac{9}{8}\lambda\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho(\vartheta) - \mathrm{KL}(\varrho \mid \Pi) - \log(\delta^{-1})\Big)\Big] \\ &\quad= \mathbb{E}_{\mathcal{D}_n}\Big[\exp\Big(\int\Big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \tfrac{9}{8}\lambda\mathcal{E}(\vartheta) - \log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(\vartheta)\Big) - \log(\delta^{-1})\Big)\,\mathrm{d}\varrho(\vartheta)\Big)\Big] \\ &\quad\leqslant \mathbb{E}_{\mathcal{D}_n,\vartheta\sim\varrho}\Big[\exp\Big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \tfrac{9}{8}\lambda\mathcal{E}(\vartheta) - \log\Big(\frac{\mathrm{d}\varrho}{\mathrm{d}\Pi}(\vartheta)\Big) - \log(\delta^{-1})\Big)\Big] \\ &\quad\leqslant \mathbb{E}_{\mathcal{D}_n}\Big[\int\exp\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta) - \big(1 + C_{n,\lambda} + \tfrac{\lambda}{n}(\sigma C+\sigma^2)\big)\lambda\mathcal{E}(\vartheta) - \log(\delta^{-1})\big)\,\mathrm{d}\Pi(\vartheta)\Big] \leqslant \delta.\end{aligned}$$

Using $\mathbb{1}_{[0,\infty)}(x) \leqslant \mathrm{e}^{\lambda x}$ again, we conclude with probability of at least $1-\delta$:

$$\int \widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta) \leqslant \frac{9}{8}\int \mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta) + \lambda^{-1}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(\delta^{-1})\big).$$

Therefore, we conclude with probability of at least $1-2\delta$

$$\mathcal{E}(\widetilde{\vartheta}) \leqslant 9\int \mathcal{E}(\vartheta)\,\mathrm{d}\varrho(\vartheta) + \frac{16}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(\delta^{-1})\big). \qquad \square$$

3.4.3 Proof of Theorem 3.3

We fix a radius $\eta \in (0,1]$ and apply Proposition 3.13 with $\varrho = \varrho_\eta$ defined via

$$\frac{\mathrm{d}\varrho_\eta}{\mathrm{d}\Pi}(\vartheta) \propto \mathbb{1}_{\{|\vartheta-\vartheta^*|_\infty \leqslant \eta\}}$$

with ϑ^* from (3.13). Note that indeed $C_{n,\lambda} + \frac{\lambda}{n}(\sigma C + \sigma^2) \leqslant \frac{1}{8}$ for Q_0 sufficiently large. In order to control the integral term, we decompose

$$\begin{aligned}
\int \mathcal{E}\,\mathrm{d}\varrho_\eta &= \mathcal{E}(\vartheta^*) + \int \mathbb{E}\big[(f_\vartheta(\mathbf{X}) - f(\mathbf{X}))^2 - (f_{\vartheta^*}(\mathbf{X}) - f(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta) \\
&= \mathcal{E}(\vartheta^*) + \int \mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta) \\
&\qquad + 2\int \mathbb{E}\big[(f(\mathbf{X}) - f_{\vartheta^*}(\mathbf{X}))(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))\big]\,\mathrm{d}\varrho_\eta(\vartheta) \\
&\leqslant \mathcal{E}(\vartheta^*) + \int \mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta) \\
&\qquad + 2\int \mathbb{E}\big[(f(\mathbf{X}) - f_{\vartheta^*}(\mathbf{X}))^2\big]^{1/2}\mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))^2\big]^{1/2}\,\mathrm{d}\varrho_\eta(\vartheta) \\
&\leqslant \frac{4}{3}\mathcal{E}(\vartheta^*) + 4\int \mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta), \qquad (3.22)
\end{aligned}$$

using $2ab \leqslant \frac{a^2}{3} + 3b^2$ in the last step. To bound the remainder, we use the Lipschitz continuity of the map $\vartheta \mapsto f_\vartheta(\mathbf{x})$ for fixed $\mathbf{x} \in \mathbb{R}^p$:

Lemma 3.14. *Let $\vartheta, \widetilde{\vartheta} \in [-B,B]^P$. Then we have for $\mathbf{x} \in \mathbb{R}^p$ that*

$$|f_\vartheta(\mathbf{x}) - f_{\widetilde{\vartheta}}(\mathbf{x})| \leqslant 4(2rB)^L(|\mathbf{x}|_1 \vee 1)|\vartheta - \widetilde{\vartheta}|_\infty.$$

We obtain

$$\int \mathcal{E}\,\mathrm{d}\varrho_\eta \leqslant \frac{4}{3}\mathcal{E}(\vartheta^*) + \frac{4}{n^2} \qquad \text{for} \qquad \eta = \frac{1}{8K(2rB)^L p\,n}. \qquad (3.23)$$

It remains to bound the Kullback-Leibler term in (3.18) which can be done with the following lemma.

Lemma 3.15. *We have* $\mathrm{KL}(\varrho_\eta \mid \Pi) \leqslant P \log(2B/\eta)$.

Inserting (3.23) and the bound from Lemma 3.15 into the PAC-Bayes bound (3.18), we conclude

$$\begin{aligned}\mathcal{E}(\widetilde{\vartheta}_{\lambda,\rho}) &\leqslant 12\mathcal{E}(\vartheta^*) + \frac{36}{n^2} + \frac{16}{\lambda}\big(P\log\big(16BK(2rB)^L pn\big) + \log(2/\delta)\big)\\ &\leqslant 12\mathcal{E}(\vartheta^*) + \frac{Q_1}{n}\big(PL\log(n) + \log(2/\delta)\big)\end{aligned}$$

for some constant Q_1 only depending on C, σ, Γ. □

3.4.4 Proof of Theorem 3.5

Due to Remark 3.12, we can prove the following PAC-Bayes bound under Assumption 3.A analogously to Proposition 3.13: For any sample-dependent (in a measurable way) probability measure $\varrho \ll \Pi$ and any $\lambda \in (0, n/V)$ and $\rho \in (0,1]$ such that $C_{n,\lambda} + \frac{\lambda}{n\rho} 4(\sigma C + \sigma^2) \leqslant \frac{1}{4}$, we have

$$\mathcal{E}(\widehat{\vartheta}_\lambda) \leqslant \frac{5}{2}\int \mathcal{E}\,\mathrm{d}\varrho + \frac{4}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(2/\delta)\big)$$

with probability of at least $1-\delta$. From here, we can continue as in Section 3.4.3. □

3.4.5 Proof of Theorem 3.10

Choosing $\lambda = \frac{n}{2Q_0}$, Theorem 3.3 and Corollary 3.6 yield

$$\min\big\{\mathbb{E}[\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : \|f_\vartheta - f\|_{L^2(\mathbb{P}^{\mathbf{x}})} \leqslant s_n\} \mid \mathcal{D}_n)], \mathbb{P}(\|f - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{x}})} \leqslant s_n)\big\} \geqslant 1 - \frac{\alpha^2}{2}$$

with $s_n^2 \coloneqq 2r_n^2 + \frac{4(Q_1 \vee Q_2)}{n}\log\frac{2}{\alpha}$. We conclude

$$\begin{aligned}\mathbb{P}\big(\operatorname{diam}(\widehat{C}(\tau_\alpha)) \leqslant 4s_n\big) &= \mathbb{P}\Big(\sup_{g,h\in\widehat{C}(\tau_\alpha)} \|g-h\|_{L^2(\mathbb{P}^{\mathbf{x}})} \leqslant 4s_n\Big)\\ &\geqslant \mathbb{P}\Big(\sup_{g,h\in\widehat{C}(\tau_\alpha)} \|g-\bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{x}})} + \|\bar{f}_{\lambda,\rho}-h\|_{L^2(\mathbb{P}^{\mathbf{x}})} \leqslant 4s_n\Big)\\ &\geqslant \mathbb{P}\big(\tau_\alpha \leqslant 2s_n\big)\\ &= \mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : \|f_\vartheta - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{x}})} \leqslant 2s_n\} \mid \mathcal{D}_n) > 1-\alpha\big)\end{aligned}$$

$$
\begin{aligned}
&\geqslant \mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : \|f_\vartheta - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} > 2s_n\} \mid \mathcal{D}_n) < \alpha\big)\\
&= 1 - \mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : \|f_\vartheta - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} > 2s_n\} \mid \mathcal{D}_n) \geqslant \alpha\big)\\
&\geqslant 1 - \alpha^{-1}\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : \|f_\vartheta - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} > 2s_n\} \mid \mathcal{D}_n)\big]\\
&\geqslant 1 - \alpha^{-1}\big(\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : \|f_\vartheta - f\|_{L^2(\mathbb{P}^{\mathbf{X}})} > s_n\} \mid \mathcal{D}_n)\big]\\
&\qquad\qquad + \mathbb{P}\big(\|\bar{f}_{\lambda,\rho} - f\|_{L^2(\mathbb{P}^{\mathbf{X}})} > s_n\big)\big)\\
&\geqslant 1 - \alpha.
\end{aligned}
$$

The first statement in Theorem 3.10 is thus verified.

For the coverage statement, we denote $\bar{\xi} \coloneqq \xi\Delta(L,r) = \xi(2rB)^L$ and bound

$$
\begin{aligned}
\mathbb{P}\big(f \in \widehat{C}(\xi\tau_\alpha^\vartheta)\big) &= \mathbb{P}\big(\|f - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \xi\tau_\alpha^\vartheta\big)\\
&\geqslant \mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : |\vartheta|_\infty \leqslant \bar{\xi}^{-1}\|f - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})}\} \mid \mathcal{D}_n) < 1 - \alpha\big)\\
&\geqslant \mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : |\vartheta|_\infty \leqslant \bar{\xi}^{-1}s_n\} \mid \mathcal{D}_n) < 1 - \alpha\big) - \alpha^2\\
&= 1 - \alpha^2 - \mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\{\vartheta : |\vartheta|_\infty \leqslant \bar{\xi}^{-1}s_n\} \mid \mathcal{D}_n) \geqslant 1 - \alpha\big)\\
&\geqslant 1 - \alpha^2 - (1-\alpha)^{-1}\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(B_n \mid \mathcal{D}_n)\big]
\end{aligned}
$$

with

$$
B_n \coloneqq \big\{\vartheta : |\vartheta|_\infty \leqslant \bar{\xi}^{-1}s_n\big\}.
$$

In terms of $\widetilde{\mathcal{E}}_n(\vartheta) = \widetilde{R}_{n,\rho}(\vartheta) - \widetilde{R}_{n,\rho}(f)$ and $\widetilde{D}_\lambda = \int \exp\big(-\lambda\widetilde{R}_{n,\rho}(\vartheta)\big)\,\Pi(\mathrm{d}\vartheta)$ the inequalities by Cauchy-Schwarz and Jensen imply

$$
\begin{aligned}
\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(B_n \mid \mathcal{D}_n)\big] &= \mathbb{E}\Big[\widetilde{D}_\lambda^{-1}\int_{B_n} \mathrm{e}^{-\lambda\widetilde{R}_{n,\rho}(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big]\\
&= \mathbb{E}\Big[\widetilde{D}_\lambda^{-1}\mathrm{e}^{-\lambda\widetilde{R}_{n,\rho}(f)}\int_{B_n} \mathrm{e}^{-\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big]\\
&\leqslant \mathbb{E}\big[\widetilde{D}_\lambda^{-2}\mathrm{e}^{-2\lambda\widetilde{R}_{n,\rho}(f)}\big]^{1/2}\mathbb{E}\Big[\Big(\int_{B_n} \mathrm{e}^{-\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big)^2\Big]^{1/2}\\
&\leqslant \mathbb{E}\big[\widetilde{D}_\lambda^{-2}\mathrm{e}^{-2\lambda\widetilde{R}_{n,\rho}(f)}\big]^{1/2}\mathbb{E}\Big[\Pi(B_n)\int_{B_n} \mathrm{e}^{-2\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big]^{1/2}.
\end{aligned}
$$

The smaller choice of $\lambda = n/(2Q_0)$ instead of n/Q_0 ensures $C_{n,2\lambda} + \frac{2\lambda}{n}(\sigma C + \sigma^2) \leqslant \frac{1}{8}$ allowing us to apply Proposition 3.11 with 2λ. With Fubini's theorem and the uniform distribution of the prior, the second factor can thus be bounded using

$$
\begin{aligned}
\mathbb{E}\Big[\int_{B_n} \mathrm{e}^{-2\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big] &= \int_{B_n} \mathbb{E}\big[\mathrm{e}^{-2\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\big]\,\Pi(\mathrm{d}\vartheta)\\
&\leqslant \int_{B_n} \exp\big(2\big(C_{n,2\lambda} + \tfrac{3}{4} + \tfrac{2\lambda}{n}(\sigma C + \sigma^2) - 1\big)\lambda\mathcal{E}(\vartheta)\big)\,\Pi(\mathrm{d}\vartheta)
\end{aligned}
$$

$$\begin{aligned}&\leqslant \Pi(B_n)\\&\leqslant \exp\Big(P\log\frac{s_n}{B\bar{\xi}}\Big).\end{aligned}$$

Based on (3.21), we conclude

$$\begin{aligned}\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(B_n \mid \mathcal{D}_n)\big] &\leqslant \exp\Big(P\log\frac{s_n}{B\bar{\xi}}\Big)\mathbb{E}\big[\widetilde{D}_\lambda^{-2}\mathrm{e}^{-2\lambda\widetilde{R}_{n,\rho}(f)}\big]^{1/2}\\&= \exp\Big(P\log\frac{s_n}{B\bar{\xi}}\Big)\mathbb{E}\Big[\exp\Big(\inf_{\varrho\ll\Pi}\Big(2\,\mathrm{KL}(\varrho\mid\Pi)+2\int\lambda\widetilde{R}_{n,\rho}(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big)-2\lambda\widetilde{R}_{n,\rho}(f)\Big)\Big]^{1/2}\\&= \exp\Big(P\log\frac{s_n}{B\bar{\xi}}\Big)\mathbb{E}\Big[\exp\Big(\inf_{\varrho\ll\Pi}\Big(2\,\mathrm{KL}(\varrho\mid\Pi)+\int 2\lambda\widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big)\Big)\Big]^{1/2}.\end{aligned}$$

For $\varrho_{\eta'}$ defined via

$$\frac{\mathrm{d}\varrho_{\eta'}}{\mathrm{d}\Pi}(\vartheta)\propto\mathbb{1}_{\{|\vartheta-\vartheta^*|_\infty\leqslant\eta'\}},\qquad \eta'=\frac{s_n}{8K\Delta(L,r)p\sqrt{L\log n}},$$

we can use (3.22), Lemma 3.14, and Lemma 3.15 to obtain

$$\begin{aligned}&\inf_{\varrho\ll\Pi}\Big(\mathrm{KL}(\varrho\mid\Pi)+\int\lambda\widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big)\\&\qquad\leqslant \mathrm{KL}(\varrho_{\eta'}\mid\Pi)+\frac{4}{3}\lambda\mathcal{E}(\vartheta^*)+3\lambda\int\mathbb{E}\big[\big(f_{\vartheta^*}(\mathbf{X})-f_\vartheta(\mathbf{X})\big)^2\big]\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\&\qquad\qquad+\lambda\int\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\&\qquad\leqslant P\log\frac{2B}{\eta'}+\frac{4}{3}\lambda\mathcal{E}(\vartheta^*)+3L^{-1}\lambda s_n^2+\lambda\int\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta).\end{aligned}$$

In the sequel, $Q_{10}, Q_{11} > 0$ are numerical constants which may depend on C, Γ, σ, K, p, and α. Since $L\log(n)\mathcal{E}(\vartheta^*)\leqslant s_n^2\leqslant Q_{10}PL\log(n)/\lambda$ by assumption, we obtain

$$\begin{aligned}\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(B_n\mid\mathcal{D}_n)\big]&\leqslant\exp\big(-P\log\bar{\xi}+P\log\big(16K\Delta(L,r)p\sqrt{L\log n}\big)+5Q_{10}P\big)\\&\qquad\cdot\mathbb{E}\Big[\exp\Big(2\lambda\int\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\Big)\Big]^{1/2}\\&\leqslant\exp\big(-P\log\xi+P(Q_{11}+\log\sqrt{L\log n})\big)\\&\qquad\cdot\mathbb{E}\Big[\int\exp\big(2\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\Big]^{1/2},\end{aligned}$$

applying Jensen's inequality in the last line. To bound the expectation in the previous line, we apply Fubini's theorem, Proposition 3.11 with $C_{n,2\lambda}+\frac{2\lambda}{n}(\sigma C+\sigma^2)\leqslant\frac{1}{8}$, and Lemma 3.14 to

obtain

$$\begin{aligned}
\mathbb{E}\Big[\int \exp\big(2\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\Big] &= \int \mathbb{E}\big[\exp\big(2\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\big)\big]\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\
&\leqslant \int \exp\big(2\lambda\big(C_{n,2\lambda}+\tfrac{2\lambda}{n}(\sigma C+\sigma^2)\big)\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\
&\leqslant \int \exp\big(\tfrac{1}{2}\lambda\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\
&\leqslant \int \exp\big(\lambda\big(\mathcal{E}(\vartheta^*)+\|f_\vartheta-f_{\vartheta^*}\|^2_{L^2(\mathbb{P}^{\mathbf{X}})}\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\
&\leqslant \int \exp\big(\lambda\big(\mathcal{E}(\vartheta^*)+s_n^2/L\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\
&\leqslant \mathrm{e}^{2Q_{10}P\log n}.
\end{aligned}$$

We conclude

$$\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(B_n\mid\mathcal{D}_n)\big]\leqslant\exp\big(-P\big(\log\xi-Q_{11}-Q_{10}\log n-\log\sqrt{L\log n}\big)\big).$$

For a sufficiently large $\xi\geqslant\sqrt{L\log n}$, we obtain $\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(B_n\mid\mathcal{D}_n)\big]\leqslant\alpha(1-\alpha)^2$ and thus

$$\mathbb{P}\big(f\in\widehat{C}(\xi r_\alpha^\vartheta)\big)\geqslant 1-\alpha^2-\alpha(1-\alpha)\geqslant 1-\alpha.$$ □

3.4.6 Remaining proofs for Section 3.2

3.4.6.1 Proof of Lemma 3.1

Define

$$D_\lambda\coloneqq\int\exp\big(-\lambda R_n(\vartheta)\big)\,\Pi(\mathrm{d}\vartheta),\qquad \bar{D}_\lambda\coloneqq\int\exp\big(-\lambda\bar{R}_{n,\rho}(\vartheta)\big)\,\Pi(\mathrm{d}\vartheta).$$

For the first part of the lemma, we write

$$\begin{aligned}
\mathrm{KL}\big(\bar{\Pi}_{\lambda,\rho}(\cdot\mid\mathcal{D}_n)\,\big|\,\Pi_\lambda(\cdot\mid\mathcal{D}_n)\big) &= \int\log\frac{\mathrm{d}\bar{\Pi}_{\lambda,\rho}(\vartheta\mid\mathcal{D}_n)}{\mathrm{d}\Pi_\lambda(\cdot\mid\mathcal{D}_n)}\,\bar{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta\mid\mathcal{D}_n)\\
&= \lambda\int S_n(\vartheta)\,\bar{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta\mid\mathcal{D}_n)+\log\frac{D_\lambda}{\bar{D}_\lambda}\qquad\text{with}\\
S_n(\vartheta)&\coloneqq R_n(\vartheta)-\bar{R}_{n,\rho}(\vartheta).
\end{aligned}$$

By concavity of the logarithm we have

$$\frac{1}{\lambda}\sum_{i=1}^n\log\big(\rho\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}+1-\rho\big)\geqslant\frac{1}{\lambda}\sum_{i=1}^n\rho\log\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}+(1-\rho)\log 1=-\frac{1}{n}\sum_{i=1}^n\ell_i(\vartheta)=-R_n(\vartheta).$$

Hence, $S_n(\vartheta) \geqslant 0$ and $D_\lambda \leqslant \bar{D}_\lambda$. We conclude

$$\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \Pi_\lambda(\cdot \mid \mathcal{D}_n)\right) \leqslant \lambda \int S_n(\vartheta)\, \bar{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n).$$

Moreover, $\log(x+1) \leqslant x$ for all $x > -1$ and a second order Taylor expansion of $x \mapsto \mathrm{e}^x$ yield

$$\begin{aligned} S_n(\vartheta) &= \frac{1}{\lambda} \sum_{i=1}^n \left(\log\left(\rho(\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} - 1) + 1\right) + \tfrac{\lambda}{n}\ell_i(\vartheta)\right) \\ &\leqslant \frac{\rho}{\lambda} \sum_{i=1}^n \left(\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} - 1 + \tfrac{\lambda}{n\rho}\ell_i(\vartheta)\right) \\ &\leqslant \frac{\rho}{2\lambda} \sum_{i=1}^n \left(\tfrac{\lambda}{n\rho}\ell_i(\vartheta)\right)^2 \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} \\ &\leqslant \frac{\lambda}{n\rho} \cdot \frac{1}{2n} \sum_{i=1}^n |\ell_i(\vartheta)|^2. \end{aligned}$$

For $\ell_i(\vartheta) = |Y_i - f_\vartheta(\mathbf{X}_i)|^2 \leqslant 2|f(\mathbf{X}_i) - f_\vartheta(\mathbf{X}_i)|^2 + 2\varepsilon_i^2 \leqslant 8C^2 + 2\varepsilon_i^2$, we obtain

$$S_n(\vartheta) \leqslant \frac{\lambda}{n\rho}\Big(64C^4 + \frac{4}{n}\sum_{i=1}^n \varepsilon_i^4\Big)$$

and thus,

$$\begin{aligned} \frac{1}{\lambda}\,\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \Pi_\lambda(\cdot \mid \mathcal{D}_n)\right) &\leqslant \frac{\lambda}{n\rho}\Big(64C^4 + \frac{4}{n}\sum_{i=1}^n \varepsilon_i^4\Big) \int \bar{\Pi}(\mathrm{d}\vartheta \mid \mathcal{D}_n) \\ &= \frac{\lambda}{n\rho}\Big(64C^4 + \frac{4}{n}\sum_{i=1}^n \varepsilon_i^4\Big). \end{aligned}$$

In the regime $\rho \to 0$, define

$$T_n(\vartheta) \coloneqq -\rho n \cdot \frac{1}{n}\sum_{i=1}^n \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} \qquad \text{and} \qquad D_{\varpi,\lambda} \coloneqq \int \exp\big(-T_n(\vartheta)\big)\,\Pi(\mathrm{d}\vartheta),$$

such that

$$\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)\right) = \int \big(T_n(\vartheta) - \lambda \bar{R}_{n,\rho}(\vartheta)\big)\,\bar{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n) + \log\Big(\frac{D_{\varpi,\lambda}}{\bar{D}_\lambda}\Big).$$

We have

$$\begin{aligned}\lambda\bar{R}_{n,\rho}(\vartheta) - T_n(\vartheta) &= -\sum_{i=1}^{n} \log\big(\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\big) - T_n(\vartheta) \\ &= -n\log(1-\rho) - \sum_{i=1}^{n} \big(\log\big(\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\big) - \log(1-\rho)\big) - T_n(\vartheta) \\ &= -n\log(1-\rho) - \sum_{i=1}^{n} \rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} \int_0^1 \big(t\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\big)^{-1} \mathrm{d}t - T_n(\vartheta) \\ &= -n\log(1-\rho) - \sum_{i=1}^{n} \rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} \int_0^1 \Big(\frac{1}{t\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho} - 1\Big) \mathrm{d}t,\end{aligned}$$

where $(t\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho)^{-1} - 1 \in [0, \frac{\rho}{1-\rho}]$. Therefore,

$$-\frac{\rho^2}{(1-\rho)} \sum_{i=1}^{n} \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} \leqslant \lambda\bar{R}_{n,\rho}(\vartheta) - T_n(\vartheta) + n\log(1-\rho) \leqslant 0.$$

This implies $\log\big(\frac{D_{\varpi,\lambda}}{D_\lambda}\big) \leqslant -n\log(1-\rho)$ and thus,

$$\mathrm{KL}\left(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)\right) \leqslant \frac{\rho^2}{1-\rho} \int \sum_{i=1}^{n} \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}\, \bar{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n) \leqslant \frac{\rho^2 n}{1-\rho}. \qquad \square$$

3.4.6.2 Proof of Lemma 3.2

Recall $\psi_\rho(x) = -\log(\mathrm{e}^{-x} + 1 - \rho)$, $\psi'_\rho(x) = \frac{1}{1+(1-\rho)\mathrm{e}^x}$, and $\psi''_\rho(x) = -\frac{(1-\rho)\mathrm{e}^x}{(1+(1-\rho)\mathrm{e}^x)^2} \in [-1/4, 0]$. Since

$$\begin{aligned}\widetilde{R}_{n,\rho}(\vartheta) &= \frac{1}{\lambda}\sum_{i=1}^{n} \psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big) \\ &= \frac{n}{\lambda}\psi_\rho(0) + \frac{1}{\lambda}\sum_{i=1}^{n} \tfrac{\lambda}{n}\ell_i(\vartheta)\psi'_\rho\big(\xi_i\tfrac{\lambda}{n}\ell_i(\vartheta)\big) \\ &= \frac{n}{\lambda}\psi_\rho(0) + \frac{\psi'_\rho(0)}{n}\sum_{i=1}^{n}\ell_i(\vartheta) + \frac{1}{n}\sum_{i=1}^{n}\ell_i(\vartheta)\big(\psi'_\rho(\xi_i\tfrac{\lambda}{n}\ell_i(\vartheta)) - \psi'_\rho(0)\big) \\ &= -\frac{n}{\lambda}\log(2-\rho) + \frac{1}{2-\rho}R_n(\vartheta) + \frac{\lambda}{n^2}\sum_{i=1}^{n}\ell_i(\vartheta)^2\xi_i\psi''_\rho\big(\xi'_i\tfrac{\lambda}{n}\ell_i(\vartheta)\big),\end{aligned}$$

we have

$$-\frac{\lambda^2}{4n^2}\sum_{i=1}^{n}\ell_i(\vartheta)^2 \leqslant \lambda\widetilde{R}_{n,\rho}(\vartheta) - \frac{\lambda}{2-\rho}R_n(\vartheta) + n\log(2-\rho) \leqslant 0.$$

Therefore, we have with $\widetilde{D}_\lambda$ from (3.20) that

$$\begin{aligned}
&\mathrm{KL}\left(\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \,\middle|\, \Pi_{\lambda/(2-\rho)}(\cdot \mid \mathcal{D}_n)\right) \\
&\qquad = \int \Big(\frac{\lambda}{2-\rho} R_n(\vartheta) - \lambda \widetilde{R}_{n,\rho}(\vartheta)\Big)\, \widetilde{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n) + \log\Big(\frac{D_{\lambda/(2-\rho)}}{\widetilde{D}_\lambda}\Big) \\
&\qquad \leqslant \int \Big(\frac{\lambda}{2-\rho} R_n(\vartheta) - \lambda \widetilde{R}_{n,\rho}(\vartheta) - n\log(2-\rho)\Big)\, \widetilde{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n) \\
&\qquad \leqslant \frac{\lambda^2}{4n} \int \frac{1}{n} \sum_{i=1}^n \ell_i(\vartheta)^2\, \widetilde{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n) \\
&\qquad \leqslant \frac{\lambda^2}{n}\Big(32C^4 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i^4\Big). \qquad \square
\end{aligned}$$

3.4.6.3 Proof of Corollary 3.6

Jensen's and Markov's inequality yield for r_n^2 from (3.15) that

$$\begin{aligned}
&\mathbb{P}\Big(\mathcal{E}(\bar{f}_{\lambda,\rho}) > r_n^2 + \frac{Q_1}{n} + \frac{Q_1}{n}\log(2/\delta)\Big) \\
&\qquad = \mathbb{P}\Big(\|\mathbb{E}[f_{\widetilde{\vartheta}_{\lambda,\rho}} \mid \mathcal{D}_n] - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} > r_n^2 + \frac{Q_1}{n} + \frac{Q_1}{n}\log(2/\delta)\Big) \\
&\qquad \leqslant \mathbb{P}\Big(\mathbb{E}\big[\|f_{\widetilde{\vartheta}_{\lambda,\rho}} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} \,\big|\, \mathcal{D}_n\big] > r_n^2 + \frac{Q_1}{n} + \frac{Q_1}{n}\log(2/\delta)\Big) \\
&\qquad = \mathbb{P}\Big(\int_{\frac{Q_1}{n}\log(2/\delta)}^\infty \widetilde{\Pi}_{\lambda,\rho}\big(\|f_{\widetilde{\vartheta}_{\lambda,\rho}} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} > r_n^2 + t \,\big|\, \mathcal{D}_n\big)\,\mathrm{d}t > \frac{Q_1}{n}\Big) \\
&\qquad \leqslant \frac{n}{Q_1}\int_{\frac{Q_1}{n}\log(2/\delta)}^\infty \mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}\big(\|f_{\widetilde{\vartheta}_\lambda} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} > r_n^2 + t \,\big|\, \mathcal{D}_n\big)\big]\,\mathrm{d}t.
\end{aligned}$$

Using Theorem 3.3, we thus obtain

$$\mathbb{P}\Big(\mathcal{E}(\bar{f}_{\lambda,\rho}) > r_n^2 + \frac{Q_1}{n} + \frac{Q_1}{n}\log(2/\delta)\Big) \leqslant \frac{2n}{Q_1}\int_{\frac{Q_1}{n}\log(2/\delta)}^\infty \mathrm{e}^{-nt/Q_1}\,\mathrm{d}t = \delta. \qquad \square$$

3.4.6.4 Proof of Proposition 3.7

We combine arguments from Schmidt-Hieber (2020) with the approximation results from Kohler & Langer (2021). By rescaling, we can rewrite

$$f = f_q \circ \cdots \circ f_0 = h_q \circ \cdots \circ h_0$$

with $h_i = (h_{ij})_{j=1,\dots,d_{i+1}}$, where $\widetilde{h}_{0j} \in \mathcal{C}_{t_0}^{\beta_0}([0,1]^{t_0}, 1)$, $\widetilde{h}_{ij} \in \mathcal{C}_{t_i}^{\beta_i}\big([0,1]^{t_i}, (2C_0)^{\beta_i}\big)$ for $i = 1, \dots, q-1$ and $\widetilde{h}_{qj} \in \mathcal{C}_{t_q}^{\beta_q}\big([0,1]^{t_q}, C_0(2C_0)^{\beta_q}\big)$ and h_{ij} is $\widetilde{h}_{ij}$ understood as a function in d_i instead of t_i arguments.

We want to show that there exists a constant C_i such that for any $M_i \in \mathbb{N}$ we can find sufficiently large $L_i, r_i \in \mathbb{N}$ and a neural network $\widetilde{g}_{ij} \in \mathcal{G}(t_i, L_i, r_i)$ with $P_{L_i,r_i} = c_i M^{t_i}$ parameters and

$$\|\widetilde{h}_{ij} - \widetilde{g}_{ij}\|_{L^\infty([0,1]^{t_i})} \leqslant C_i M_i^{-2\beta_i}. \tag{3.24}$$

To construct such g_{ij}, we use Theorem 2(a) from Kohler & Langer (2021). Their conditions

(a) $L_i \geqslant 5 + \lceil \log_4(M^{2\beta_i}) \rceil \big(\lceil \log_2(\max\{\lfloor \beta_i \rfloor, t_i\} + 1) \rceil + 1 \big)$ and

(b) $r_i \geqslant 2^{t_i+6} \binom{t_i + \lfloor \beta_i \rfloor}{t_i} t_i^2 (\lfloor \beta_i \rfloor + 1) M_i^{t_i}$

can be satisfied for $L_i = C_i \log(M_i), r_i = C_i M_i^{t_i}$, where C_i only depends on upper bounds for t_i and β_i. Hence, there exists a neural network $\widetilde{g}_{ij} \in \mathcal{G}(t_i, L_i, r_i)$ with (3.24). Careful inspection of the proof of this theorem reveals that the weights and shifts of $\widetilde{g}_{ij}$ grow at most polynomially in M. Since $t_i \leqslant d_i, r_i$, we can easily embed $\widetilde{g}_{ij}$ into the class $\mathcal{G}(d_i, L_i, r_i)$ by setting $g_{ij} = \widetilde{g}_{ij}(W_{ij}\,\cdot)$, where the matrix $W_{ij} \in \mathbb{R}^{t_i \times d_i}$ is chosen such that g_{ij} depends on the same t_i many arguments as h_{ij}. Note that the approximation accuracy of $\widetilde{g}_{ij}$ carries over to g_{ij}, that is

$$\|h_{ij} - g_{ij}\|_{L^\infty([0,1]^{d_i})} \leqslant \|\widetilde{h}_{ij} - \widetilde{g}_{ij}\|_{L^\infty([0,1]^{t_i})} \leqslant C_i M_i^{-2\beta_i}. \tag{3.25}$$

Setting $g = g_q \circ \cdots \circ g_0$ with $g_i = (g_{ij})_j$, we obtain a neural network $g \in \mathcal{G}(p, L, r)$ with $r = \max_{i=0,\dots,q,} r_i d_{i+1}$ and $L = \sum_{i=0}^q L_i$.

Counting the number of parameters of g and using $L_i = C_i M_i^{t_i}$, we get $P_{L,r} \leqslant Q_{12} \sum_{i=0}^q L_i r_i^2$ for some $Q_{12} > 0$.

It follows from Schmidt-Hieber (2020, Lemma 3) and (3.25) that

$$\|f - g\|_{L^\infty([0,1]^p)} \leqslant C_0 \prod_{l=0}^{q-1} (2C_0)^{\beta_{l+1}} \sum_{i=0}^q \big\| |h_i - g_i|_\infty \big\|_{L^\infty([0,1]^{d_i})}^{\prod_{l=i+1}^q \beta_l \wedge 1} \leqslant Q_{13} \sum_{i=0}^q M_i^{-2\beta_i},$$

for some $Q_{13} > 0$.

Applying Theorem 3.3 together with $\mathcal{E}(f_{\vartheta^*}) \leqslant \|f - g\|^2_{L^\infty([0,1]^p)}$, we now obtain for some Q_{14} that

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant Q_{14} \sum_{i=0}^q M_i^{-4\beta_i} + \frac{Q_{14}}{n} \sum_{i=0}^q M_i^{2t_i} (\log n)^3 + Q_{14} \frac{\log(2/\delta)}{n} \tag{3.26}$$

with probability of at least $1-\delta$. Choosing $M_i = \lceil (n/(\log n)^3)^{1/(4\beta_i+2t_i)} \rceil$ ensures $L, r \leqslant n$ for sufficiently large n, balances the first two terms in the upper bound (3.26) and thus yields the asserted convergence rate for $\widetilde{f}_{\lambda,\rho}$.

The convergence rate for the posterior mean is proved analogously using Corollary 3.6. □

3.4.7 Proofs of the auxiliary results

3.4.7.1 Proof of Lemma 3.14

Set $\eta := |\vartheta - \widetilde{\vartheta}|_\infty$ and let $W^{(1)}, \dots, W^{(L+1)}$, $v^{(1)}, \dots, v^{(L+1)}$ and $\widetilde{W}^{(1)}, \dots, \widetilde{W}^{(L+1)}$, $\widetilde{v}^{(1)}, \dots, \widetilde{v}^{(L+1)}$ be the weights and shifts associated with ϑ and $\widetilde{\vartheta}$, respectively. Define $\widetilde{\mathbf{x}}^{(0)}, \dots, \widetilde{\mathbf{x}}^{(L+1)}$ analogously to (3.12). We can recursively deduce from the Lipschitz continuity of φ that for $l = 2, \dots, L$:

$$
\begin{aligned}
|\mathbf{x}^{(1)}|_1 &\leqslant |W^{(1)}\mathbf{x}|_1 + |v^{(1)}|_1 \leqslant 2rB(|\mathbf{x}|_1 \vee 1),\\
|\mathbf{x}^{(1)} - \widetilde{\mathbf{x}}^{(1)}|_1 &\leqslant |W^{(1)}\mathbf{x}^{(0)} + v^{(1)} - \widetilde{W}^{(1)}\widetilde{\mathbf{x}}^{(0)} - \widetilde{v}^{(1)}|_1 \leqslant \eta 2r(|\mathbf{x}|_1 \vee 1),\\
|\mathbf{x}^{(l)}|_1 &\leqslant |W^{(l)}\mathbf{x}^{(l-1)}|_1 + |v^{(l)}|_1 \leqslant 2rB(|\mathbf{x}^{(l-1)}|_1 \vee 1) \qquad \text{and}\\
|\mathbf{x}^{(l)} - \widetilde{\mathbf{x}}^{(l)}|_1 &\leqslant |W^{(l)}\mathbf{x}^{(l-1)} + v^{(l)} - \widetilde{W}^{(l)}\widetilde{\mathbf{x}}^{(l-1)} - \widetilde{v}^{(l)}|_1\\
&\leqslant |(W^{(l)} - \widetilde{W}^{(l)})\mathbf{x}^{(l-1)}|_1 + |\widetilde{W}^{(l)}(\mathbf{x}^{(l-1)} - \widetilde{\mathbf{x}}^{(l-1)})|_1 + |v^{(l)} - \widetilde{v}^{(l)}|_1\\
&\leqslant \eta 2r(|\mathbf{x}^{(l-1)}|_1 \vee 1) + rB|\mathbf{x}^{(l-1)} - \widetilde{\mathbf{x}}^{(l-1)}|_1.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
|\mathbf{x}^{(L)}|_1 &\leqslant (2rB)^{L-1}(|\mathbf{x}^{(1)}|_1 \vee 1) \leqslant (2rB)^L(|\mathbf{x}|_1 \vee 1) \qquad \text{and}\\
|\mathbf{x}^{(L)} - \widetilde{\mathbf{x}}^{(L)}|_1 &\leqslant \eta 2r \sum_{k=1}^{L-1} (rB)^{k-1}(|\mathbf{x}^{(L-k)}|_1 \vee 1) + (rB)^{L-1}|\mathbf{x}^{(1)} - \widetilde{\mathbf{x}}^{(1)}|_1\\
&\leqslant \eta 2^{(L+1)} r(|\mathbf{x}|_1 \vee 1)(rB)^{L-1}.
\end{aligned}
$$

Since the clipping function $y \mapsto (-C) \vee (y \wedge C)$ has Lipschitz constant 1, we conclude

$$
\begin{aligned}
|f_\vartheta(\mathbf{x}) - f_{\widetilde{\vartheta}}(\mathbf{x})| &\leqslant |g_\vartheta(\mathbf{x}) - g_{\widetilde{\vartheta}}(\mathbf{x})|\\
&= |\mathbf{x}^{(L+1)} - \widetilde{\mathbf{x}}^{(L+1)}|\\
&= |W^{(L+1)}\mathbf{x}^{(L)} + v^{(L+1)} - \widetilde{W}^{(L+1)}\widetilde{\mathbf{x}}^{(L)} - \widetilde{v}^{(L+1)}|\\
&\leqslant |(W^{(L+1)} - \widetilde{W}^{(L+1)})\mathbf{x}^{(L)}| + |\widetilde{W}^{(L+1)}(\mathbf{x}^{(L)} - \widetilde{\mathbf{x}}^{(L)})| + |v^{(L+1)} - \widetilde{v}^{(L+1)}|\\
&\leqslant r|W^{(L+1)} - \widetilde{W}^{(L+1)}|_\infty |\mathbf{x}^{(L)}|_1 + r|\widetilde{W}^{(L+1)}|_\infty |\mathbf{x}^{(L)} - \widetilde{\mathbf{x}}^{(L)}|_1 + \eta
\end{aligned}
$$

$$\leqslant \eta r(2rB)^L(|\mathbf{x}|_1 \vee 1) + \eta(rB)^L 2^{L+1}(|\mathbf{x}|_1 \vee 1) + \eta$$
$$\leqslant \eta 4(2rB)^L(|\mathbf{x}|_1 \vee 1). \qquad \square$$

3.4.7.2 Proof of Lemma 3.15

Since ϱ_η and Π are product measures, their KL-divergence is equal to the sum of the KL-divergences in each of the P factors. For each such factor, we compare

$$\mathcal{U}([(\vartheta^*)_i - \eta, (\vartheta^*)_i + \eta] \cap [-B, B]) \qquad \text{with} \qquad \mathcal{U}([-B, B]),$$

where $(\vartheta^*)_i$ denotes the i-th entry of ϑ^*. The KL-divergence of these distributions is equal to

$$\log\Big(\frac{\lambda([-B,B])}{\lambda([(\vartheta^*)_i - \eta, (\vartheta^*)_i + \eta] \cap [-B,B])}\Big) \leqslant \log\Big(\frac{\lambda([-B,B])}{\lambda([0,\eta])}\Big) = \log(2B/\eta).$$

Thus,

$$\mathrm{KL}(\varrho_\eta \mid \Pi) = \sum_{i=1}^{P} \mathrm{KL}\left(\mathcal{U}([(\vartheta^*)_i - \eta, (\vartheta^*)_i + \eta] \cap [-B,B]) \,\middle|\, \mathcal{U}([-B,B])\right) \leqslant P\log(2B/\eta). \qquad \square$$

4 Stochastic neural networks with mixing priors

In this chapter, we combine our insights from the multi-index analysis and with the stochastic neural networks from the previous chapter. In Chapter 2, we have seen that an estimator based on the Gibbs posterior with respect to a mixing prior can adapt to various structural properties of multi-index models. In Chapter 3, we have demonstrated that a stochastic neural network based on the Gibbs posterior with respect to a uniform prior allows us to estimate hierarchical regression functions with the minimax-optimal rate (up to a logarithmic factor). Now, we will merge these ideas by considering stochastic neural networks based on the Gibbs posterior with respect to mixing priors. In Section 4.1, we construct a prior which allows the Gibbs posterior to adaptively choose the optimal network architecture. In Section 4.2, we demonstrate how we can use the concept of sparsity as in Section 2.2 to apply stochastic neural networks to high-dimensional regression problems.

While these methods have strong theoretical properties, an efficient implementation is challenging and left for future research.

These extensions are based on Bieringer et al. (2023, Section 4) and Steffen & Trabs (2023).

Throughout this chapter, we again consider the regression setting from the introduction of Chapter 2.

4.1 Learning the width

To balance the approximation error term and the stochastic error term in (3.15), we have to choose an optimal network width. In this section we present a fully data-driven approach to this hyperparameter optimization problem which avoids evaluating competing network architectures on a validation set. To account for the model selection problem, we augment the approach with a mixing prior, which prefers narrower neural networks. Equivalently, this approach can

be understood as a hierarchical Bayes method where we put a geometric distribution on the hyperparameter r.

We set

$$\breve{\Pi} = \sum_{r=1}^{n} 2^{-r} \Pi_r \Big/ (1 - 2^{-n}),$$

where $\Pi_r = \mathcal{U}([-B, B]^{P_r})$ with $P_r \coloneqq (p+1)r + (L-1)(r+1)r + r + 1$. As in Section 2.2, the basis 2 of the geometric weights is arbitrary. It can be replaced by a larger constant to assign even less weight to wide networks, but the theoretical results remain the same up to constants.

We obtain our adaptive estimator $\breve{f}_{\lambda,\rho}$ by drawing a parameter ϑ from the surrogate-posterior distribution with respect to this prior, i.e.

$$\breve{f}_{\lambda,\rho} \coloneqq f_{\breve{\vartheta}_{\lambda,\rho}} \quad \text{for} \quad \breve{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n \sim \breve{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \quad \text{with} \quad \breve{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) \propto \exp\big(-\lambda \widetilde{R}_{n,\rho}(\vartheta)\big) \breve{\Pi}(\mathrm{d}\vartheta)$$

where $\widetilde{R}_{n,\rho}$ is from (3.9). It should be noted that our results extend to the estimator based on the posterior mean as in Chapter 3 and are only omitted for the sake of conciseness.

This modification of the prior allows the estimator to adapt to the optimal network width and we can compare its performance with that of the network corresponding the oracle choice of the parameter

$$\vartheta_r^* \in \operatorname*{arg\,min}_{\vartheta \in [-B,B]^{P_r}} R(\vartheta) \tag{4.1}$$

given any width r. We obtain the following adaptive version of Theorem 3.3:

Theorem 4.1 (Width-adaptive oracle inequality)**.** *Under Assumption 3.A there is a constant $Q_5 > 0$ depending only on C, Γ, σ such that for $\lambda = n/Q_0$ (with Q_0 from Theorem 3.3) and sufficiently large n we have for all $\delta \in (0,1)$ with probability of at least $1 - \delta$ that*

$$\mathcal{E}(\breve{f}_{\lambda,\rho}) \leqslant \min_{r=1,\dots,n} \Big(12\mathcal{E}(f_{\vartheta_r^*}) + \frac{Q_5}{n}\big(P_r L \log(n) + \log(2/\delta)\big)\Big).$$

Since the modified estimator mimics the performance of the optimal network choice regardless of width, we obtain the following width-adaptive version of Proposition 3.7 with no additional loss in the convergence rate:

Corollary 4.2 (Width-adaptive rates of convergence)**.** *Let $\mathbf{X} \in [0,1]^p$. In the situation of Theorem 4.1, there exists a network depth $L = C_3 \log n$ with $C_3 > 0$ only depending on upper bounds for $q, |\mathbf{d}|_\infty, |\beta|_\infty, C_0$ such that the estimator $\breve{f}_{\lambda,\rho}$ satisfies for sufficiently large n uniformly*

over all hierarchical functions $f \in \mathcal{H}(q, \mathbf{d}, \mathbf{t}, \beta, C_0)$

$$\mathcal{E}(\check{f}_{\lambda,\rho}) \leqslant Q_6\Big(\frac{(\log n)^3}{n}\Big)^{2\beta^*/(2\beta^*+t^*)} + Q_6\frac{\log(2/\delta)}{n}$$

with probability of at least $1-\delta$*, where* β^* *and* t^* *are as in Proposition 3.7. The constant* Q_6 *only depends on upper bounds for* $q, |\mathbf{d}|_\infty, |\beta|_\infty$ *and* C_0 *as well as the constants* C, Γ, σ *from Assumption 3.A.*

It has to be noted that we cannot hope to construct credible sets with coverage as in Theorem 3.10 based on the adaptive posterior distribution. It is well known that adaptive honest confidence sets are only possible under additional assumptions on the regularity of the regression function, e.g. self-similarity or polished tail conditions, see Hoffmann & Nickl (2011) and we remark that such conditions with respect to the network parameterization seem infeasible.

4.2 High-dimensional regression using sparse neural networks

While the rate in Proposition 3.7 only depends on the intrinsic dimension of the hierarchical regression function, the constants Q_3, Q_4 depend on the dimension p of the data. To circumvent this, we consider sparse neural networks.

A network is sparse, or more precisely *connection sparse*, if many weights in the network are zero and thus some links between nodes are inactive. For some active set $\mathcal{I} \subset \{1, \dots, P\}$, the corresponding class of sparse networks is defined by

$$\mathcal{G}(p, L, r, \mathcal{I}) \coloneqq \{g_\vartheta \in \mathcal{G}(p, L, r) : \vartheta_i = 0 \text{ if } i \notin \mathcal{I}\}.$$

We denote by $|\mathcal{I}|$ the cardinality of a set $\mathcal{I}$, which is useful for quantifying the sparsity of a neural network in $\mathcal{G}(p, L, r, \mathcal{I})$.

For some $C \geqslant 1$, we also introduce the class of clipped networks

$$\mathcal{F}(p, L, r, C) \coloneqq \big\{ f_\vartheta \coloneqq (-C) \vee (g_\vartheta \wedge C) \,\big|\, g_\vartheta \in \mathcal{G}(p, L, r)\big\}$$

and similarly we denote clipped networks with active set $\mathcal{I}$ by $\mathcal{F}(p, L, r, \mathcal{I}, C)$.

In order to adopt csMALA to sparse neural networks, we again modify the prior. For a given active set $\mathcal{I}$ the prior $\Pi_\mathcal{I}$ on the parameter set of the class $\mathcal{G}(p, L, r, \mathcal{I})$ is defined as the uniform

distribution on

$$\mathcal{S}_{\mathcal{I}} \coloneqq \{\vartheta \in [-B, B]^P \mid \vartheta_i = 0 \text{ if } i \notin \mathcal{I}\} \tag{4.2}$$

for some $B \geqslant 1$. To allow for a data-driven choice of the active set, we define the prior $\mathring{\Pi}$ as a mixture of the uniform priors $\Pi_{\mathcal{I}}$:

$$\mathring{\Pi} \coloneqq \sum_{i=1}^{P} 2^{-i} \sum_{\substack{\mathcal{I} \subseteq \{1,\dots,P\}, \\ |\mathcal{I}|=i}} \binom{P}{i}^{-1} \Pi_{\mathcal{I}} \Big/ C_P \qquad \text{with} \qquad C_P \coloneqq (1 - 2^{-P}).$$

Similarly to Section 2.2, the basis 2 of the geometric weights is arbitrary and can be replaced by a larger constant leading to a stronger preference of sparse networks. The theoretical results remain unchanged up to constants. The prior $\mathring{\Pi}$ can be understood as a hierarchical prior, where we first draw a geometrically distributed sparsity i, given i we uniformly choose an active set $\mathcal{I} \subset \{1, \dots, P\}$ with $|\mathcal{I}| = i$ and on $\mathcal{I}$ the uniform prior $\Pi_{\mathcal{I}}$ is applied. We define our estimator via

$$\mathring{f}_{\lambda,\rho} \coloneqq f_{\mathring{\vartheta}_{\lambda,\rho}} \quad \text{for} \quad \mathring{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n \sim \mathring{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \quad \text{with} \quad \mathring{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) \propto \exp\big(-\lambda \widetilde{R}_{n,\rho}(\vartheta)\big)\mathring{\Pi}(\mathrm{d}\vartheta). \tag{4.3}$$

As a benchmark for the performance of the method, we define the *oracle choice* on $\mathcal{S}_{\mathcal{I}}$ from (4.2) for some active set $\mathcal{I}$ as

$$\vartheta^*_{\mathcal{I}} \in \operatorname*{arg\,min}_{\vartheta \in \mathcal{S}_{\mathcal{I}}} R(\vartheta). \tag{4.4}$$

The oracle is not accessible to the practitioner because $R(\vartheta)$ depends on the unknown distribution of $(\mathbf{X}, Y)$. A solution to the minimization problem in (4.4) always exists since $\mathcal{S}_{\mathcal{I}}$ is compact and $\vartheta \mapsto R(f_\vartheta)$ is continuous. If there is more than one solution, we choose one of them. Our main result gives a theoretical guarantee that the PAC-Bayes estimator $\mathring{f}_{\lambda,\rho}$ from (4.3) is almost as good as the oracle $\vartheta^*_{\mathcal{I}}$ in terms of the excess risk.

We obtain the following non-asymptotic oracle inequality:

Theorem 4.3 (PAC-Bayes oracle inequality)**.** *Under Assumption 3.A there is a constant $Q_7 > 0$ only depending on C, Γ, σ such that for $\lambda = n/Q_0$ (with Q_0 from Theorem 3.5) and sufficiently large n, we have with probability of at least $1 - \delta$ that*

$$\mathcal{E}(\mathring{f}_{\lambda,\rho}) \leqslant \min_{\mathcal{I}} \Big(12\mathcal{E}(f_{\vartheta^*_{\mathcal{I}}}) + \frac{Q_7}{n}\big(|\mathcal{I}| L \log(p \vee n) + \log(2/\delta)\big)\Big).$$

Remark 4.4. The dependence of Q_7 on C, Γ, σ is at most quadratic and $n \geqslant n_0 \coloneqq 2 \vee r \vee b \vee K$ is sufficiently large.

The benefit of this theorem over Theorem 3.3 is that the terms in the upper bound only depend on the number of nonzero weights of the oracle and that we take the minimum over all possible sets of nonzero weights. In particular, we recover $\frac{|\mathcal{I}|}{n} \log p$ as the typical error term for estimating high-dimensional vectors with sparsity $|\mathcal{I}|$ similarly to Theorem 2.5.

Theorem 4.3 is in line with classical PAC-Bayes oracle inequalities, see Alquier (2021). Chérief-Abdellatif (2020) has obtained a similar oracle inequality for a variational approximation of the Gibbs posterior distribution, but without the minimum over the active sets. For penalized empirical risk minimization, Taheri et al. (2021) have obtained another oracle inequality with a different dependence on the depth L.

Corollary 4.5 (Rates of convergence)**.** *Let* $\log p \leqslant n/(\log^2 n)$ *and* $\mathbf{X} \in [0,1]^p$*. In the situation of Theorem 4.3, there exists a network architecture* $(L, r) = (C_1 \lceil \log_2 n \rceil, C_2 n)$ *with* C_1 *and* C_2 *only depending on upper bounds for* $q, |(d_1, \dots, d_q)|_\infty, |\mathbf{t}|_\infty, |\beta|_\infty$ *and* C_0 *such that for sufficiently large* n*, the estimator* $\mathring{f}_{\lambda,\rho}$ *yields an excess risk of at most*

$$\mathcal{E}(\mathring{f}_{\lambda,\rho}) \leqslant Q_8 \Big(\frac{\log(p \vee n) \log^2(n)}{n} \Big)^{2\beta^*/(2\beta^*+t^*)} + Q_8 \frac{\log(2/\delta)}{n}$$

with probability of at least $1 - \delta$*, where* β^* *and* t^* *are given by*

$$\beta^* \coloneqq \beta^*_{i^*}, \qquad t^* \coloneqq t^*_{i^*} \qquad \text{for} \qquad i^* \in \operatorname*{arg\,min}_{i=0,\dots,q} \frac{2\beta^*_i}{2\beta^*_i + t^*_i} \qquad \text{and} \qquad \beta^*_i \coloneqq \beta_i \prod_{l=i+1}^{q} (\beta_l \wedge 1).$$

The constant Q_8 *only depends on* $q, (d_1, \dots, d_q), \mathbf{t}, \beta$ *and* C_0 *and* C, Γ, σ*.*

Since the class of multi-index models can be embedded into the class of hierarchical functions, we recover the same rate of convergence (up to a logarithmic factor) when applying $\mathring{f}_{\lambda,\rho}$ to data generated from a multi-index model, cf. Corollary 2.8.

Corollary 4.6 (Sparse neural networks applied to multi-index models)**.** *Grant Assumption 2.B and Assumption 2.C with* $g^* \in \mathcal{C}^\beta_{d^*}([-B_1, B_1]^p, C_0)$*. In the situation of Theorem 4.3, there exists a network architecture* $(L, r) = (C_1' \lceil \log_2 n \rceil, C_2' n)$ *with* C_1' *and* C_2' *only depending on upper bounds for* d^*, β *and* C_0 *such that the estimator* $\mathring{f}_{\lambda,\rho}$ *yields an excess risk for sufficiently large* n *uniformly over all hierarchical functions* $f \in \mathcal{H}(q, \mathbf{d}, \mathbf{t}, \beta, C_0)$ *of at most*

$$\mathcal{E}(\mathring{f}_{\lambda,\rho}) \leqslant Q_9 \Big(\frac{\log(p \vee n) \log^2(n)}{n} \Big)^{2\beta/(2\beta+d^*)} + Q_9 \frac{\log(2/\delta)}{n}$$

with a probability of at least $1 - \delta$*.*

4.3 Proofs

The proofs of Theorem 4.1 and Theorem 4.3 are structurally similar to that of Theorem 3.3. Note that the only property of the prior that we used in the proof of Proposition 3.13 is that Π is a probability measure on the space of network weights. Hence, it is straightforward to see that the analogous statement still holds when replacing Π with $\breve{\Pi}$ and $\mathring{\Pi}$, respectively. To account for the change in the prior, we also choose a different probability measure ϱ in the resulting upper bound and modify Lemma 3.15 accordingly.

4.3.1 Proof of Theorem 4.1

As outlined in the previous paragraph, we have for any sample-dependent probability measure $\varrho \ll \breve{\Pi}$ with probability of at least $1 - \delta$

$$\mathcal{E}(\breve{\vartheta}_{\lambda,\rho}) \leqslant 9 \int \mathcal{E} \, \mathrm{d}\varrho + \frac{16}{\lambda}\big(\mathrm{KL}(\varrho \mid \breve{\Pi}) + \log(2/\delta)\big). \tag{4.5}$$

For a width $r \in \mathbb{N}$ and some radius $\eta \in (0, 1]$, we now choose $\varrho = \varrho_{r,\eta}$ defined via

$$\frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\Pi_r}(\vartheta) \propto \mathbb{1}_{\{|\vartheta - \vartheta_r^*|_\infty \leqslant \eta\}}$$

with ϑ_r^* from (4.1). Replacing ϑ^* with ϑ_r^* in the arguments from the proof of Theorem 3.3, we find

$$\int \mathcal{E} \, \mathrm{d}\varrho_{r,\eta} \leqslant \frac{4}{3}\mathcal{E}(\vartheta_r^*) + \frac{4}{n^2} \qquad \text{for} \qquad \eta = \frac{1}{8K(2rB)^L p \, n}.$$

To bound the Kullback-Leibler term in (4.5), we employ the following modification of Lemma 3.15:

Lemma 4.7. *We have* $\mathrm{KL}(\varrho_{r,\eta} \mid \breve{\Pi}) \leqslant P_r \log(2B/\eta) + r$.

Therefore, we have with probability of at least $1 - \delta$ that

$$\mathcal{E}(\breve{\vartheta}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta_r^*}) + \frac{Q_5}{n}\big(P_r L \log(n) + \log(2/\delta)\big),$$

for some $Q_5 > 0$ only depending on C, Γ, σ. Choosing r to minimize the upper bound in the last display yields the assertion. □

4.3.2 Proof of Theorem 4.3

The arguments analogous to the proof of Theorem 4.1 yield with probability of at least $1-\delta$

$$\mathcal{E}(\mathring{\vartheta}_{\lambda,\rho}) \leqslant 9 \int \mathcal{E} \, \mathrm{d}\varrho + \frac{16}{\lambda}\big(\mathrm{KL}(\varrho \mid \mathring{\Pi}) + \log(2/\delta)\big). \tag{4.6}$$

Now, for some fixed index set $\mathcal{I}$, a radius $\eta \in (0,1]$ and $\varrho = \varrho_{\mathcal{I},\eta}$ defined via

$$\frac{\mathrm{d}\varrho_{\mathcal{I},\eta}}{\mathrm{d}\mathring{\Pi}}(\vartheta) \propto \mathbb{1}_{\{|\vartheta - \vartheta^*_{\mathcal{I}}|_\infty \leqslant \eta\}}$$

with $\vartheta^*_{\mathcal{I}}$ from (4.4), we deduce

$$\int \mathcal{E} \, \mathrm{d}\varrho_{\mathcal{I},\eta} \leqslant \frac{4}{3}\mathcal{E}(\vartheta^*_{\mathcal{I}}) + \frac{4}{n^2} \qquad \text{for} \qquad \eta = \frac{1}{8K(2rB)^L p\, n}.$$

The Kullback-Leibler term in (4.6) is controlled with the following further modification of Lemma 3.15:

Lemma 4.8. *We have*

(i) $\mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi) = \mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi_{\mathcal{I}}) + \log(C_{\mathcal{I}})$ *where* $C_{\mathcal{I}} \coloneqq C_P 2^{|\mathcal{I}|}\binom{P}{|\mathcal{I}|}$,

(ii) $\mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi_{\mathcal{I}}) \leqslant |\mathcal{I}| \log(2B/\eta)$.

In particular,

$$\mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi) \leqslant |\mathcal{I}| \log(2B/\eta) + \log(C_{\mathcal{I}}).$$

Together with $\binom{P}{|\mathcal{I}|} \leqslant \frac{P^{|\mathcal{I}|}}{(|\mathcal{I}|)!} \leqslant (\mathrm{e}P)^{|\mathcal{I}|}$, the previous lemma yields

$$\mathrm{KL}(\rho \mid \Pi) \leqslant |\mathcal{I}| \log\big(4C_P BP\mathrm{e}/\eta\big). \tag{4.7}$$

Plugging (4.6) and (4.7) into the PAC-Bayes bound (4.6), we conclude

$$\begin{aligned}\mathcal{E}(\mathring{\vartheta}_{\lambda,\rho}) &\leqslant 12\mathcal{E}(\vartheta^*_{\mathcal{I}}) + \frac{4}{n^2} + \frac{4}{\lambda}\big(|\mathcal{I}| \log\big(32BP\mathrm{e}(2rB)^L pKn\big) + \log(2/\delta)\big) \\ &\leqslant 12\mathcal{E}(\vartheta^*_{\mathcal{I}}) + \frac{Q_7}{n}\big(|\mathcal{I}|L\log(p \vee n) + \log(2/\delta)\big) \end{aligned} \tag{4.8}$$

for $n \geqslant n_0 \coloneqq 2 \vee r \vee B \vee K$ and some constant Q_7 only depending on C, Γ, σ. Note that the upper bound in (4.8) is deterministic and $\mathcal{I}$ is arbitrary. Therefore, we can choose $\mathcal{I}$ such that this bound is minimized, which completes the proof. □

4.3.3 Proof of Corollary 4.2

The statement follows by choosing r in the upper bound from Theorem 4.1 as in the statement of Proposition 3.7 and then using the same approximation result to control the excess-risk of the corresponding oracle choice ϑ_r^*. □

4.3.4 Proof of Corollary 4.5

Throughout, denote by $C_i, i = 1, 2, \dots$ constants only depending on upper bounds for q, $|(d_1, \dots, d_q)|_\infty$, $|\mathbf{t}|_\infty$, $|\beta|_\infty$ and C_0.

We will verify that for any sufficiently large $n, M \in \mathbb{N}$ there exists a sparse ReLU neural network $g = g_\vartheta \in \mathcal{G}(p, C_1\lceil \log_2 n \rceil, C_2 M, \mathcal{I})$ with $|\mathcal{I}| \leqslant C_3 M \lceil \log_2 n \rceil$ and $|\vartheta|_\infty \leqslant 1 \leqslant B$ such that

$$\|g - f\|_{L^\infty([0,1]^p)} \leqslant C_4 M^{-\beta^*/t^*}. \tag{4.9}$$

Careful inspection of the proof of Schmidt-Hieber (2020, Theorem 1) reveals that there exists a sparse ReLU neural network $g \in \mathcal{G}(p, L, r, \mathcal{J})$ with weights and shifts absolutely bounded by 1 and

$$\begin{aligned} L &= 3(q-1) + \sum_{i=0}^{q} \big(8 + (\lceil \log_2 n \rceil + 5)(1 + \lceil \log_2(t_i \vee \beta_i) \rceil)\big), \\ r &= 6M \max_{i=0,\dots,q} d_{i+1}(t_i + \lceil \beta_i \rceil) \qquad \text{and} \\ |\mathcal{J}| &\leqslant \sum_{i=0}^{q} d_{i+1}\big(141(t_i + \beta_i + 1)^{3+t_i} M(\lceil \log_2 n \rceil + 6) + 4\big) \end{aligned}$$

such that (4.9) holds with

$$M = \Big\lceil \Big(\frac{n}{\log^2(n) \log(p \vee n)} \Big)^{t^*/(2\beta^* + t^*)} \Big\rceil \leqslant n,$$

provided $M \geqslant \max_{i=0,\dots,q} (\beta_i + 1)^{t_i} \vee (C_0(2C_0)^{\beta_i} + 1)\mathrm{e}^{t_i}$. Hence, it remains to show that g can also be represented as a ReLU neural network in

$$\mathcal{G}(p, C_1\lceil \log_2 n \rceil, C_2 n, \mathcal{I}). \tag{4.10}$$

To do this, we employ the *embedding properties of network function classes* from Schmidt-Hieber (2020, Section 7.1).

Note that L, r and the upper bound for $|\mathcal{J}|$ are independent of $d_0 = p$ and monotonically increasing in q, $|(d_1, \dots, d_q)|_\infty$, $|\mathbf{t}|_\infty$ and $|\beta|_\infty$. Also, r is of order M, L is of order $\lceil \log_2 n \rceil$ and the upper bound for $|\mathcal{J}|$ is of order $M\lceil \log_2 n \rceil$. Using the *enlarging* and the *depth synchronization properties*, g can indeed be written as a ReLU neural network in (4.10). Note that to ensure the depth of the network, we added additional layers after the last hidden layer, instead of right after the input to preserve the order of the sparsity.

Theorem 4.3 together with $\mathcal{E}(f_{\vartheta^*_{\mathcal{I}}}) \leqslant \|g_\vartheta - f\|^2_{L^\infty([0,1]^p)}$ now yields

$$\mathcal{E}(\widehat{\vartheta}_\lambda) \leqslant 4C_4 M^{-2\beta^*/t^*} + \frac{Q_7 C_3}{n} M \lceil \log_2 n \rceil^2 \log(p \vee n) + Q_7 \frac{\log(2/\delta)}{n}$$

with probability of at least $1 - \delta$. □

4.3.5 Proof of Corollary 4.6

The idea of the proof is similar to that of Corollary 4.5: we construct a sparse neural network which approximates the true regression function $f = g^*(W^*\cdot)$. We treat g^* and W^* separately. $\mathbf{x} \mapsto W^*\mathbf{x}$ can be replicated exactly by $W^*\mathbf{x} = W^{(2)}\phi(W^{(1)}\mathbf{x})$ with $W^{(1)} = (W^*, -W^*)^\top$, $W^{(2)} = (E_{d^*}, -E_{d^*})$. The corresponding network has $L_{W^*} = 1$ hidden layer, a width of $r_{W^*} = d^*$ and sparsity level $|\mathcal{J}| = 2\|W^*\|_0 + 2d^* \leqslant 4\|W^*\|_0$. Since the rows of W^* are ℓ^2-standardized, all weights are absolutely bounded by 1.

To approximate the link function, we write $g^*(W^*\mathbf{x}) = g(W^*\mathbf{x}/(2B_1) + v)$ for $v = (1/2, \dots, 1/2)^\top \in \mathbb{R}^{d^*}$ and $g = g^*(2B_1(\cdot - v))$. As in the proof of Corollary 4.5, there exists for any sufficiently large $n, M \in \mathbb{N}$ a sparse ReLU neural network g_ϑ with $|\vartheta|_\infty \leqslant 1$, $L_{g^*} \lesssim \log_2 n$ hidden layers, width $r_{g^*} \lesssim M$ and sparsity $\mathcal{J}_{g^*} \lesssim \log_2 n$ such that

$$\|g - g_\vartheta\|_{L^\infty([0,1]^{d^*})} \leqslant cM^{-\beta/d^*}$$

for some $c > 0$ only depending on upper bounds for d^*, β and C_0. Using the *embedding properties of network function classes* from Schmidt-Hieber (2020, Section 7.1), we can also write $g_\vartheta(2B_1(\cdot - v))$ as a ReLU neural network $g_{\widetilde{\vartheta}}$ with architecture and sparsity of the same order as for g_ϑ and with $|\widetilde{\vartheta}|_\infty \leqslant B_1$. The composition of $g_{\widetilde{\vartheta}}$ with the network which replicates $\mathbf{x} \mapsto W^*\mathbf{x}$ is a ReLU neural network with architecture and sparsity of the same order as for $g_{\widetilde{\vartheta}}$. Its weights are absolutely bounded by B_1. Since $|W^*\mathbf{X}|_\infty \leqslant B_1$, we obtain

$$\mathcal{E}(f_{\vartheta^*_{\mathcal{I}}}) \leqslant \|g^* - g_{\widetilde{\vartheta}}\|^2_{L^\infty([-B_1,B_1]^{d^*})} = \|g - g_\vartheta\|^2_{L^\infty([0,1]^{d^*})} \leqslant cM^{-2\beta/d^*}.$$

Using Theorem 4.3 yields for some $c > 0$ only depending on upper bounds for d^*, β and C_0

that

$$\mathcal{E}(\widehat{\vartheta}_\lambda) \leqslant 4c' M^{-2\beta/d^*} + \frac{c' Q_7}{n} M \lceil \log_2 n \rceil^2 \log(p \vee n) + Q_7 \frac{\log(2/\delta)}{n}$$

with a probability of at least $1 - \delta$. Choosing

$$M = \Big\lceil \Big(\frac{n}{\log^2(n) \log(p \vee n)} \Big)^{d^*/(2\beta + d^*)} \Big\rceil \leqslant n$$

yields the assertion. □

4.3.6 Proofs of the auxiliary results

The proofs of the auxiliary results in this section use similar arguments to the proofs of Lemma 2.12 and Lemma 3.15, see Section 2.3.5.3 and Section 3.4.7.2, respectively.

4.3.6.1 Proof of Lemma 4.7

The key step is to verify

$$\frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\breve{\Pi}} = 2^r (1 - 2^{-n}) \frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\Pi_r}, \tag{4.11}$$

from which we can deduce

$$\begin{aligned} \mathrm{KL}(\varrho_{r,\eta} \mid \breve{\Pi}) = \int \log\Big(\frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\breve{\Pi}}\Big) \,\mathrm{d}\varrho_{r,\eta} &= \int \log\Big(\frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\Pi_r}\Big) \,\mathrm{d}\varrho_{r,\eta} + \log(2^r(1 - 2^{-n})) \\ &\leqslant \mathrm{KL}(\varrho_{L,\eta} \mid \Pi_L) + r. \end{aligned}$$

Since the arguments from the proof of Lemma 3.15, yield $\mathrm{KL}(\varrho_{r,\eta} \mid \Pi_r) \leqslant P_r \log(2B/\eta)$, the lemma follows.

For (4.11), note how $\varrho_{r,\eta}$ only assigns positive probability to sets $A \subseteq [-B, B]^{P_r}$. Hence,

$$\varrho_{r,\eta}(A) = \int_A \frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\breve{\Pi}} \,\mathrm{d}\breve{\Pi} = (1 - 2^{-n})^{-1} \sum_{l=1}^{n} 2^{-l} \int_A \frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\breve{\Pi}} \,\mathrm{d}\Pi_l = (1 - 2^{-n})^{-1} 2^{-r} \int_A \frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\breve{\Pi}} \,\mathrm{d}\Pi_r. \quad \square$$

4.3.6.2 Proof of Lemma 4.8

(i) We will show that

$$\frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi} = C_{\mathcal{I}} \frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi_{\mathcal{I}}} \tag{4.12}$$

to deduce

$$\mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi) = \int \log\Big(\frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi}\Big)\,\mathrm{d}\rho_{\mathcal{I},\eta} = \int \log\Big(\frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi_{\mathcal{I}}}\Big)\,\mathrm{d}\rho_{\mathcal{I},\eta} + \log(C_{\mathcal{I}}) = \mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi_{\mathcal{I}}) + \log(C_{\mathcal{I}}).$$

For (4.12), we need to check

$$\rho_{\mathcal{I},\eta}(A) = \int_A C_{\mathcal{I}}^{-1} \frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi}\,\mathrm{d}\Pi_{\mathcal{I}}$$

for all Borel-measurable set $A \subseteq \mathbb{R}^p$. Observe that for the sets $\mathcal{S}_{\mathcal{J},\Leftrightarrow} \coloneqq \{\vartheta \in \mathcal{S}_{\mathcal{J}} \mid \vartheta_i \neq 0 \Leftrightarrow i \in \mathcal{J}\}$ with $\emptyset \neq \mathcal{J} \subseteq \{1,\dots,P\}$, we have

$$\Pi_{\mathcal{J}}(\mathcal{S}_{\mathcal{J},\Leftrightarrow}) = 1. \tag{4.13}$$

In particular, (4.13) holds for $\mathcal{J} = \mathcal{I}$. Since also $\rho_{\mathcal{I},\eta}(\mathcal{S}_{\mathcal{I},\Leftrightarrow}) = 1$, no generality is lost in additionally assuming $A \subseteq \mathcal{S}_{\mathcal{I},\Leftrightarrow}$. Note how

$$\mathcal{S}_{\mathcal{J},\Leftrightarrow} \cap \mathcal{S}_{\mathcal{I},\Leftrightarrow} = \emptyset \qquad \forall \mathcal{J} \neq \mathcal{I}. \tag{4.14}$$

Combining (4.13) with (4.14), we see that

$$\int_A \frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi}\,\mathrm{d}\Pi_{\mathcal{J}} = 0 \qquad \forall \mathcal{J} \neq \mathcal{I}.$$

Therefore,

$$\rho_{\mathcal{I},\eta}(A) = \int_A \frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi}\,\mathrm{d}\Pi = \int_A C_{\mathcal{I}}^{-1} \frac{\mathrm{d}\rho_{\mathcal{I},\eta}}{\mathrm{d}\Pi}\,\mathrm{d}\Pi_{\mathcal{I}}.$$

(ii) $\rho_{\mathcal{I},\eta}$ and $\Pi_{\mathcal{I}}$ are product measures and thus their KL-divergence is equal to the sum of the KL-divergences in each of the P factors. For terms with index $i \in \mathcal{I}$, the corresponding KL-divergence can be bounded by $\log(2B/\eta)$ as in the proof of Lemma 3.15. For terms with index $i \notin \mathcal{I}$, the corresponding KL-divergence is zero, as both factors have all their mass in 0. Thus,

$$\mathrm{KL}(\rho_{\mathcal{I},\eta} \mid \Pi_{\mathcal{I}}) = \sum_{i \in \mathcal{I}} \mathrm{KL}\left(\mathcal{U}([(\vartheta^*_{\mathcal{I}})_i - \eta, (\vartheta^*_{\mathcal{I}})_i + \eta] \cap [-B,B]) \mid \mathcal{U}([-B,B])\right) \leqslant |\mathcal{I}| \log(2B/\eta). \qquad \square$$

5 Estimating a multivariate Lévy density

In this chapter, we present our results on the estimation of a multivariate Lévy density based on Steffen (2023a). The estimator is based on the spectral method, see Belomestny & Reiß (2015) for an introduction and Trabs (2015) for the one-dimensional case.

Let us recall that a Lévy process $(L_t)_{t\geqslant 0}$ is an $\mathbb{R}^d$-valued stochastically continuous stochastic process in continuous time with independent and stationary increments, (a.s.) càdlàg paths and $L_0 = 0$. The distribution of the entire process is characterized by its Lévy triplet (Σ, γ, ν) with a drift parameter $\gamma \in \mathbb{R}^d$, a positive semi-definite volatility matrix $\Sigma \in \mathbb{R}^{d\times d}$ and a Lévy measure ν. The aim of this chapter is to estimate the density of the Lévy measure, called the Lévy density, which we also denote by ν assuming that it exists. Throughout, the Lévy process is observed in the form of $n \in \mathbb{N}$ increments at equidistant time points with time difference $\delta > 0$ and overall time horizon $T \coloneqq n\delta$:

$$Y_k \coloneqq L_{\delta k} - L_{\delta(k-1)}, \qquad k = 1, \dots, n.$$

Since, by definition, $(L_t)_{t\geqslant 0}$ has independent and stationary increments, $Y_1, \dots Y_n$ are i.i.d.

The Lévy-Khinchine formula, see e.g. Sato (1999), allows for an explicit representation of the characteristic function of the process at any time point. If $\int |x|^2 \, \nu(\mathrm{d}x) < \infty$, this representation reads

$$\varphi_t(u) \coloneqq \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle u, L_t\rangle}] = \mathrm{e}^{t\psi(u)} \quad \text{with} \quad \psi(u) \coloneqq \mathrm{i}\langle\gamma, u\rangle - \frac{1}{2}\langle u, \Sigma u\rangle + \int \left(\mathrm{e}^{\mathrm{i}\langle u,x\rangle} - 1 - \mathrm{i}\langle u, x\rangle\right) \nu(\mathrm{d}x). \tag{5.1}$$

Denote the gradient and the Laplacian of a function $g\colon \mathbb{R}^d \to \mathbb{C}$ by ∇g and Δg, respectively, assuming they exist. We have

$$\nabla\psi(u) = \mathrm{i}\gamma - \Sigma u + \mathrm{i}\int x\left(\mathrm{e}^{i\langle u,x\rangle} - 1\right) \nu(\mathrm{d}x), \tag{5.2}$$

$$\Delta\psi(u) = -\operatorname{tr}(\Sigma) - \int |x|^2 \mathrm{e}^{\mathrm{i}\langle u,x\rangle} \, \nu(\mathrm{d}x) = -\operatorname{tr}(\Sigma) - \mathscr{F}[|x|^2\nu](u) = \frac{\varphi_t(u)\Delta\varphi_t(u) - (\nabla\varphi_t(u))^2}{t\varphi_t^2(u)}, \tag{5.3}$$

where the integral in the first line is component-wise, $\mathscr{F}[|x|^2\nu] \coloneqq \int \mathrm{e}^{\mathrm{i}\langle\cdot,x\rangle}|x|^2\,\nu(\mathrm{d}x)$ and for vectors $x, y \in \mathbb{C}^d$, we set $x \cdot y = \sum_{k=1}^d x_k y_k$ and then $x^2 = x \cdot x$. In particular, $(\nabla\varphi_t(u))^2 = \sum_{k=1}^d \big(\frac{\partial\varphi_t}{\partial u_k}(u)\big)^2$.

The chapter is organized as follows. In Section 5.1, we introduce the estimation method and state our main results along with a short outline of the proof and the key tools used. The empirical performance of our estimator is illustrated in simulation examples in Section 5.2. The full proofs are postponed to Section 5.3.

5.1 Estimation method and main results

Just to motivate our estimator for ν, suppose $\Sigma = 0$. In view of (5.3), we then have $\nu = -|\cdot|^{-2}\mathscr{F}^{-1}[\Delta\psi]$ and $\Delta\psi$ can naturally be estimated using the empirical characteristic function $\widehat{\varphi}_{\delta,n}(u) \coloneqq \frac{1}{n}\sum_{k=1}^n \mathrm{e}^{\mathrm{i}\langle u, Y_k\rangle}$ leading to

$$\widehat{\Delta\psi_n}(u) \coloneqq \frac{\widehat{\varphi}_{\delta,n}(u)\Delta\widehat{\varphi}_{\delta,n}(u) - (\nabla\widehat{\varphi}_{\delta,n}(u))^2}{\delta\widehat{\varphi}_{\delta,n}(u)}\mathbb{1}_{\{|\widehat{\varphi}_{\delta,n}(u)|\geqslant T^{-1/2}\}} \tag{5.4}$$

with the indicator ensuring a well-defined expression. Therefore, granted ν has a Lévy density also denoted by ν, it is reasonable to propose the estimator

$$\widehat{\nu}_h(x) \coloneqq -|x|^{-2}\mathscr{F}^{-1}\big[\mathscr{F}K_h\widehat{\Delta\psi_n}\big](x), \qquad x \in \mathbb{R}^d \setminus \{0\}, \tag{5.5}$$

where K is bandwidth limited kernel with bandwidth $h > 0$ ($K_h \coloneqq h^{-d}K(\cdot/h)$). We assume that the kernel satisfies, for some order $p \in \mathbb{N}$, that for any multi-index $0 \neq \beta \in \mathbb{N}_0^d$ with $|\beta|_1 \leqslant p$ we have

$$\int_{\mathbb{R}^d} K(x)\,\mathrm{d}x = 1, \qquad \int_{\mathbb{R}^d} x^\beta K(x)\,\mathrm{d}x = 0 \qquad \text{and} \qquad \operatorname{supp}\mathscr{F}K \subseteq [-1,1]^d. \tag{5.6}$$

For $d = 1$, we recover the jump density estimator by Trabs (2015) to estimate quantiles of Lévy measures.

A suitable kernel can be constructed as $K \coloneqq (\mathscr{F}^{-1}g)/g(0)$ from an integrable even function $g\colon C^\infty(\mathbb{R}^d) \to \mathbb{R}$ with support contained in $[-1,1]^d$, $g(0) \neq 0$, and vanishing mixed partial derivatives of order up to p at 0. For the theoretical analysis, it is useful to consider a kernel with product structure $K(x) = \prod_{j=1}^d K^j(x_j)$ for kernels K^j on $\mathbb{R}$, each with order p, i.e. for all $q \in \mathbb{N}$, $q \leqslant p$

$$\int_{\mathbb{R}} K^j(x_j)\,\mathrm{d}x_j = 1, \qquad \int x_j^q K^j(x_j)\,\mathrm{d}x_j = 0, \qquad \text{and} \qquad \operatorname{supp}\mathscr{F}K^j \subseteq [-1,1].$$

Obviously, such a product kernel also fulfills (5.6).

5.1.1 Convergence rates

To control the estimation error, we need to impose smoothness and moment conditions on the Lévy density. To this end, we introduce for a number of moments $m > 0$, a regularity index $s > 0$, an open subset $U \subseteq \mathbb{R}^d$, and a universal constant $C > 0$,

$$\mathcal{J}^s(m, U, C) \coloneqq \Big\{(\Sigma, \gamma, \nu) \,\Big|\, \Sigma \in \mathbb{R}^{d\times d} \text{ positive-semidefinite}, \operatorname{tr}(\Sigma) \leqslant C,\ \gamma \in \mathbb{R}^d, \int |x|^m \nu(\mathrm{d}x) \leqslant C, \nu \text{ has a Lebesgue density with } |\cdot|^2\nu \in \mathcal{C}^s(U, C)\Big\},$$

where $\mathcal{C}^s(U, C)$ denotes the ball of all Hölder regular functions on U with regularity index $s > 0$.

Since we require regularity of the Lévy density in a small ζ-neighborhood beyond U for a uniform rate, we set $U_\zeta \coloneqq \{x \in \mathbb{R}^d \mid \exists u \in U : |x - u| < \zeta\}$ for some radius $\zeta > 0$.

In view of (5.4) it is natural that the estimation error also depends on the decay behavior of the characteristic function, which in turn, is affected by the presence of a Gaussian component. Therefore, we distinguish between the following two classes of Lévy processes. First, is the so-called mildly ill-posed case for a decay exponent $\alpha > 0$

$$\mathcal{D}^s(\alpha, m, U, C, \zeta) \coloneqq \big\{(0, \gamma, \nu) \in \mathcal{J}^s(m, U_\zeta, C) \,\big|\, \|(1 + |\cdot|_\infty)^{-\alpha}/\varphi_1\|_\infty \leqslant C,\ \||x|\nu\|_\infty \leqslant C\big\}.$$

As alluded to in the introduction, a Gaussian component overclouds the jumps in addition to the discrete observations and is therefore treated as the severely ill-posed case for $\alpha, r, \eta > 0$

$$\mathcal{G}^s(\alpha, m, U, r, C, \zeta, \eta) \coloneqq \big\{(\Sigma, \gamma, \nu) \in \mathcal{J}^s(m, U_\zeta, C) \,\big|\, \|\exp(-r|\cdot|_\infty^\alpha)/\varphi_1\|_\infty \leqslant C, |x|^{3-\eta}\nu(x) \leqslant C\ \forall |x| \leqslant 1\big\}.$$

The parameters α and r control the exponential decay of the characteristic function. Note that $\Sigma \neq 0$ already implies $\alpha = 2$. In this case, the assumption $|x|^{3-\eta}\nu(x) \leqslant C$ for $|x| \leqslant 1$ allows us to control the behavior of the small jumps.

In the mildly ill-posed case, the Blumenthal-Getoor index of the Lévy process is at most 1, whereas in the severely ill-posed case it is at most $\big((3-\eta)\wedge 2\big)\vee 0$, where we set $a\wedge b \coloneqq \min\{a, b\}$ and $a \vee b \coloneqq \max\{a, b\}$ for $a, b \in \mathbb{R}$.

For these regularity classes, we are able to quantify the estimation error as follows.

Theorem 5.1. *Let $\alpha, r, C, \zeta > 0, s > 1, m > 4$, and let the kernel satisfy $|\cdot|^{p+d}K \in L^1(\mathbb{R}^d)$ and (5.6) with order $p \geqslant s$. Let $U \subseteq \mathbb{R}^d$ be an open set which is bounded away from 0. We have for $0 < \delta \leqslant C$, $n \to \infty$:*

(M) If U is bounded and $h = h_{\delta,n} = (\log(T)/T)^{1/(2s+2\delta\alpha+d)}$, then uniformly in $(\Sigma, \gamma, \nu) \in \mathcal{D}^s(\alpha, m, U, C, \zeta)$

$$\sup_{x^* \in U} |\widehat{\nu}_h(x^*) - \nu(x^*)| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{T}\Big)^{s/(2s+2\delta\alpha+d)}\Big).$$

If $\delta = n^{-\varepsilon}$ with $\varepsilon \in (0,1)$, the choice $h = (\log(T)/T)^{1/(2s+d)}$ yields the rate $(\log(T)/T)^{s/(2s+d)}$.

(S) If $|\cdot|^{p+d}K \in L^1(\mathbb{R}^d)$, $\eta > 0$ and $h = h_{\delta,n} = (\log(T)/(4r\delta))^{-1/\alpha}$, then uniformly in $(\Sigma, \gamma, \nu) \in \mathcal{G}^s(\alpha, m, U, r, C, \zeta, \eta)$

$$\sup_{x^* \in U} |\widehat{\nu}_h(x^*) - \nu(x^*)| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{4r\delta}\Big)^{-s/\alpha}\Big).$$

If $\delta = n^{-\varepsilon}$ with $\frac{3}{2(s+d)+1} \vee \frac{\alpha}{2(s+d)+\alpha} < \varepsilon < 1$, the choice $h = T^{-1/(2(s+d))}$ yields the rate $T^{-s/(2(s+d))}$.

This theorem generalizes Trabs (2015, Proposition 2) to the multivariate case and additionally allows for high-frequency observations. Figs. 5.1 and 5.2 illustrate a simulation example of the estimation method.

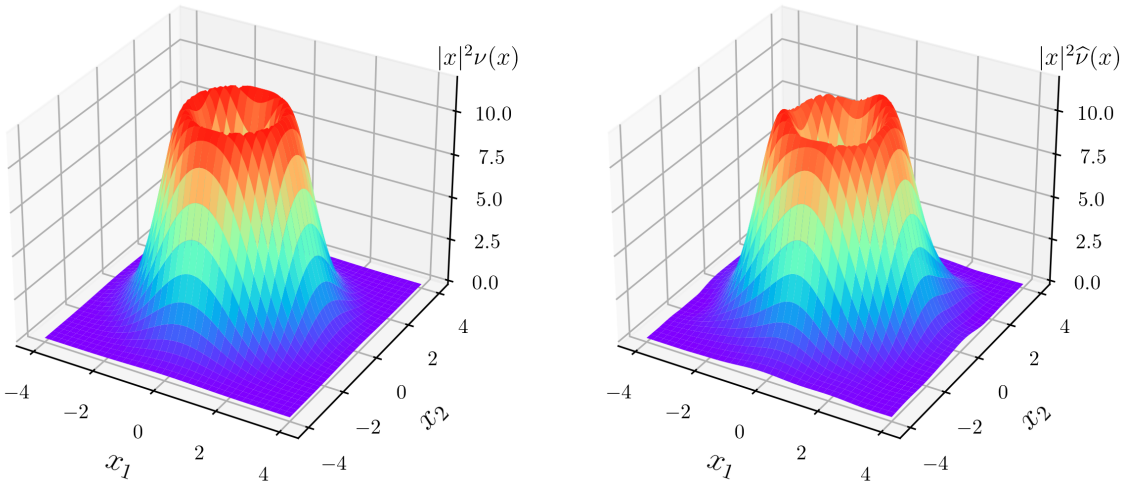

Figure 5.1: 3D plot of $|\cdot|^2\nu$ (left) and its estimate (right) for a two-dimensional compound Poisson process with Gaussian jumps.

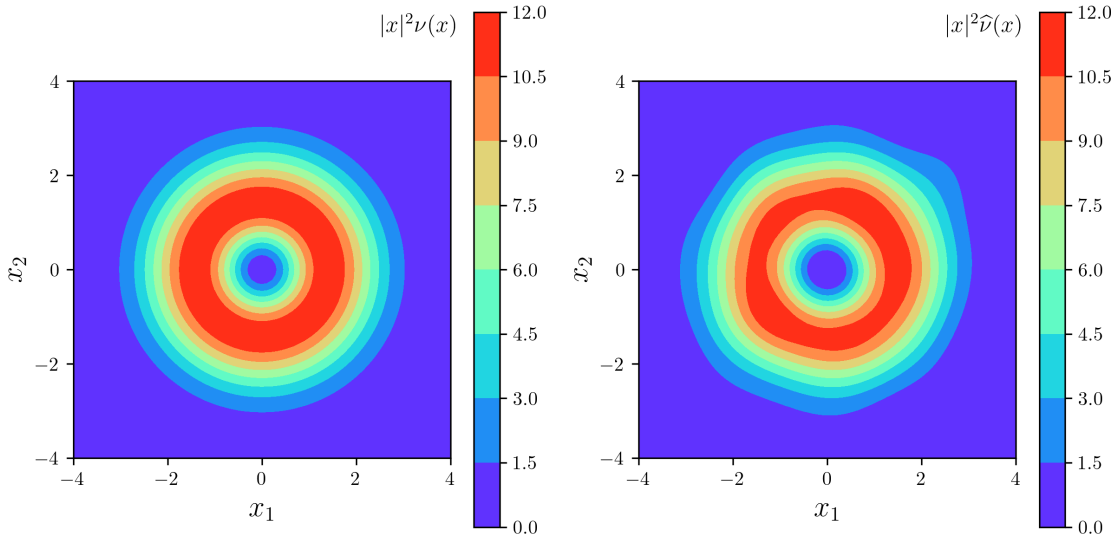

Figure 5.2: Heatmap of $|\cdot|^2\nu$ (left) and its estimate (right) for a two-dimensional compound Poisson process with Gaussian jumps.

In the mildly ill-posed case, one can easily attain the same rates without the logarithm when considering the pointwise loss.

We first discuss the low-frequency regime: For $d = 1$, our rates coincide with the proved minimax-optimal rates in the corresponding nonparametric deconvolution problems, see Fan (1991). In the mildly ill-posed case with $d = 1$, the pointwise variant of our rate has been shown to be minimax-optimal under the assumption that $x\nu$ is s-Sobolev regular, see Kappus (2012). In the severely ill-posed case with $d = 1$ and $\alpha = \{1, 2\}$, our rates coincide with the minimax-optimal rates of Neumann & Reiß (2009), who consider the integrated risk in the estimation of $\Sigma\delta_0(\mathrm{d}x) + |x|^2(1 + |x|^2)^{-1}\nu(\mathrm{d}x)$ against test functions with Sobolev regularity s. This measure has an atom in 0 and is therefore not smooth. Hence, the regularity in the rate comes purely from the test function. By considering U bounded away from 0, we can profit from the regularity of the Lévy density outside the origin. We do not even suffer an additional loss for the dimension in the rate, only in the constant. Therefore, the above suggests its optimality.

One sees that the rates improve as the time grid narrows. If this refinement happens at an appropriate order compared to the growth of the sample, the ill-posedness vanishes completely in the mildly ill-posed case and the rate becomes polynomial in the severely ill-posed case. In the mildly ill-posed case with high-frequency observations, the rate corresponds to the minimax-optimal rate in a nonparametric regression.

It is straightforward to see from our proof that when estimating $|\cdot|^2\nu$, we can forgo the exclusion

of the origin from U while achieving the same rates in the mildly ill-posed case. In the severely ill-posed case, the unknown volatility of the Brownian component of the Lévy process obstructs the observation of the small jumps. Hence, we can benefit from a pilot estimator for Σ. As discussed earlier, even with a minimax-optimal estimator for Σ, we would suffer a loss in the overall rate. However, in view of (5.3), it suffices to estimate the one-dimensional parameter $\operatorname{tr}(\Sigma)$ which is easier compared to the $d \times d$-matrix Σ. Following the spectral approach again, we propose the estimator

$$\widehat{\operatorname{tr}(\Sigma)} \coloneqq \widehat{\operatorname{tr}(\Sigma)}_h \coloneqq -\int W_h(u)\widehat{\Delta\psi_n}(u)\,\mathrm{d}u, \tag{5.7}$$

where $W_h = h^d W(h\cdot)$ for a bandwidth $h > 0$ (corresponding to the threshold h^{-1}) and a weight function $W\colon \mathbb{R}^d \to \mathbb{R}$ with

$$\int W(u)\,\mathrm{d}u = 1 \qquad \text{and} \qquad \operatorname{supp} W \subseteq [-1,1]^d.$$

This estimator achieves a rate of $(\log T)^{-(s+d)/\alpha}$ and is incorporated into the estimator for $|\cdot|^2\nu$ via

$$\widehat{|\cdot|^2\nu}_h \coloneqq -\mathscr{F}^{-1}\big[\mathscr{F}K_h\big(\widehat{\Delta\psi_n} + \widehat{\operatorname{tr}(\Sigma)}_h\big)\big]$$

leading to the following extension of Theorem 5.1.

Proposition 5.2. *Let $\alpha, r, C, \zeta, \eta > 0, 1 < s \in \mathbb{N}, m > 4$, and let the kernel satisfy $|\cdot|^{p+d}K \in L^1(\mathbb{R}^d)$ and (5.6) with order $p \geqslant s$. Assume $\|\mathscr{F}^{-1}[W(x)/x_k^s]\|_{L^1} < \infty$ for some k. Choosing $h = (\log(T)/(4r\delta))^{-1/\alpha}$, we have uniformly in $(\Sigma, \gamma, \nu) \in \mathcal{G}^s(\alpha, m, \mathbb{R}^d, r, C, \zeta, \eta)$*

$$\sup_{x^* \in \mathbb{R}^d} \big|\big(\widehat{|\cdot|^2\nu}_h\big)(x^*) - |x^*|^2\nu(x^*)\big| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{4r\delta}\Big)^{-s/\alpha}\Big).$$

5.1.2 Independent components

Compared to the one-dimensional case, we need to take the dependence structure of the components of the process into account. In particular, our previous assumption about ν having a Lebesgue density on $\mathbb{R}^d$, rules out Lévy processes where all components are independent, since the corresponding Lévy measure would only have mass on the coordinate cross. Similarly, Lévy processes consisting of multiple mutually independent blocks of components, where the components within the same block depend on each other, are not covered. For the sake of notational simplicity, we focus on the case of two equisized independent blocks: Let d be even and $L = (L^{(1)}, L^{(2)})$, where $L^{(1)}$ and $L^{(2)}$ are two independent Lévy processes on $\mathbb{R}^{d/2}$ with characteristic triplets $(\Sigma_1, \gamma_1, \nu_1)$ and $(\Sigma_2, \gamma_2, \nu_2)$, respectively. Denoting by δ_0 the Dirac measure in

$0 \in \mathbb{R}^{d/2}$, it holds that

$$\nu(\mathrm{d}x) = \nu_1(\mathrm{d}x^{(1)}) \otimes \delta_0(\mathrm{d}x^{(2)}) + \delta_0(\mathrm{d}x^{(1)}) \otimes \nu_2(\mathrm{d}x^{(2)}), \qquad x = (x^{(1)}, x^{(2)}),\ x^{(1)}, x^{(2)} \in \mathbb{R}^{d/2}. \tag{5.8}$$

We summarize the class of such Lévy processes as

$$\widetilde{\mathcal{J}}(m, C) \coloneqq \Big\{ (\Sigma, \gamma, \nu) \,\Big|\, \Sigma = \left(\begin{smallmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{smallmatrix}\right), \operatorname{tr}(\Sigma) \leqslant C, \Sigma_1, \Sigma_2 \in \mathbb{R}^{d/2 \times d/2} \text{ positive-semidefinite}, \gamma \in \mathbb{R}^d, \int |x|^m \, \nu(\mathrm{d}x) \leqslant C, \nu \text{ has the form (5.8), and } \nu_1, \nu_2 \text{ have Lebesgue densities} \Big\}$$

for $m, C > 0$. A simple example of such a Lévy measure and its estimate are illustrated in Fig. 5.3.

As before, we distinguish between the mildly ill-posed case with $\alpha > 0$

$$\widetilde{\mathcal{D}}(\alpha, m, C) \coloneqq \big\{ (0, \gamma, \nu) \in \widetilde{\mathcal{J}}(m, C) \,\big|\, \|(1 + |\cdot|_\infty)^{-\alpha} / \varphi_1\|_\infty \leqslant C,\ \||x_k| \nu_k\|_\infty \leqslant C, k = 1, 2 \big\}$$

and the severely ill-posed case with $\alpha, r, \eta > 0$

$$\widetilde{\mathcal{G}}(\alpha, m, r, C, \eta) \coloneqq \big\{ (\Sigma, \gamma, \nu) \in \widetilde{\mathcal{J}}(m, C) \,\big|\, \| \exp(-r|\cdot|_\infty^\alpha) / \varphi_1\|_\infty \leqslant C, \ |x_k|^{3-\eta} \nu_k(x_k) \leqslant C \ \forall |x_k| \leqslant 1, k = 1, 2 \big\}$$

based on the decay behavior of the characteristic function and the presence of a Gaussian component.

If the dependence structure were known, we could separate the blocks in the observations, apply our method to each block, and obtain an estimator for the overall Lévy measure. Since this is not the case, we are left with applying our initial method. In spite of the unknown dependence structure, we are able to quantify the estimation error. Due to the structure of the Lévy measure, we cannot hope for a pointwise quantitative bound. Instead, we consider the error in a functional sense. To this end, we introduce the following class of test functions for $\varrho > 0$ and $U \subseteq \mathbb{R}^d$

$$F_\varrho(U, C) \coloneqq \{ f \colon \mathbb{R}^d \to \mathbb{R} \mid f \in \mathcal{C}^\varrho(\mathbb{R}^d), \|f\|_{\mathcal{C}^\varrho(\mathbb{R}^d)}, \|f\|_{L^1(\mathbb{R}^d)} \leqslant C, \operatorname{supp} f \subseteq U \}.$$

Theorem 5.3. *Let $\alpha, r, C > 0, \varrho > 1, m > 4$, let the kernel have product structure and satisfy $|\cdot|^{p+d} K \in L^1(\mathbb{R}^d)$ and (5.6) with order $p \geqslant \varrho$. Then, we have for $0 < \delta \leqslant C, n \to \infty$:*

(M) If $U \subseteq \mathbb{R}^d$ is bounded and $h = (\log(T)/T)^{1/(2\varrho + 2\delta\alpha + 3d/2)}$, then uniformly in $(\Sigma, \gamma, \nu) \in$

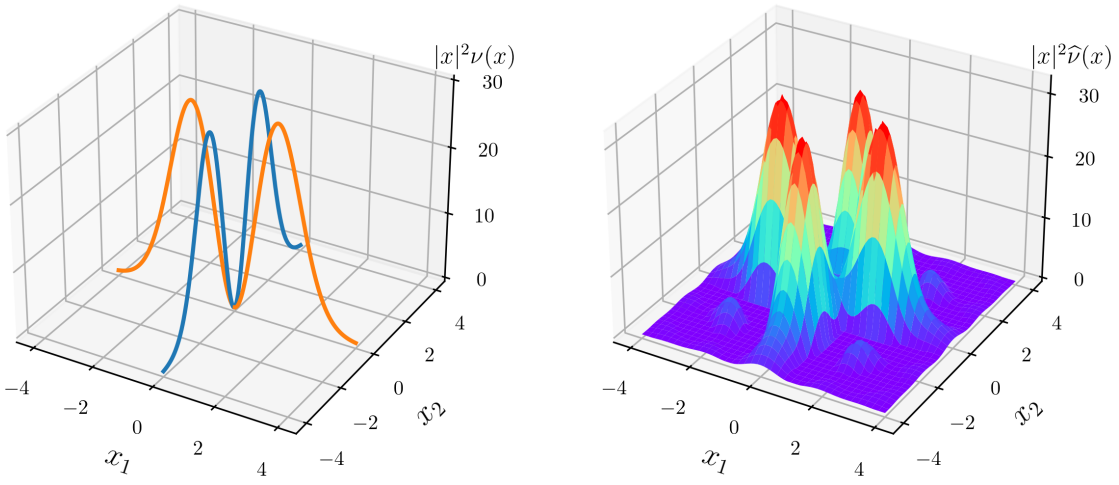

Figure 5.3: 3D plot of $|\cdot|^2\nu$ (left) and its estimate (right) for a Lévy process where both components are independent compound Poisson processes with Gaussian jumps.

$\widetilde{\mathcal{D}}(\alpha, m, C)$

$$\sup_{f\in F_\varrho(U,C)} \Big| \int_U f(x)|x|^2\big(\nu(\mathrm{d}x) - \widehat{\nu}_h(\mathrm{d}x)\big)\Big| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{T}\Big)^{\varrho/(2\varrho+2\delta\alpha+3d/2)}\Big).$$

(S) If $U \subseteq \mathbb{R}^d$ is bounded away from 0, $|\cdot|^{p+d}K \in L^1(\mathbb{R}^d)$, $\eta > 0$ and $h = (\log(T)/(4r\delta))^{-1/\alpha}$, then uniformly in $(\Sigma, \gamma, \nu) \in \widetilde{\mathcal{G}}(\alpha, m, r, C, \eta)$

$$\sup_{f\in F_\varrho(U,C)} \Big| \int_U f(x)|x|^2\big(\nu(\mathrm{d}x) - \widehat{\nu}_h(\mathrm{d}x)\big)\Big| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{4r\delta}\Big)^{-\varrho/\alpha}\Big).$$

Note that the regularity parameter ϱ in the rates comes from the smoothness of the test functions as compared to the smoothness s of the Lévy measure in Theorem 5.1. In the severely ill-posed case, the result is analogous to the well-specified. In the mildly ill-posed case, we pay for the dependence structure with an $d/2$ in the rate. Morally, one can interpret this as the model dimension being $3d/2$ instead of d.

Remark 5.4. The product kernel is compatible with any dependence structure of blocks, regardless of their size. For instance, if all components of the process are independent, one still obtains the analogous result in the severely ill-posed case. In the mildly ill-posed case, the dimension appearing in the rate is $2d - 1$ instead of $3d/2$. Comparing the dependence structures, one finds that the two independent blocks are an in-between case of no independent blocks and fully independent components.

5.1.3 A uniform risk-bound for the characteristic function and linearization

A key ingredient in the proofs of our preceding results is the following moment-uniform risk bound for the multivariate characteristic function and its partial derivatives. It generalizes the existing results in the univariate case (see Kappus & Reiß 2010, Theorem 1) and the multivariate non-uniform case (see Belomestny & Trabs 2018, Proposition A.1).

Proposition 5.5. *Let $X_1, X_2, \dots$ be $\mathbb{R}^d$-valued i.i.d. random variables with characteristic function φ and empirical characteristic function $\widehat{\varphi}_n$ such that $\mathbb{E}[|X_1|^{2\beta}|X_1|^\tau] \lesssim \rho^{|\beta|_1\wedge 1}$ and $\mathbb{E}[|X_1|^{2\beta}] \lesssim \rho^{|\beta|_1\wedge 1}$ for some multi-index $\beta \in \mathbb{N}_0^d$ and $\tau, \rho > 0$. For the inverse softplus-type weight function $w(u) = \log(\mathrm{e} + |u|)^{-(1+\chi)/2}$ with $\chi > 0$, we have*

$$\mathbb{E}\big[\big\|w(u)(\widehat{\varphi}_n - \varphi)^{(\beta)}(u)\big\|_\infty\big] \lesssim \rho^{(|\beta|_1\wedge 1)/2} n^{-1/2}.$$

As a direct consequence of Proposition 5.5, the indicator in the definition (5.4) equals 1 on the support of $\mathscr{F}K_h$, with probability converging to 1 for the bandwidths we consider.

To prove our rates for $\widehat{\nu}_h$, we decompose the error into

$$\begin{aligned}\widehat{\nu}_h(x^*) - \nu(x^*) &= |x^*|^{-2}\big(\big(K_h * (|\cdot|^2\nu) - |\cdot|^2\nu\big)(x^*) \\ &\qquad - \mathscr{F}^{-1}\big[\mathscr{F}K_h\big(\widehat{\Delta\psi_n} - \Delta\psi\big)\big](x^*) + \operatorname{tr}(\Sigma)K_h(x^*)\big) \\ &= |x^*|^{-2}\big(\underbrace{\big(K_h * (|\cdot|^2\nu) - |\cdot|^2\nu\big)(x^*)}_{=:B^\nu(x^*)} \\ &\qquad - \underbrace{\mathscr{F}^{-1}\big[\mathscr{F}K_h\delta^{-1}\Delta\big((\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta\big)\big](x^*)}_{=:L^\nu_{\delta,n}(x^*)} + R_{\delta,n} + \operatorname{tr}(\Sigma)K_h(x^*)\big) \quad (5.9)\end{aligned}$$

into a bias term B^ν, the linearized stochastic error $L^\nu_{\delta,n}$, the error $\operatorname{tr}(\Sigma)K_h$ due to the volatility, and a remainder term $R_{\delta,n}$. Proposition 5.5 applied to the increments of the Lévy process leads to the following linearization.

Lemma 5.6. *Let $\int |x|^{4+\tau}\,\nu(\mathrm{d}x) \leqslant C$ for some $\tau > 0$. If $n^{-1/2}(\log h^{-1})^{(1+\chi)/2}\|\varphi_\delta^{-1}\|_{L^\infty(I_h)} \to 0$ as $n \to \infty$ for $h \in (0,1), \chi > 0$, it holds*

$$\sup_{|u|_\infty \leqslant h^{-1}} \big|\widehat{\Delta\psi_n}(u) - \Delta\psi(u) - \delta^{-1}\Delta\big((\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta\big)(u)\big| = \mathcal{O}_\mathbb{P}(a_n), \qquad \text{where}$$

$$a_n \coloneqq n^{-1}(\log h^{-1})^{1+\chi}\|\varphi_\delta^{-1}\|^2_{L^\infty(I_h)}\delta^{-1/2}\big(\delta\big\||\nabla\psi|\big\|_{L^\infty(I_h)} + \delta^{3/2}\big\||\nabla\psi|\big\|^2_{L^\infty(I_h)} + 1\big).$$

As a direct consequence, the remainder term is of the order

$$|R_{\delta,n}| = \mathcal{O}_{\mathbb{P}}\big(h^{-d}a_n\big). \tag{5.10}$$

After treating the four terms in (5.9), the asserted rates follow from our bandwidth choices. The full proofs are postponed to Section 5.3.

5.2 Simulation examples

We demonstrate the estimation of the Lévy density for $d = 2$ with three examples: a compound Poisson process, a variance gamma process and two independent compound Poisson processes.

A challenge is to find examples of multivariate Lévy processes for which paths can be simulated and the true Lévy measure is accessible (at least numerically). To compensate for the possible singularity of the Lévy density at the origin, we plot $|\cdot|^2\nu$ and its estimate. Throughout, we use the flat-top-kernel K, see McMurry & Politis (2004), as defined by its Fourier transform

$$\mathscr{F}K(u) := \begin{cases} 1, & |u| \leqslant c, \\ \exp\Big(-\frac{b\exp(-b/(|u|-c))^2}{(|u|-1)^2}\Big), & c < |u| < 1, \\ 0, & |u| \geqslant 1, \end{cases}$$

whose decay behavior is controlled by $b > 0$ and $0 < c < 1$. In our simulations, $b = 1$, $c = 1/50$ deliver stable results. While a product kernel is convenient for theoretical reasons in Section 5.1.2, it does not seem necessary in practice. Throughout, we simulate increments of the processes with a time difference of $\delta = 0.001$ and fix the bandwidth at $h = 4T^{-1/2}$. To conquer this ill-posed problem, we use large samples of $n = 500000$ increments. From the definition (5.5) of the estimator, it is not guaranteed that $\widehat{\nu} \geqslant 0$, and for numerical reasons even $\widehat{\nu}(x) \in \mathbb{C} \setminus \mathbb{R}$ is possible for some $x \in \mathbb{R}^d$ in practice. Therefore, we consider the estimator $\mathrm{Re}(\widehat{\nu}) \vee 0$ in our simulations.

The most straightforward example under consideration is the compound Poisson process with intensity $\lambda = 100$ and two-dimensional standard-Gaussian jumps. In this case, the Lévy density is just the standard normal density, rescaled with the intensity λ. Fig. 5.1 illustrates that the method captures the overall shape of the density. The heatmap in Fig. 5.2 provides a more detailed view especially around the origin. We observe that the decay for $|x| \to \infty$ and $|x| \searrow 0$ is well-estimated, with slight problems only arising on an annulus around the origin.

A practical way to construct easy-to-simulate multivariate Lévy processes is to subordinate multivariate Brownian motion. In particular, we use a gamma process with variance $\kappa = 1$ to subordinate a two-dimensional standard Brownian motion. To access the Lévy measure of the resulting variance gamma process, we approximate the theoretical expression from Cont & Tankov (2004, Theorem 4.2) numerically. The results are again illustrated in a 3D plot (Fig. 5.4) and as a heatmap (Fig. 5.5). In this example, the estimator suffers from oscillations around the true density which are to be expected from spectral-based methods.

To demonstrate the method under the dependence structure discussed in Section 5.1.2, we consider a Lévy process comprised of two independent compound Poisson processes, each with intensity $\lambda = 100$ and one-dimensional standard-Gaussian jumps. In contrast to the two-dimensional compound Poisson process considered at the beginning of this section, the jumps in both components are driven by independent Poisson processes. The corresponding Lévy measure takes the form (5.8), where ν_1 and ν_2 are one-dimensional standard-Gaussian densities, rescaled with λ, as illustrated on the left-hand side of Fig. 5.3. It is important to emphasize that the blue and the orange line represent the Lebesgue densities of both components on $\mathbb{R}$, not $\mathbb{R}^2$. The right-hand side of the aforementioned figure reveals a strong performance of the estimator on the coordinate cross. Around the axes, we observe a smearing effect due to the singularity of the true Lévy measure on the coordinate cross before the estimate drops off farther away from the origin.

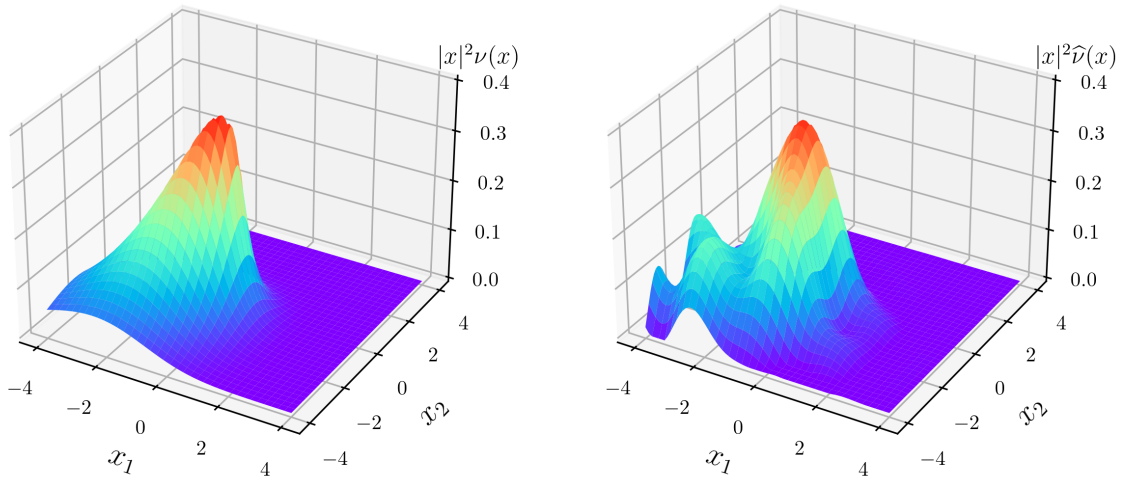

Figure 5.4: 3D plot of $|\cdot|^2\nu$ (left) and its estimate (right) for a two-dimensional variance gamma process.

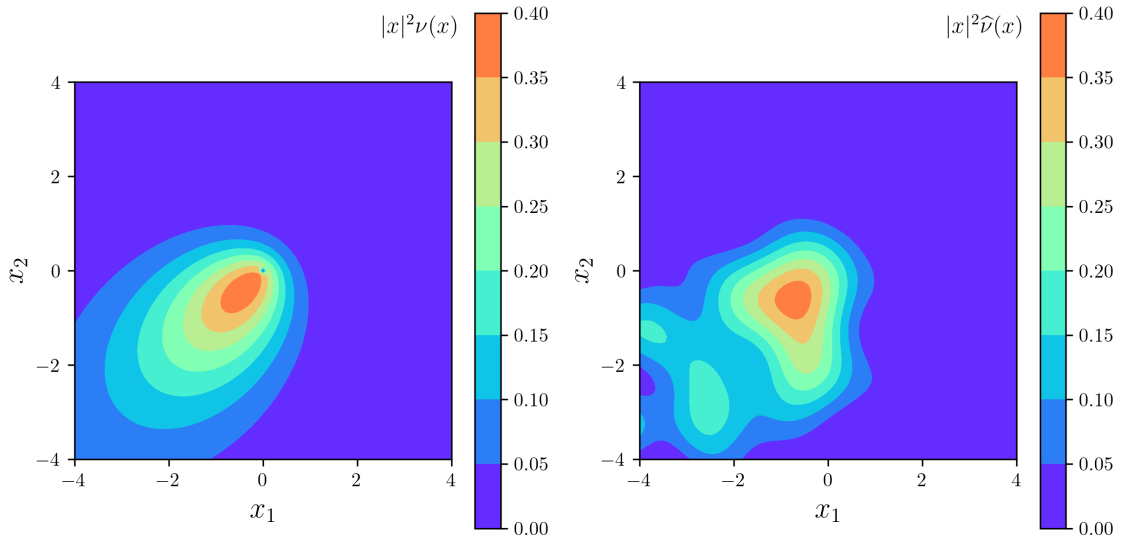

Figure 5.5: Heatmap of $|\cdot|^2\nu$ (left) and its estimate (right) for a two-dimensional variance gamma process.

5.3 Proofs

Throughout, set $I_h := [-h^{-1}, h^{-1}]^d$ for $h > 0$. Note that $\Sigma, \nu, \Delta\psi$ and $\widehat{\Delta\psi_n}$ do not change if we consider increments based on the Lévy process $(L_t - t\gamma_0)_{t\geqslant 0}$ for some $\gamma_0 \in \mathbb{R}^d$. Hence, no generality is lost if we choose γ_0 such that in the mildly ill-posed case

$$\nabla\psi = \mathrm{i}\mathscr{F}[x\nu] \qquad \text{and} \qquad \Delta\psi = -\mathscr{F}[|x|^2\nu] \tag{5.11}$$

and in the severely ill-posed case $\gamma = 0$, see Nickl et al. (2016, Lemma 12) for a similar argument in the one-dimensional case.

Further, due to the infinite divisibility of ν, the decay behavior of φ_1 governs that of φ_δ. In particular, we have for $0 < \delta \leqslant C$

$$\|(1 + |\cdot|_\infty)^{-\delta\alpha}/\varphi_\delta\|_\infty \leqslant (1 \vee C)^C \qquad \text{and} \qquad \|\exp(-r\delta|\cdot|_\infty^\alpha)/\varphi_\delta\|_\infty \leqslant (1 \vee C)^C$$

in the mildly and the severely ill-posed case, respectively.

5.3.1 Proof of Theorem 5.1

We extend the proof strategy by Trabs (2015) to accommodate for the multivariate setting. To allow for the application to high-frequency observations, we carefully keep track of δ throughout.

Subsequently, we will analyze the four terms in (5.9).

5.3.1.1 Controlling the linearized stochastic error

To control

$$L^{\nu}_{\delta,n} \coloneqq \mathscr{F}^{-1}\big[\mathscr{F}K_h\delta^{-1}\Delta\big((\widehat{\varphi}_{\delta,n}-\varphi_\delta)/\varphi_\delta\big)\big],$$

we need to get a grip on the partial derivatives of $\widehat{\varphi}_{\delta,n}-\varphi_\delta$ in the Laplacian of $(\widehat{\varphi}_{\delta,n}-\varphi_\delta)/\varphi_\delta$. In particular, we show that

$$\sup_{u\in\mathbb{R}^d}\mathbb{E}\Big[\Big|\frac{\partial^l}{\partial u_k^l}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(u)\Big|\Big] \leqslant \sup_{u\in\mathbb{R}^d}\mathbb{E}\Big[\Big|\frac{\partial^l}{\partial u_k^l}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(u)\Big|^2\Big]^{1/2} \overset{!}{\lesssim} n^{-1/2}\delta^{(l\wedge 1)/2}, \qquad l=0,1,2. \tag{5.12}$$

Since

$$\mathbb{E}\Big[\Big|\frac{\partial^l}{\partial u_k^l}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(u)\Big|^2\Big] \leqslant n^{-1}\mathbb{E}[Y_{1,k}^{2l}] = n^{-1}\Big|\frac{\partial^{2l}}{\partial u_k^{2l}}\varphi_\delta(0)\Big| \qquad \forall u\in\mathbb{R}^d,$$

where $Y_{1,k}$ denotes the k-th entry of Y_1, the case $l=0$ is obvious and for $l=1,2$ it remains to show that

$$\Big|\frac{\partial^{2l}}{\partial u_k^{2l}}\varphi_\delta(0)\Big| \lesssim \delta.$$

In the mildly ill-posed case, we have

$$\Big|\frac{\partial}{\partial u_k}\psi(0)\Big| \lesssim \int |x_k|\,\nu(\mathrm{d}x) \leqslant \int |x|\,\nu(\mathrm{d}x) \lesssim 1$$

and in the severely ill-posed case

$$\frac{\partial}{\partial u_k}\psi(0) = 0.$$

The product rule for higher order derivatives yields

$$\begin{aligned}
\Big|\frac{\partial^2\varphi_\delta}{\partial u_k^2}(0)\Big| &= \Big|\delta\varphi_\delta(u)\Big(\delta\Big(\frac{\partial}{\partial u_k}\psi(u)\Big)^2+\frac{\partial^2}{\partial u_k^2}\psi(u)\Big)\Big|_{u=0} \lesssim \delta\Big(1+\int|x|^2\,\nu(\mathrm{d}x)\Big)\lesssim\delta \qquad \text{and}\\
\Big|\frac{\partial^4\varphi_\delta}{\partial u_k^4}(0)\Big| &= \Big|\frac{\partial^2}{\partial u_k^2}\Big(\frac{\partial^2\varphi_\delta}{\partial u_k^2}(u)\Big)(0)\Big|\\
&\lesssim \delta\sum_{j=0}^{2}\binom{2}{j}|\mathbb{E}[Y_{1,k}^{2-j}]|\Big(\Big|\frac{\partial^j}{\partial u_k^j}\Big(\frac{\partial\psi}{\partial u_k}(u)\Big)^2\Big|_{u=0}+\Big|\frac{\partial^{j+2}\psi}{\partial u_k^{j+2}}(0)\Big|\Big)\lesssim\delta,
\end{aligned}$$

where all emerging partial derivatives can again be absolutely and uniformly bounded using our assumptions on ν.

To simplify the notation, set $m_{\delta,h}\coloneqq\mathscr{F}K_h/\varphi_\delta$ and recall $x\cdot y\coloneqq\sum_{k=1}^d x_k y_k$ for $x,y\in\mathbb{C}^d$.

For the severely ill-posed case, we have $|\Delta\psi(u)| \lesssim 1$ and that $|\mathrm{e}^{\mathrm{i}\langle u,x\rangle} - 1| \leqslant |x||u|$ implies $|\nabla\psi(u)| \lesssim |u|$. Together with (5.12), we obtain

$$\begin{aligned}
\mathbb{E}\Big[\sup_{x^*\in\mathbb{R}^d} |L^\nu_{\delta,n}(x^*)|\Big] &\leqslant \delta^{-1}\mathbb{E}\big[\big\|\mathscr{F}^{-1}[m_{\delta,h}\Delta(\widehat{\varphi}_{\delta,n}-\varphi_\delta)]\big\|_\infty\big] + 2\mathbb{E}\big[\big\|\mathscr{F}^{-1}[m_{\delta,h}\nabla(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\cdot\nabla\psi]\big\|_\infty\big] \\
&\qquad + \mathbb{E}\big[\big\|\mathscr{F}^{-1}[m_{\delta,h}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(\delta(\nabla\psi)^2-\Delta\psi)]\big\|_\infty\big] \\
&\lesssim (2\pi)^{-d}\int_{I_h}\Big(\delta^{-1}\mathbb{E}\big[\big|\Delta(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(u)\big|\big] + \mathbb{E}\big[\big|\nabla(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(u)\cdot\nabla\psi(u)\big|\big] \\
&\qquad + \mathbb{E}\big[\big|(\widehat{\varphi}_{\delta,n}-\varphi_\delta)(u)\big|\big]\big(|\Delta\psi(u)| + \delta|\nabla\psi(u)|^2\big)\Big)\exp(r\delta|u|_\infty^\alpha)\,\mathrm{d}u \\
&\lesssim \pi^{-d}T^{-1/2}\int_{I_h}\big(1+\delta|u|+\delta^{1/2}+\delta^{3/2}|u|^2\big)\exp(r\delta|u|_\infty^\alpha)\,\mathrm{d}u \\
&\lesssim T^{-1/2}\big(h^{-d}+\delta h^{-d-1}+\delta^{3/2}h^{-d-2}\big)\exp(r\delta h^{-\alpha}),
\end{aligned} \tag{5.13}$$

which is dominated by the bias.

In the mildly ill-posed case, the stochastic error needs to be decomposed further into the main stochastic error

$$M^\nu_{\delta,n} \coloneqq -\frac{1}{T}\sum_{k=1}^n \mathscr{F}^{-1}\Big[m_{\delta,h}\big(|Y_k|^2\mathrm{e}^{\mathrm{i}\langle u,Y_k\rangle} - \mathbb{E}\big[|Y_k|^2\mathrm{e}^{\mathrm{i}\langle u,Y_k\rangle}\big]\big)\Big] \qquad \text{and}$$

$$M^\nu_{\delta,n} - L^\nu_{\delta,n} = 2\mathscr{F}^{-1}\big[m_{\delta,h}\nabla(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\cdot\nabla\psi\big] + \mathscr{F}^{-1}\big[m_{\delta,h}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\big(\Delta\psi-\delta(\nabla\psi)^2\big)\big]. \tag{5.14}$$

To control the difference (5.14), note that $\||x|\nu\|_\infty \leqslant C$ and $\||x|^m\nu\|_{L^1} \leqslant C$ imply $\||x|\nu\|_{L^1}$, $\||x|\nu\|_{L^2}$, $\||x|^2\nu\|_{L^2} \lesssim 1$. Further, the support of $\mathscr{F}K$ and the decay behavior of φ_δ ensure

$$\|m_{\delta,h}\|_{L^2}^2 \lesssim \int |\mathscr{F}K(hu)|^2(1+|u|)^{2\delta\alpha}\,\mathrm{d}u \lesssim (1+h^{-1})^{2\delta\alpha}h^{-d} \lesssim h^{-2\delta\alpha-d}. \tag{5.15}$$

Hence, (5.12) and the Cauchy-Schwarz inequality together with (5.11), and the Plancherel theorem lead to

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{x^*\in U}\big|M^\nu_{\delta,n}(x^*) - L^\nu_{\delta,n}(x^*)\big|\Big] \\
&\qquad\leqslant (2\pi)^{-d}\big(2\mathbb{E}\big[\big\|m_{\delta,h}\nabla(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\cdot\nabla\psi\big\|_{L^1}\big] + \mathbb{E}\big[\big\|m_{\delta,h}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\big(\Delta\psi-\delta(\nabla\psi)^2\big)\big\|_{L^1}\big]\big) \\
&\qquad\lesssim n^{-1/2}\|m_{\delta,h}\|_{L^2}\big(\delta^{1/2}\||x|\nu\|_{L^2} + \||x|^2\nu\|_{L^2} + \delta d^2\||x|\nu\|_{L^2}\||x|\nu\|_{L^1}\big) \\
&\qquad\lesssim n^{-1/2}h^{-\delta\alpha-d/2}.
\end{aligned} \tag{5.16}$$

Being the sum of centered i.i.d. random variables, the main stochastic error for fixed x is controlled by Bernstein's inequality, as summarized in the following lemma.

Lemma 5.7. *Let $\alpha, C, \zeta > 0, m > 4$ and $x \in \mathbb{R}^d$ and let the kernel satisfy $|\cdot|^{p+d}K \in L^1(\mathbb{R}^d)$ and (5.6) for $p \geqslant 1$. If $(\Sigma, \gamma, \nu) \in \mathcal{D}^s(\alpha, m, U, C, \zeta)$, then there exists some constant $c > 0$ depending only on C, α and d such that for any $\kappa_0 > 0$ and any $n \in \mathbb{N}, 0 < \delta \leqslant C, h \in (0,1)$*

$$\mathbb{P}\big(|M^\nu_{\delta,n}(x)| \geqslant \kappa_0 T^{-1/2} h^{-\delta\alpha - d/2}\big) \leqslant 2\exp\Big(-\frac{c\kappa_0^2}{(1+|x|^3)\big(1+\kappa_0(h^d T)^{-1/2}\big)}\Big).$$

To establish a uniform bound for $x^* \in U$, a union bound extends this lemma to a discretization of the bounded set U and Lipschitz continuity of $x \mapsto M^\nu_{\delta,n}(x)$ allows us to control the discretization error. In particular, a standard covering argument yields a discretization $x_1, \dots, x_{N_n} \in \mathbb{R}^d$ of U such that $\sup_{x^* \in U} \min_{l=1,\dots,N_n} |x^* - x_l| \leqslant T^{-2}$, $N_n \lesssim T^{2d}$ and $\max_{l=1,\dots,N_n} |x_l| \leqslant Q$ with some $Q > 0$ independent of n. Since

$$M^\nu_{\delta,n} = K_h * g \qquad \text{with} \qquad g \coloneqq \delta^{-1}\mathscr{F}^{-1}\big[\mathbb{1}_{I_h}\varphi_\delta^{-1}\Delta(\widehat{\varphi}_{\delta,n} - \varphi_\delta)\big],$$

the fundamental theorem of calculus together with the order of the kernel ensures the Lipschitz continuity of $M^\nu_{\delta,n}$ via

$$\begin{aligned}|M^\nu_{\delta,n}(x) - M^\nu_{\delta,n}(y)| = \Big|\int_0^1 (x-y)\cdot\nabla(K_h * g)(y + \tau(x-y))\,\mathrm{d}\tau\Big| &\leqslant |x-y|\|g\|_\infty \big\||\nabla K_h|_1\big\|_{L^1} \\ &\lesssim |x-y|h^{-1}\|g\|_\infty.\end{aligned}$$

Therefore, the discretization error is upper bounded by

$$\mathbb{E}\Big[\sup_{x^* \in U}\min_{l=1,\dots,N_n} |M^\nu_{\delta,n}(x^*) - M^\nu_{\delta,n}(x_l)|\Big] \lesssim T^{-5/2}h^{-1}\int_{I_h} |\varphi_\delta^{-1}(u)|\,\mathrm{d}u \lesssim T^{-5/2}h^{-\delta\alpha-d-1}.$$

Combining the above with Markov's inequality yields for any κ_0 such that $2d < \frac{c}{6}\kappa_0^2/(1+Q^3)$ with c from Lemma 5.7 and T with $\kappa_0^2 \log(T)/(Th^d) \leqslant 1$

$$\begin{aligned}&\mathbb{P}\Big(\sup_{x^* \in U}|M^\nu_{\delta,n}(x^*)| > \kappa_0\Big(\frac{\log T}{T}\Big)^{1/2} h^{-\delta\alpha-d/2}\Big) \\ &\quad\leqslant \mathbb{P}\Big(\max_{l=1,\dots,N_n}|M^\nu_{\delta,n}(x_l)| + \sup_{x^*\in U}\min_{l=1,\dots,N_n}|M^\nu_{\delta,n}(x^*) - M^\nu_{\delta,n}(x_l)| > \kappa_0\Big(\frac{\log T}{T}\Big)^{1/2}h^{-\delta\alpha-d/2}\Big) \\ &\quad\leqslant \frac{2}{\kappa_0}\Big(\frac{\log T}{T}\Big)^{-1/2}h^{\delta\alpha+d/2}\mathbb{E}\Big[\sup_{x^*\in U}\min_{l=1,\dots,N_n}|M^\nu_{\delta,n}(x^*) - M^\nu_{\delta,n}(x_l)|\Big] \\ &\qquad\qquad + 2N_n\exp\Big(-\frac{c\kappa_0^2\log T}{2(1+Q^3)\big(2+\kappa_0(\log(T)/(Th^d))^{1/2}\big)}\Big) \\ &\quad\lesssim \frac{2}{\kappa_0}\Big(\frac{\log T}{T}\Big)^{-1/2}h^{-d/2-1}T^{-5/2} + 2\exp\Big(\Big(2d - \frac{c\kappa_0^2}{6(1+Q^3)}\Big)\log T\Big). \qquad (5.17)\end{aligned}$$

The second term obviously converges to 0 as $T \to \infty$. For the first term, $3d/2 \geqslant d/2 + 1$ implies

$$\frac{2}{\kappa_0}\Big(\frac{\log T}{T}\Big)^{-1/2} h^{-d/2-1} T^{-5/2} \leqslant \frac{2}{\kappa_0}\Big(\frac{\log T}{T}\Big)^{-1/2} h^{-3d/2} T^{-5/2} = \frac{2}{\kappa_0}\Big(\frac{\log T}{Th^d}\Big)^{3/2} T^{-1/2} (\log T)^{-2}$$

and the right-hand side converges to 0 by our choice of bandwidth.

Overall, (5.16) and (5.17) show

$$|L^{\nu}_{\delta,n}| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{T}\Big)^{1/2} h^{-\delta\alpha - d/2}\Big),$$

which our bandwidths balance with the bias.

5.3.1.2 Controlling the error due to the volatility

We now consider the last term in (5.9). The mildly ill-posed case is trivial since $\Sigma = 0$. Turning to the severely ill-posed case, we first aim to bound $|x|^{p+d}|K(x)|$. To this end, consider

$$|x_k|^{p+d}|K(x)| \leqslant \frac{1}{(2\pi)^d}\Big\|\frac{\partial^{p+d}\mathscr{F}K}{\partial u_k^{p+d}}\Big\|_{L^1(I_1)} = \frac{1}{(2\pi)^d}\int_{I_1}\Big|\int \mathrm{e}^{\mathrm{i}\langle u,z\rangle} K(z) z_k^{p+d}\,\mathrm{d}z\Big|\,\mathrm{d}u \lesssim 1.$$

It follows from the equivalence of norms $|x| \lesssim |x|_{p+d}$ that

$$|x|^{p+d}|K(x)| \lesssim |x|_{p+d}^{p+d}|K(x)| = \sum_{k=1}^{d} |x_k|^{p+d}|K(x)| \lesssim 1. \tag{5.18}$$

Thus,

$$|K_h(x^*)| \leqslant h^{-d} \sup_{|x|\geqslant|x^*|/h} |K(x)| \leqslant h^{-d} \sup_{|x|\geqslant|x^*|/h} \frac{|x|^{p+d}}{|x^*/h|^{p+d}}|K(x)| \lesssim h^p |x^*|^{-p-d}$$

and since U is bounded away from 0, this gives a uniform bound in x^* of the order h^s as $p \geqslant s$.

5.3.1.3 Controlling the bias

For $x^* \in U$ and $h|x| < \zeta$, we use a multivariate Taylor expansion of $g \coloneqq |\cdot|^2\nu \in \mathcal{C}^s(U_\zeta)$ around x^* to obtain

$$g(x^* - hx) - g(x^*) = \sum_{0<|\beta|_1<\lfloor s\rfloor_<} \frac{g^{(\beta)}(x^*)}{\beta!}(-hx)^\beta + \sum_{|\beta|_1=\lfloor s\rfloor_<} \frac{g^{(\beta)}(x^* - \tau_{x^*-hx}hx)}{\beta!}(-hx)^\beta,$$

for some $\tau_{x^*-hx} \in [0,1]$. The order of the kernel and the Hölder regularity of g yield

$$\begin{aligned}
|B^\nu(x^*)| &= \big|\big(K_h * (|\cdot|^2\nu) - |\cdot|^2\nu\big)(x^*)\big| \\
&= \Big|\int \big(g(x^* - hx) - g(x^*)\big)K(x)\,\mathrm{d}x\Big| \\
&\leqslant \Big|\int_{|x|\geqslant\zeta/h} \Big(g(x^* - hx) - \sum_{|\beta|_1\leqslant\lfloor s\rfloor_<} \frac{g^{(\beta)}(x^*)}{\beta!}(-hx)^\beta\Big)K(x)\,\mathrm{d}x\Big| \\
&\qquad + \Big|\int_{|x|<\zeta/h} \sum_{|\beta|_1=\lfloor s\rfloor_<} \frac{(-hx)^\beta}{\beta!}\big(g^{(\beta)}(x^* - \tau_{x^*-hx}hx) - g^{(\beta)}(x^*)\big)K(x)\,\mathrm{d}x\Big| \\
&\lesssim \int_{|x|\geqslant\zeta/h} |g(x^* - hx)K(x)|\,\mathrm{d}x + \sum_{|\beta|_1\leqslant\lfloor s\rfloor_<} \frac{\|g^{(\beta)}\|_{L^\infty(U)}}{\beta!}\frac{h^s}{\zeta^{s-|\beta|_1}}\int_{|x|\geqslant\zeta/h} |x|^s|K(x)|\,\mathrm{d}x \\
&\qquad + \sum_{|\beta|_1=\lfloor s\rfloor_<} \frac{1}{\beta!}\int_{|x|<\zeta/h} |hx|^{\lfloor s\rfloor_<}|\tau_{x^*-hx}hx|^{s-\lfloor s\rfloor_<}|K(x)|\,\mathrm{d}x.
\end{aligned}$$

The second and the third term are clearly of the order h^s. To establish the same for the first term, we proceed slightly differently for the two cases of ill-posedness.

In the severely ill-posed case, we separate the behavior of the small and the large jumps. On the one hand, (5.18) yields

$$\begin{aligned}
\int_{\substack{|x|\geqslant\zeta/h,\\ |x^*-hx|>1}} |g(x^* - hx)K(x)|\,\mathrm{d}x &\leqslant \frac{h^{p+d}}{\zeta^{p+d}}\int_{\substack{|x|\geqslant\zeta/h,\\ |x^*-hx|>1}} |x^* - hx|^2\nu(x^* - hx)|x|^{p+d}|K(x)|\,\mathrm{d}x \\
&\lesssim h^s\int_{|x|>1} |x|^2\nu(x)\,\mathrm{d}x.
\end{aligned}$$

On the other hand, the assumption $|y|^{3-\eta}\nu(y) \leqslant C$ for $|y| \leqslant 1$ and (5.18) gives

$$\begin{aligned}
\int_{\substack{|x|\geqslant\zeta/h,\\ |x^*-hx|\leqslant 1}} |g(x^* - hx)K(x)|\,\mathrm{d}x \lesssim \int_{\substack{|x|\geqslant\zeta/h,\\ |x^*-hx|\leqslant 1}} \frac{|x|^{-(p+d)}}{|x^* - hx|^{1-\eta}}\,\mathrm{d}x &\lesssim h^p\int_{\substack{|x^*-x|\geqslant\zeta,\\ |x|\leqslant 1}} \frac{|x^* - x|^{-(p+d)}}{|x|^{1-\eta}}\,\mathrm{d}x \\
&\lesssim h^s\zeta^{-(p+d)}\int_{|x|\leqslant 1} |x|^{\eta-1}\,\mathrm{d}x.
\end{aligned}$$

In the mildly ill-posed case, one uses $\||x|\nu\|_\infty \leqslant C$ and the triangle inequality to find

$$\int_{|x|\geqslant\zeta/h} |g(x^* - hx)K(x)|\,\mathrm{d}x \leqslant C \int_{|x|\geqslant\zeta/h} |x^* - hx||K(x)|\,\mathrm{d}x \lesssim h^s \frac{1+\zeta}{\zeta^s} \int |x|^s |K(x)|\,\mathrm{d}x.$$

5.3.1.4 Controlling the remainder term

To bound $R_{\delta,n}$ in (5.9), we first show that, with a_n from Lemma 5.6,

$$g \coloneqq \mathscr{F}^{-1}\big[\mathscr{F}K_h\big(\widehat{\Delta\psi_n} - \Delta\psi - \delta^{-1}\Delta((\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta)\big)\big] = \mathcal{O}_{\mathbb{P}}\big(h^{-d}a_n\big).$$

Let $\varepsilon > 0$. Owing to Lemma 5.6 we can choose $N, M > 0$ (w.l.o.g. $M > \|K\|_{L^1}$) such that the probability of the event $A_n \coloneqq \{\sup_{|u|_\infty \leqslant h^{-1}} |\widetilde{g}(u)| > a_n M\}$ with $\widetilde{g} \coloneqq \widehat{\Delta\psi_n} - \Delta\psi - \delta^{-1}\Delta((\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta)$ is less than ε for $n > N$. Due to the support of $\mathscr{F}K$, we have on A_n^{c}

$$|g(x^*)| = \frac{1}{(2\pi)^d}\Big|\int_{I_h} \mathrm{e}^{-\mathrm{i}\langle u, x^*\rangle} \mathscr{F}K(hu)\widetilde{g}(u)\,\mathrm{d}u\Big| \leqslant \frac{a_n M}{(2\pi)^d}\int_{I_h} |\mathscr{F}K(hu)|\,\mathrm{d}u \leqslant h^{-d}a_n M\|K\|_{L^1}.$$

For $M' \coloneqq M\|K\|_{L^1}$, we obtain

$$\mathbb{P}\Big(\sup_{x^*\in U} |g(x^*)| > h^{-d}a_n M'\Big) \leqslant \varepsilon,$$

whereby the remainder term has the order proposed in (5.10).

In the mildly ill-posed case, we have $\|\varphi_\delta^{-1}\|_{L^\infty(I_h)} \lesssim h^{-\delta\alpha}$ and (5.11) implies $\big\||\nabla\psi|\big\|_{L^\infty(I_h)} \lesssim 1$. Thus, we have

$$|R_{\delta,n}| = \mathcal{O}_{\mathbb{P}}\big(n^{-1}\delta^{-1/2}(\log h^{-1})^{1+\chi}h^{-2\delta\alpha-d}\big).$$

In the severely ill-posed case, $\|\varphi_\delta^{-1}\|_{L^\infty(I_h)} \lesssim \exp(r\delta h^{-\alpha})$ holds and (5.2) implies $\big\||\nabla\psi|\big\|_{L^\infty(I_h)} \lesssim h^{-1}$. Hence,

$$|R_{\delta,n}| = \mathcal{O}_{\mathbb{P}}\big(n^{-1}\delta^{-1/2}(\log h^{-1})^{1+\chi}\big(h^{-d} + \delta h^{-d-1} + \delta^{3/2}h^{-d-2}\big)\exp(2r\delta h^{-\alpha})\big).$$

In both cases, the remainder term is dominated by the linearized stochastic error.

This completes the proof of Theorem 5.1. □

5.3.2 Proof of Proposition 5.2

For the modified estimator, we have to replace $\operatorname{tr}(\Sigma)K_h$ with $\big(\operatorname{tr}(\Sigma) - \widehat{\operatorname{tr}(\Sigma)}\big)K_h$ in the decomposition (5.9). All other terms are treated as before. Since we can bound $|K_h(x^*)|$ by h^{-d} uniformly in x^* and $\delta \leqslant C$ is fixed, we only need to prove

$$\big|\widehat{\operatorname{tr}(\Sigma)} - \operatorname{tr}(\Sigma)\big| = \mathcal{O}_{\mathbb{P}}\big((\log n)^{-(s+d)/\alpha}\big).$$

Similarly to (5.9), the error for estimating the trace of Σ can be decomposed into

$$\begin{aligned}\operatorname{tr}(\Sigma) - \widehat{\operatorname{tr}(\Sigma)} &= \int W_h(u)\big(\widehat{\Delta\psi_n} - \Delta\psi\big)(u)\,\mathrm{d}u + \int W_h(u)\big(\Delta\psi(u) + \operatorname{tr}(\Sigma)\big)\,\mathrm{d}u \\ &= \underbrace{\int W_h(u)\delta^{-1}\Delta\Big(\frac{\widehat{\varphi}_{\delta,n}(u) - \varphi_\delta(u)}{\varphi_\delta(u)}\Big)(u)\,\mathrm{d}u}_{=:\widetilde{L}^\nu_{\delta,n}} + \widetilde{R}_{\delta,n} - \underbrace{\int W_h(u)\mathscr{F}[|x|^2\nu](u)\,\mathrm{d}u}_{=:\widetilde{B}^\nu_h}\end{aligned}$$

with the linearized stochastic error $\widetilde{L}^\nu_{\delta,n}$, the bias $\widetilde{B}^\nu_h$ and a remainder term $\widetilde{R}_{\delta,n}$. Using the techniques from Section 5.3.1.1, it is straightforward to see

$$\begin{aligned}\mathbb{E}[|\widetilde{L}^\nu_{\delta,n}|] &\lesssim \delta^{-1}n^{-1/2}\int |W_h(u)||\varphi_\delta^{-1}(u)|\big(\delta^{1/2} + \delta^{3/2}|\nabla\psi(u)| + \delta^2|\nabla\psi(u)|^2 + \delta\big)\,\mathrm{d}u \\ &\lesssim \delta^{-1}n^{-1/2}\|\varphi_\delta^{-1}\|_{L^\infty(I_h)}\big(\delta^{1/2} + \delta^{3/2}h^{-1} + \delta^2h^{-2}\big)\int |W_h(u)|\,\mathrm{d}u \\ &\lesssim n^{-1/2}\exp(r\delta h^{-\alpha})h^{-2}\int |W(u)|\,\mathrm{d}u,\end{aligned}$$

which is of the order $n^{-1/4}(\log n)^{2/\alpha}$ by our choice of h and will be dominated by the bias.

Using the Plancherel theorem in a similar fashion to Belomestny & Reiß (2015, Section 4.2.1), we have for $g \coloneqq |\cdot|^2\nu$ and $\beta \in \mathbb{N}_0^d$ with $\beta_l = s\mathbb{1}_{\{l=k\}}$

$$|\widetilde{B}^\nu_h| \lesssim \Big|\int \mathscr{F}^{-1}[W_h(x)/x_k^s]\mathscr{F}^{-1}\big[x_k^s\mathscr{F}g(x)\big](u)\,\mathrm{d}u\Big| \lesssim \|g^{(\beta)}\|_\infty\|\mathscr{F}^{-1}[W_h(x)/x_k^s]\|_{L^1}.$$

By substitution

$$\mathscr{F}^{-1}[W_h(x)/x_k^s](u) = \frac{h^s}{(2\pi)^d}\mathscr{F}^{-1}\big[W(x)/x_k^s\big](u/h)$$

and therefore

$$|\widetilde{B}^\nu_h| \lesssim h^s\|g^{(\beta)}\|_\infty\|\mathscr{F}^{-1}[W(x)/x_k^s](\cdot/h)\|_{L^1} \lesssim h^{(s+d)}\|g^{(\beta)}\|_\infty\|\mathscr{F}^{-1}[W(x)/x_k^s]\|_{L^1} \lesssim h^{(s+d)}.$$

Together with Lemma 5.6, we have

$$\begin{aligned}|\widetilde{R}_{\delta,n}| &\lesssim \sup_{|u|_\infty \leqslant h^{-1}} \big|\widehat{\Delta\psi_n}(u) - \Delta\psi(u) - \delta^{-1}\Delta\big((\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta\big)(u)\big| \int |W_h(u)|\,\mathrm{d}u \\ &= \mathcal{O}_{\mathbb{P}}\big(n^{-1}(\log h^{-1})^{1+\chi}\|\varphi_\delta^{-1}\|^2_{L^\infty(I_h)}h^{-2}\big),\end{aligned}$$

which is dominated by the linearized stochastic error. □

5.3.3 Proof of Theorem 5.3

The distributional analogon to (5.9) is

$$\begin{aligned}\int f(x)|x|^2\widehat{\nu}_h(\mathrm{d}x) - \int f(x)|x|^2\,\nu(\mathrm{d}x) \\ &= \int f(x)\big(K_h * (|\cdot|^2\nu)\big)(x)\,\mathrm{d}x - \int f(x)|x|^2\nu(\mathrm{d}x) \\ &\quad + \int f(x)\mathscr{F}^{-1}\big[\mathscr{F}K_h(\Delta\psi - \widehat{\Delta\psi}_n)\big](x)\,\mathrm{d}x + \int f(x)\,\mathrm{tr}(\Sigma)K_h(x)\,\mathrm{d}x \\ &= \int f(x)\big(K_h * (|\cdot|^2\nu)\big)(x)\,\mathrm{d}x - \int f(y)|y|^2\nu(\mathrm{d}y) \\ &\quad + \int f(x)L^\nu_{\delta,n}(x)\,\mathrm{d}x + \int f(x)R_{\delta,n}\,\mathrm{d}x + \int f(x)\,\mathrm{tr}(\Sigma)K_h(x)\,\mathrm{d}x\end{aligned}$$

with the same $L^\nu_{\delta,n}$ and $R_{\delta,n}$, for which we will derive uniform upper bounds on U which directly translate into bounds when integrating against test functions due to their regularity. For the integrated bias, we use Fubini's theorem to obtain

$$\begin{aligned}B^\nu_I &= \int f(x)\big(K_h * (|\cdot|^2\nu)\big)(x)\,\mathrm{d}x - \int f(y)|y|^2\nu(\mathrm{d}y) \\ &= \int f(x)\Big(\int K_h(x-y)|y|^2\,\nu(\mathrm{d}y)\Big)\,\mathrm{d}x - \int f(y)|y|^2\,\nu(\mathrm{d}y) \\ &= \int \big((f * K_h(-\cdot))(y) - f(y)\big)|y|^2\,\nu(\mathrm{d}y) \\ &= \int \big((K_h(-\cdot) * f)(y) - f(y)\big)|y|^2\,\nu(\mathrm{d}y).\end{aligned}$$

$\big((K_h(-\cdot) * f)(y) - f(y)\big)$ is of the order $(|y| \vee 1)h^\varrho$ which follows from the arguments in (5.9) with $g = f$, ϱ and $K_h(-\cdot)$ instead of $|\cdot|^2\nu$, s and K_h, respectively. Therefore,

$$|B^\nu_I| \lesssim h^\varrho \int (|y| \vee 1)|y|^2\,\nu(\mathrm{d}y) \lesssim h^\varrho \int |y|^3\,\nu(\mathrm{d}y) \lesssim h^\varrho.$$

A key tool to control the linearized stochastic error $L^{\nu}_{\delta,n}$ in Section 5.3.1.1 was (5.12), which we can still establish here by bounding the first four partial derivatives of ψ at the origin. Indeed, by (5.3) we have $\frac{\partial \psi}{\partial u_k}(0) = 0$ and similarly

$$\Big|\frac{\partial \psi}{\partial u_k^j}(u)\Big| \leqslant \operatorname{tr}(\Sigma) + \int |x|^j \, \nu(\mathrm{d}x) = \operatorname{tr}(\Sigma) + \sum_{l=1}^{2} \int |x_l|^j \, \nu_l(\mathrm{d}x_l) \lesssim 1, \quad j = 2,3,4,\ k = 1,\dots,d. \tag{5.19}$$

Hence, (5.12) holds. Additionally, (5.19) implies $|\Delta\psi(u)| \lesssim 1$.

In the severely ill-posed case, we can still bound the gradient of ψ by

$$|\nabla\psi(u)| \lesssim |\Sigma u| + \int |x||\mathrm{e}^{\mathrm{i}\langle u,x\rangle} - 1| \, \nu(\mathrm{d}x) \leqslant |u| + \int |\langle u,x\rangle||x| \, \nu(\mathrm{d}x) \lesssim |u|,$$

and then apply the arguments from (5.13). Hence, the linearized stochastic error is of the same order as before.

In the severely ill-posed case, (5.19) holds even for $j = 1$ and therefore $|\nabla\psi(u)| \lesssim 1$. Continuing from (5.12) in the mildly ill-posed case requires the most significant changes. (5.11) now reads as

$$\nabla\psi(u) = \big(\mathrm{i}\mathscr{F}[x^{(1)}\nu_1](u^{(1)}), \mathrm{i}\mathscr{F}[x^{(2)}\nu_2](u^{(2)})\big)^{\top}$$

and the main crux is that

$$\int \mathrm{e}^{\mathrm{i}\langle u^{(k)},x^{(k)}\rangle}|x^{(k)}|^j \, \nu_k(\mathrm{d}x^{(k)}), \qquad j,k = 1,2$$

are constant in half of their arguments. Therefore, they cannot be finitely integrable as functions on $\mathbb{R}^d$. In (5.14), a way out is to consider

$$\big\|m_{\delta,h}|\nabla\psi|\big\|_{L^1} \leqslant \sum_{k=1}^{2} \big\|m_{\delta,h}(u)|\mathscr{F}[x^{(k)}\nu_k](u^{(k)})|\big\|_{L^1}. \tag{5.20}$$

Then, we apply the Cauchy-Schwarz inequality and Plancherel's theorem only on $L^2(\mathbb{R}^{d/2})$ to obtain

$$\begin{aligned}
&\big\|m_{\delta,h}(u)|\mathscr{F}[x^{(1)}\nu_1](u^{(1)})|\big\|_{L^1} \\
&\qquad = \int\int \Big|\frac{\mathscr{F}K(hu)}{\varphi_\delta(u)}|\mathscr{F}[x^{(1)}\nu_1](u^{(1)})|\Big| \, \mathrm{d}u^{(1)} \, \mathrm{d}u^{(2)} \\
&\qquad \leqslant \||\mathscr{F}[x^{(1)}\nu_1]|\|_{L^2(\mathbb{R}^{d/2})} \int_{[-h^{-1},h^{-1}]^{d/2}} \Big(\int_{[-h^{-1},h^{-1}]^{d/2}} \big|\varphi_\delta^{-1}(u)\big|^2 \, \mathrm{d}u^{(1)} \Big)^{1/2} \mathrm{d}u^{(2)} \\
&\qquad \lesssim h^{-\delta\alpha - 3d/4}.
\end{aligned} \tag{5.21}$$

Analogously, the second summand in (5.20) has the same order. As a direct consequence,

$$\big\|m_{\delta,h}|\nabla\psi|^2\big\|_{L^1} \leqslant \sum_{k=1}^{2} \||x^{(k)}|\nu_k\|_{L^1}\big\|m_{\delta,h}|\nabla\psi|\big\|_{L^1} \lesssim h^{-\delta\alpha-3d/4}$$

and similarly to (5.20) and (5.21) the same holds for $\|m_{\delta,h}\Delta\psi\|_{L^1}$. Recalling (5.14), we have

$$\begin{aligned}\mathbb{E}\Big[\sup_{x^*\in U}\big|M^\nu_{\delta,n}(x^*) - L^\nu_{\delta,n}(x^*)\big|\Big] &\leqslant (2\pi)^{-d}\big(2\mathbb{E}\big[\big\|m_{\delta,h}\nabla(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\cdot\nabla\psi\big\|_{L^1}\big] \\ &\qquad + \mathbb{E}\big[\big\|m_{\delta,h}(\widehat{\varphi}_{\delta,n}-\varphi_\delta)\big(\Delta\psi - \delta(\nabla\psi)^2\big)\big\|_{L^1}\big]\big) \\ &\lesssim n^{-1/2}h^{-\delta\alpha-3d/4}.\end{aligned}$$

Note that we pay for the dependence structure with an additional $h^{-d/4}$ compared to (5.16). The same occurs when applying Bernstein's inequality to obtain the following adaptation of Lemma 5.7.

Lemma 5.8. *Let $\alpha, C, \zeta > 0, m > 4, U \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$, let the kernel have product structure and satisfy $|\cdot|^{p+d}K \in L^1(\mathbb{R}^d)$ and (5.6) for $p \geqslant 1$. If $(\Sigma,\gamma,\nu) \in \widetilde{\mathcal{D}}(\alpha,m,C)$, then there exists some constant $c > 0$ depending only on C, α and d such that for any $\kappa_0 > 0$ and any $n \in \mathbb{N}, 0 < \delta \leqslant C, h \in (0,1)$*

$$\mathbb{P}\big(|M^\nu_{\delta,n}(x)| \geqslant \kappa_0 T^{-1/2}h^{-\delta\alpha-3d/4}\big) \leqslant 2\exp\Big(-\frac{c\kappa_0^2}{(1+|x|^3)(1+\kappa_0(Th^{d/2})^{-1/2})}\Big).$$

Carrying out the discretization argument from before, the linearized stochastic error in the mildly ill-posed case is of the order

$$|L^\nu_{\delta,n}| = \mathcal{O}_{\mathbb{P}}\Big(\Big(\frac{\log T}{T}\Big)^{1/2}h^{-\delta\alpha-3d/4}\Big).$$

The term $\operatorname{tr}(\Sigma)K_h$ is treated as in Section 5.3.1.2 just with ϱ instead of s. No changes are necessary to treat the remainder term compared to Section 5.3.1.4. This is because when treating the linearized stochastic error, we already showed that still $|\nabla\psi(u)|, |\Delta\psi(u)| \lesssim 1$ in the mildly ill-posed case and $|\nabla\psi(u)| \lesssim |u|, |\Delta\psi(u)| \lesssim 1$ in the severely ill-posed case. This concludes the proof of Theorem 5.3. □

5.3.4 Remaining proofs

5.3.4.1 Proof of Proposition 5.5

The proof uses empirical process theory and is a combination of Kappus & Reiß (2010) and Belomestny & Trabs (2018).

To simplify the notation, write

$$C^{\beta}_{\rho,n}(u) \coloneqq n^{-1/2}\rho^{-(|\beta|_1\wedge 1)/2}\sum_{k=1}^{n}\frac{\partial^{\beta}}{\partial u^{\beta}}\big(\mathrm{e}^{\mathrm{i}\langle u,X_k\rangle}-\mathbb{E}[\mathrm{e}^{\mathrm{i}\langle u,X_k\rangle}]\big)$$

so that the assertion reads

$$\sup_{\substack{n\geqslant 1,\\ 0<\rho\leqslant C}}\mathbb{E}\big[\big\|w(u)C^{\beta}_{\rho,n}(u)\big\|_{\infty}\big]<\infty.$$

We decompose $C^{\beta}_{\rho,n}$ into its real and its imaginary part to obtain

$$\mathbb{E}\big[\big\|w(u)C^{\beta}_{\rho,n}(u)\big\|_{\infty}\big]\leqslant\mathbb{E}\big[\big\|w(u)\operatorname{Re}\big(C^{\beta}_{\rho,n}(u)\big)\big\|_{\infty}\big]+\mathbb{E}\big[\big\|w(u)\operatorname{Im}\big(C^{\beta}_{\rho,n}(u)\big)\big\|_{\infty}\big]. \tag{5.22}$$

As both parts can be treated analogously, we focus on the real part. To this end, introduce the class of

$$\mathcal{G}_{\rho,\beta}\coloneqq\{g_u : u\in\mathbb{R}^d\}\qquad\text{where}\qquad g_u\colon\mathbb{R}^d\to\mathbb{R},\qquad x\mapsto w(u)\rho^{-(|\beta|_1\wedge 1)}\frac{\partial^{\beta}}{\partial u^{\beta}}\cos(\langle u,x\rangle).$$

Since $G=\rho^{-(|\beta|_1\wedge 1)/2}|\cdot|^{\beta}$ is an envelope function for $\mathcal{G}_{\rho,\beta}$, Corollary 19.35 in van der Vaart (1998) yields

$$\mathbb{E}\big[\big\|w(u)\operatorname{Re}\big(C^{\beta}_{\rho,n}(u)\big)\big\|_{\infty}\big]\lesssim J_{[]}\big(\mathbb{E}[G(X_1)^2]^{1/2},\mathcal{G}_{\rho,\beta}\big)\coloneqq\int_0^{\mathbb{E}[G(X_1)^2]^{1/2}}\sqrt{\log N_{[]}(\varepsilon,\mathcal{G}_{\rho,\beta})}\,\mathrm{d}\varepsilon, \tag{5.23}$$

where $N_{[]}(\varepsilon,\mathcal{G}_{\rho,\beta})$ is the minimal number of ε-brackets (with respect to the distribution of X_1) needed to cover $\mathcal{G}_{\rho,\beta}$.

Since $|g_u(x)|\leqslant w(u)\rho^{-(|\beta|_1\wedge 1)/2}|x|^{\beta}$, the set $\{g_u : |u|>B\}$ is covered by the bracket

$$[g_0^-,g_0^+]\coloneqq\{g\colon\mathbb{R}^d\to\mathbb{R}\mid g_0^-(x)\leqslant g(x)\leqslant g_0^+(x)\,\forall x\in\mathbb{R}^d\}\qquad\text{for}$$
$$g_0^{\pm}\coloneqq\pm\varepsilon\rho^{-(|\beta|_1\wedge 1)/2}|\cdot|^{\beta}\qquad\text{and}\qquad B\coloneqq B(\varepsilon)\coloneqq\inf\big\{b>0:\sup_{|u|\geqslant b}w(u)\leqslant\varepsilon\big\}.$$

To cover $\{g_u : |u| \leqslant B\}$, we use for some grid $(u_{\rho,j})_{j\geqslant 1} \subseteq \mathbb{R}^d$ the functions

$$g^{\pm}_{\rho,j} \coloneqq \rho^{-(|\beta|_1\wedge 1)/2}\Big(w(u_{\rho,j})\frac{\partial^\beta}{\partial u^\beta_{\rho,j}}\cos(\langle u_{\rho,j},\cdot\rangle) \pm \varepsilon|\cdot|^\beta\Big)\mathbb{1}_{\{|\cdot|\leqslant M\}} \pm \rho^{-(|\beta|_1\wedge 1)/2}|\cdot|^\beta\mathbb{1}_{\{|\cdot|>M\}},$$

where $M \coloneqq \inf\{q : \rho^{-(|\beta|_1\wedge 1)}\mathbb{E}[|X_1|^{2\beta}\mathbb{1}_{\{|X_1|>q\}}] \leqslant \varepsilon^2\}$. Owing to $\mathbb{E}[|X_1|^{2\beta}] \lesssim \rho^{|\beta|_1\wedge 1}$, we have

$$\mathbb{E}\big[|g^+_j(X_1) - g^-_j(X_1)|^2\big] \leqslant 4\varepsilon^2\big(\rho^{-(|\beta|_1\wedge 1)}\mathbb{E}[|X_1|^{2\beta}] + 1\big) \leqslant c\varepsilon^2$$

for some $c > 0$. Denote by Q the Lipschitz constant of w and use the triangle inequality to see

$$\Big|w(u)\frac{\partial^\beta}{\partial u^\beta}\cos(\langle u,x\rangle) - w(u_j)\frac{\partial^\beta}{\partial u^\beta_{\rho,j}}\cos(\langle u_{\rho,j},x\rangle)\Big| \leqslant |x|^\beta(Q+|x|)|u-u_{\rho,j}|.$$

Thus, $g_u \in [g^-_j, g^+_j]$ as soon as $(Q+M)|u-u_{\rho,j}| \leqslant \varepsilon$. It takes at most $(\lceil B/\varepsilon_0\rceil)^d$ ℓ^2-balls of radius $d^{1/2}\varepsilon_0$ to cover the ℓ^2-ball of radius B around 0. For $\varepsilon_0 = \varepsilon d^{-1/2}/(Q+M)$, denote their centers by $(u_{\rho,j})_j$. To translate this into a cover of $\{g_u : |u| \leqslant B\}$, we fix some g_u with $|u| \leqslant B$. By construction, we can pick j such that $|u-u_{\rho,j}| \leqslant d^{1/2}\varepsilon_0 = \varepsilon/(Q+M)$. The previous calculations show that $[g^-_j, g^+_j]$ is a $c^{1/2}\varepsilon$-bracket containing g_u and therefore

$$N_{[]}(\varepsilon, \mathcal{G}_{\rho,\beta}) \leqslant \big(\lceil \varepsilon^{-1}(cd)^{1/2}B(Q+M)\rceil\big)^d + 1.$$

It is straightforward to see that $B \leqslant \exp(\varepsilon^{-2/(1+\chi)})$. Further, $q = (\varepsilon^{-2}\rho^{-(|\beta|_1\wedge 1)}\mathbb{E}[|X_1|^{2\beta}|X_1|^\tau])^{1/\tau}$ is sufficient for

$$\rho^{-(|\beta|_1\wedge 1)}\mathbb{E}[|X_1|^{2\beta}\mathbb{1}_{\{|X_1|>q\}}] \leqslant q^{-\tau}\mathbb{E}[|X_1|^{2\beta}|X_1|^\tau] \leqslant \varepsilon^2$$

and thus $M \leqslant (\varepsilon^{-2}\rho^{-(|\beta|_1\wedge 1)}\mathbb{E}[|X_1|^{2\beta}|X_1|^\tau])^{1/\tau} \leqslant (\varepsilon^{-2}dc')^{1/\tau}$ for some $c' > 0$. Hence,

$$\log N_{[]}(\varepsilon, \mathcal{G}_{\rho,\beta}) \lesssim 1 + \log(\varepsilon^{-2/\tau-1}) + \varepsilon^{-2/(1+\tau)} \lesssim 1 + \varepsilon^{-2/(1+\tau)}$$

implying

$$J_{[]}\big(\mathbb{E}[G(X_1)^2]^{1/2}, \mathcal{G}_{\rho,\beta}\big) = \int_0^{(\rho^{-(|\beta|_1\wedge 1)}\mathbb{E}[|X_1|^{2\beta}])^{1/2}} \sqrt{\log N_{[]}(\varepsilon, \mathcal{G}_{\rho,\beta})}\,\mathrm{d}\varepsilon < \infty.$$

In view of (5.22) and (5.23), the assertion follows. □

5.3.4.2 Proof of Lemma 5.6

Setting $g(y) = \log(1+y)$ for $y > -1$ (i.e. $g'(y) = (1+y)^{-1}, g''(y) = -(1+y)^{-2}$) and $\xi = (\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta$, we use

$$\nabla(g \circ \xi)(u) = g'(\xi(u))\nabla\xi(u), \qquad \Delta(g \circ \xi)(u) = g''(\xi(u))(\nabla\xi)^2(u) + g'(\xi(u))\Delta\xi(u)$$

and $|(\nabla\xi)^2(u)| \leqslant |\nabla\xi(u)|^2$ to obtain for $|\xi(u)| \leqslant 1/2$ that

$$\big|\Delta(g \circ \xi)(u) - \Delta\xi(u)\big| \leqslant |g''(\xi(u))||\nabla\xi(u)|^2 + |g'(\xi(u)) - 1||\Delta\xi(u)| \lesssim |\nabla\xi(u)|^2 + |\xi(u)||\Delta\xi(u)|, \tag{5.24}$$

because

$$|g'(y) - 1| \leqslant 2|y| \qquad \text{and} \qquad |g''(y)| \leqslant 4 \qquad \forall y \in \mathbb{C} : |y| \leqslant 1/2.$$

The latter statement holds, since $1/2 \leqslant |1+y|$. For the former statement, consider the expansion

$$g'(y) = \frac{1}{1+y} = \sum_{k=0}^{\infty}(-y)^k \qquad \forall y \in \mathbb{C} : |y| \leqslant 1/2$$

to see

$$|g'(y) - 1| = \Big|\sum_{k=1}^{\infty}(-y)^k\Big| = \Big| - y\sum_{k=0}^{\infty}(-y)^k\Big| = \Big|\frac{y}{1+y}\Big| \leqslant 2|y|.$$

Note that if the indicator $\mathbb{1}_{\{|\widehat{\varphi}_{\delta,n}(u)| \geqslant T^{-1/2}\}}$ in the definition of $\widehat{\Delta\psi_n}$ equals 1, then $\widehat{\Delta\psi_n} - \Delta\psi = \delta^{-1}\Delta\log(\widehat{\varphi}_{\delta,n}/\varphi_\delta)$. Therefore, (5.24) implies on the event $\Omega_n := \Omega_{n,1} \cap \Omega_{n,2}$ with $\Omega_{n,1} := \big\{\inf_{|u|_\infty \leqslant h^{-1}} |\widehat{\varphi}_{\delta,n}(u)| \geqslant T^{-1/2}\big\}$ and $\Omega_{n,2} := \big\{\sup_{|u|_\infty \leqslant h^{-1}} |\xi(u)| \leqslant 1/2\big\}$ that

$$\sup_{|u|_\infty \leqslant h^{-1}} \big|\delta(\widehat{\Delta\psi_n} - \Delta\psi)(u) - \Delta((\widehat{\varphi}_{\delta,n} - \varphi_\delta)/\varphi_\delta)(u)\big| \lesssim \||\nabla\xi|\|^2_{L^\infty(I_h)} + \|\xi\|_{L^\infty(I_h)}\|\Delta\xi\|_{L^\infty(I_h)}.$$

To control the ξ-terms, we invoke Proposition 5.5 applied to the increments of the Lévy process with $\rho = \delta$ after verifying that the moments are of the appropriate order. Owing to the equivalence of norms, it is sufficient to show that with $\tau = m - 4 > 0$

$$\mathbb{E}[|Y_{1,k}|^{2l+\tau}] \lesssim \delta^{l \wedge 1} \qquad \text{and} \qquad \mathbb{E}[|Y_{1,k}|^{2l}] \lesssim \delta^{l \wedge 1}, \qquad k = 1, \dots, d,\ l = 0, 1, 2, \tag{5.25}$$

where $Y_{1,k}$ is the k-th entry of Y_1 and thus an increment with time difference δ based on the Lévy process $(L_{t,k})_{t \geqslant 0}$ with Lévy measure ν_k. For $l = 1, 2$, it follows from Figueroa-López (2008,

Theorem 1.1) that

$$\lim_{\delta \searrow 0} \delta^{-1}\mathbb{E}[|Y_{1,k}|^{2l+\tau}] = \lim_{\delta \searrow 0} \delta^{-1}\mathbb{E}[|L_{\delta,k}|^{2l+\tau}] = \int |x_k|^{2l+\tau}\, \nu_k(\mathrm{d}x_k) \leqslant \int |x|^{2l+\tau}\, \nu(\mathrm{d}x) \lesssim C.$$

For $l = 0$, $\mathbb{E}[|Y_{1,k}|^{\tau}] \lesssim \mathbb{E}[|Y_{1,k}|^{m}] \lesssim 1$ holds by our moment assumptions. The second condition in (5.25) was already checked at the beginning of Section 5.3.1.1.

Therefore, $|\Delta\psi(u)| \lesssim 1$ yields

$$\begin{aligned}
\|\xi\|_{L^\infty(I_h)} &= \mathcal{O}_{\mathbb{P}}\big(n^{-1/2}(\log h^{-1})^{(1+\chi)/2}\|\varphi_\delta^{-1}\|_{L^\infty(I_h)}\big), \\
\||\nabla\xi|\|^2_{L^\infty(I_h)} &= \mathcal{O}_{\mathbb{P}}\big(n^{-1}(\log h^{-1})^{1+\chi}\|\varphi_\delta^{-1}\|^2_{L^\infty(I_h)}\big(\delta + \delta^2\||\nabla\psi|\|^2_{L^\infty(I_h)}\big)\big), \\
\|\Delta\xi\|_{L^\infty(I_h)} &= \mathcal{O}_{\mathbb{P}}\Big(n^{-1/2}(\log h^{-1})^{(1+\chi)/2}\|\varphi_\delta^{-1}\|_{L^\infty(I_h)} \\
&\qquad \cdot \big(\delta^{1/2} + \delta^{3/2}\big\||\nabla\psi|\big\|_{L^\infty(I_h)} + \delta^2\big\||\nabla\psi|\big\|^2_{L^\infty(I_h)}\big)\Big).
\end{aligned} \tag{5.26}$$

Combining (5.26) with $n^{-1/2}(\log h^{-1})^{(1+\chi)/2}\|\varphi_\delta^{-1}\|_{L^\infty(I_h)} \to 0$ gives $\mathbb{P}(\Omega_{n,2}) \to 1$. As discussed after Proposition 5.5, we also have $\mathbb{P}(\Omega_{n,1}) \to 1$ and therefore $\mathbb{P}(\Omega_n) \to 1$, which completes the proof. □

5.3.4.3 Proof of Lemma 5.7

For fixed $x \in \mathbb{R}^d$, we want to apply Bernstein's inequality to

$$M^{\nu}_{\delta,n}(x) = -\sum_{l=1}^{n}\big(\xi_l - \mathbb{E}[\xi_l]\big) \qquad \text{with} \qquad \xi_l \coloneqq T^{-1}\mathscr{F}^{-1}\big[m_{\delta,h}(u)|Y_l|^2\mathrm{e}^{\mathrm{i}\langle u, Y_l\rangle}\big](x).$$

Similar arguments to (5.15) reveal $\big\|m_{\delta,h}(u)\big\|_{L^1} \lesssim h^{-\delta\alpha-d}$, and with the quotient rule one finds the same order for $\big\|\Delta m_{\delta,h}(u)\big\|_{L^1}$ and $\big\||\nabla m_{\delta,h}|\big\|_{L^1}$ paving a deterministic bound of ξ_l via

$$\begin{aligned}
|\xi_l| &= T^{-1}|Y_l|^2\big|\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](-Y_l)\big| \\
&= T^{-1}\big|\mathscr{F}^{-1}\big[\Delta\big(m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big)\big](-Y_l)\big| \\
&\leqslant T^{-1}\big\|\Delta\big(m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big)\big\|_{L^1} \\
&\leqslant T^{-1}\big(\big\|\Delta m_{\delta,h}(u)\big\|_{L^1} + 2|x|_1\big\||\nabla m_{\delta,h}|\big\|_{L^1} + |x|^2\|m_{\delta,h}\|_{L^1}\big) \\
&\lesssim T^{-1}(1+|x|^2)h^{-\delta\alpha-d}.
\end{aligned} \tag{5.27}$$

To bound the variance of ξ_l, note that for the distribution $\mathbb{P}_\delta$ of Y_1, we have

$$\mathscr{F}[\mathrm{i}z_k\mathbb{P}_\delta] = \frac{\partial\varphi_\delta}{\partial u_k} = \delta\varphi_\delta\frac{\partial\psi}{\partial u_k} = \delta\mathscr{F}[\mathrm{i}z_k\nu]\varphi_\delta = \mathscr{F}[\mathscr{F}^{-1}[\delta\mathscr{F}[\mathrm{i}z_k\nu]\varphi_\delta]] = \mathscr{F}[(\delta\mathrm{i}z_k\nu) * \mathbb{P}_\delta]$$

and therefore $z_k\mathbb{P}_\delta = \delta\mu * \mathbb{P}_\delta$ with $\mu(\mathrm{d}z) = z_k\nu(z)\mathrm{d}z$. It follows that

$$\int g(z)|z_k|\,\mathbb{P}_\delta(\mathrm{d}z) \leqslant \delta\|z_k\nu\|_\infty\|g\|_{L^1}, \qquad \forall g \in L^1(\mathbb{R}^d).$$

Again, using similar arguments to (5.15) and the quotient rule, we also have $\big\|\Delta m_{\delta,h}(u)\big\|_{L^2}, \big\||\nabla m_{\delta,h}|\big\|_{L^2} \lesssim h^{-\delta\alpha-d/2}$. Thus, the Cauchy-Schwarz inequality and the Plancherel theorem imply

$$\begin{aligned}
\mathrm{Var}(\xi_l) \leqslant \mathbb{E}[|\xi_l|^2] &= T^{-2}\mathbb{E}\Big[|Y_l|^4\big|\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](-Y_l)\big|^2\Big] \\
&\lesssim T^{-2}\sum_{k=1}^d\int|y|^3\big|\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](-y)\big|^2|y_k|\,\mathbb{P}_\delta(\mathrm{d}y) \\
&\leqslant n^{-2}\delta^{-1}\sum_{k=1}^d\|z_k\nu\|_\infty\int|y|^3\big|\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](y)\big|^2\,\mathrm{d}y \\
&\lesssim n^{-2}\delta^{-1}\big\||y|^2\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](y)\big\|_{L^2}\big\||y|_1\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](y)\big\|_{L^2} \\
&\lesssim n^{-2}\delta^{-1}\Big(\sum_{k=1}^d\Big\|\frac{\partial^2}{\partial u_k^2}\big(m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big)\Big\|_{L^2}\Big)\Big(\sum_{k=1}^d\Big\|\frac{\partial}{\partial u_k}\big(m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big)\Big\|_{L^2}\Big) \\
&\lesssim n^{-2}\delta^{-1}h^{-2\delta\alpha-d}(1+|x|^3).
\end{aligned}$$

Now, Bernstein's inequality, e.g. van der Vaart (1998, Lemma 19.32) yields for a constant $c' > 0$ and any $\kappa > 0$ that

$$\mathbb{P}\big(|M_{\delta,n}^\nu(x)| \geqslant \kappa\big) \leqslant 2\exp\Big(-\frac{Tc'\kappa^2}{h^{-2\delta\alpha-d}(1+|x|^3)+\kappa(1+|x|^2)h^{-\delta\alpha-d}}\Big),$$

which reads as the assertion if we choose $\kappa = \kappa_0 T^{-1/2}h^{-\delta\alpha-d/2}$ for any $\kappa_0 > 0$ and set $c = c'/2$.□

5.3.4.4 Proof of Lemma 5.8

Fix $x = (x^{(1)}, x^{(2)})$ for $x^{(1)}, x^{(2)} \in \mathbb{R}^{d/2}$ and analogously split Y_l into its first and last $d/2$ entries $Y_l^{(1)}$ and $Y_l^{(2)}$ with characteristic functions $\varphi_{\delta,1}$ and $\varphi_{\delta,2}$, respectively. Due to the product kernel, we obtain

$$\xi_l = T^{-1}|Y_l|^2\big|\mathscr{F}^{-1}\big[m_{\delta,h}(u)\mathrm{e}^{-\mathrm{i}\langle u,x\rangle}\big](-Y_l)\big| = T^{-1}(A_1B_2 + A_2B_1)$$

with

$$A_k \coloneqq |Y_l^{(k)}|^2 \mathscr{F}^{-1}\Big[m_{\delta,h}^{(k)}(u^{(k)})\mathrm{e}^{\mathrm{i}\langle u^{(k)},Y_l^{(k)}\rangle}\Big](x^{(k)}), \quad B_k \coloneqq \mathscr{F}^{-1}\Big[m_{\delta,h}^{(k)}(u^{(k)})\mathrm{e}^{\mathrm{i}\langle u^{(k)},Y_l^{(k)}\rangle}\Big](x^{(k)}), \quad \text{and}$$

$$m_{\delta,h}^{(k)}(u^{(k)}) \coloneqq \varphi_{\delta,k}^{-1}(u^{(k)}) \prod_{j=1+(k-1)d/2}^{kd/2} \mathscr{F}K^j(hu_j^{(k)}), \qquad k=1,2.$$

A_1 and A_2 are the same terms that appeared in the proof of Lemma 5.7, see (5.27), just with half the dimension and therefore

$$\begin{aligned} |A_k| &\lesssim \|\varphi_{\delta,k}^{-1}\|_{L^\infty([-h^{-1},h^{-1}]^{d/2})} h^{-d/2}(1+|x^{(k)}|^2), \\ \mathbb{E}[|A_k|^2] &\lesssim \delta \|\varphi_{\delta,k}^{-1}\|^2_{L^\infty([-h^{-1},h^{-1}]^{d/2})} h^{-d/2}(1+|x^{(k)}|^3). \end{aligned}$$

In a similar vein, $m_{\delta,h}^{(1)}$ and $m_{\delta,h}^{(2)}$ can be treated like $m_{\delta,h}$ with half the dimension leading to

$$|B_k| \lesssim \|m_{\delta,h}^{(k)}\|_{L^1(\mathbb{R}^{d/2})} \lesssim \|\varphi_{\delta,k}^{-1}\|_{L^\infty([-h^{-1},h^{-1}]^{d/2})} h^{-d/2}.$$

Note that

$$\prod_{k=1}^{2} \|\varphi_{\delta,k}^{-1}\|_{L^\infty([-h^{-1},h^{-1}]^{d/2})} = \|\varphi_\delta^{-1}\|_{L^\infty(I_h)}.$$

Together, we have the deterministic bound $|\xi_l| \lesssim T^{-1}h^{-\delta\alpha-d}(1+|x|^2)$. Further, since A_1 and B_2 as well as A_2 and B_1 are independent, we obtain

$$\begin{aligned} \mathrm{Var}(\xi_l) \lesssim T^{-2}(\mathrm{Var}(A_1B_2) + \mathrm{Var}(A_2B_1)) &\leqslant T^{-2}\big(\mathbb{E}[|A_1|^2]\mathbb{E}[|B_2|^2] + \mathbb{E}[|A_2|^2]\mathbb{E}[|B_1|^2]\big) \\ &\lesssim n^{-2}\delta^{-1}h^{-2\delta\alpha-3d/2}(1+|x|^3). \end{aligned}$$

Overall, Bernstein's inequality, see e.g. the aforementioned van der Vaart (1998, Lemma 19.32), gives for a constant $c'>0$ and any $\kappa>0$ that

$$\mathbb{P}\big(|M_{\delta,n}^\nu(x)| \geqslant \kappa\big) \leqslant 2\exp\Big(-\frac{Tc'\kappa^2}{h^{-2\delta\alpha-d}(1+|x|^3)+\kappa(1+|x|^2)h^{-\delta\alpha-d}}\Big).$$

The assertion follows by choosing $\kappa = \kappa_0 T^{-1/2}h^{-\delta\alpha-3d/4}$ for any $\kappa_0>0$ and setting $c=c'/2$. □

6 Outlook

Before providing an outlook on future research, we present a short summary of the thesis.

Whenever the evolution of a time dependent system is influenced by random phenomena, stochastic processes are used for mathematical modeling. Since an important feature of data sets in modern applications is high dimensionality, statistical methods for stochastic processes are required which are custom-tailored for high-dimensional data. Moreover, it is of utmost importance to be able to estimate the underlying uncertainties especially in applications from natural science. In this thesis, we presented novel regression approaches which allow us to circumvent the curse of dimensionality via dimension reduction techniques in general, and via deep neural networks in particular. A focus was on Bayes-type methods which also allow for uncertainty quantification. To achieve a scalable MCMC method, a corrected stochastic MH algorithm was proposed. This method is computationally feasible for large samples and it satisfies an optimal bound for the prediction risk as well as uncertainty guarantees.

Additionally, the estimation of the jump distribution of multi-dimensional Lévy process was studied. Based on discrete observations the so-called Lévy density was estimated via a spectral approach. Allowing for low- and high-frequency observations, rates of convergence were proved and numerical experiments confirmed the theoretical findings. The proposed method is robust to various dependence structures which may lead to singular jump distributions.

A relevant question for future research is whether the developed regression methods can be combined with the spectral method to estimate the Lévy density of a multivariate, or even a high-dimensional, Lévy process. We provide an outlook on how to possibly approach this question.

By the Lévy-Itô decomposition, see Sato (1999), we can rewrite any Lévy process $(X_t)_{t\geqslant 0}$ as

$$X_t = t\gamma_0 + \Sigma^{1/2} B_t + L_t, \qquad t \geqslant 0$$

with $\gamma_0 \in \mathbb{R}^p$, positive semi-definite $\Sigma \in \mathbb{R}^{p\times p}$, a standard Brownian motion $(B_t)_{t\geqslant 0}$ in $\mathbb{R}^p$, and a pure jump Lévy process $(L_t)_{t\geqslant 0}$ with Lévy measure ν_L. The Lévy process $(X_t)_{t\geqslant 0}$ is observed on an equidistant time grid $\delta, 2\delta, \ldots, n\delta =: T$ with time difference $\delta > 0$. Equivalently, we

observe i.i.d. data $\mathcal{D}_n \coloneqq (Y_j)_{j=1,\dots,n}$ of increments

$$Y_j = X_{j\delta} - X_{(j-1)\delta} \overset{d}{=} X_\delta = \delta\gamma_0 + \Sigma^{1/2} B_\delta + L_\delta, \qquad j = 1, \dots, n.$$

Recall from (5.1) that, under the assumption $\int |x|^2 \, \nu_L(\mathrm{d}x) < \infty$, the characteristic function of Y_j is given by

$$\varphi_\delta(u) \coloneqq \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle u, X_\delta\rangle}] = \mathrm{e}^{\delta\psi(u)} \quad \text{with} \quad \psi(u) = \mathrm{i}\langle u, \gamma\rangle - \frac{1}{2}\langle u, \Sigma u\rangle + \int \left(\mathrm{e}^{\mathrm{i}\langle u,x\rangle} - 1 - \mathrm{i}\langle u, x\rangle\right) \nu_L(\mathrm{d}x),$$

where $\gamma = \gamma_0 + \int_{|x|>1} x \, \nu_L(\mathrm{d}x)$. As in Chapter 5, we can estimate φ_δ with the empirical characteristic function $\widehat{\varphi}_{\delta,n}(u) = \frac{1}{n}\sum_{j=1}^n \mathrm{e}^{\mathrm{i}\langle u, Y_j\rangle}$. In view of (5.3), we again estimate $\Delta\psi$ by

$$\widehat{\Delta\psi}(u) \coloneqq \frac{\widehat{\varphi}_{\delta,n}(u)\Delta\widehat{\varphi}_{\delta,n}(u) - (\nabla\widehat{\varphi}_{\delta,n}(u))^2}{\delta\widehat{\varphi}_{\delta,n}(u)} \mathbb{1}_{\{|\widehat{\varphi}_{\delta,n}(u)| \geqslant T^{-1/2}\}}.$$

As discussed after Proposition 5.5 and in the proof of Lemma 5.6, see Section 5.3.4.2, the indicator equals 1 with probability converging to 1, and if the indicator equals 1, we have $\widehat{\Delta\psi_n} - \Delta\psi = \delta^{-1}\log\left(\widehat{\varphi}_{\delta,n}/\varphi_\delta\right)$. Therefore,

$$\begin{aligned}
\widehat{\Delta\psi_n}(u) &= -\operatorname{tr}(\Sigma) - \mathscr{F}[|\cdot|^2\nu_L](u) + \frac{1}{\delta}\Delta\log\Big(\frac{\widehat{\varphi}_{\delta,n}(u)}{\varphi_\delta(u)}\Big) \\
&\approx -\mathscr{F}[|\cdot|^2\nu_L](u) - \operatorname{tr}(\Sigma) + \frac{1}{\delta}\Delta\Big(\frac{\widehat{\varphi}_{\delta,n}(u) - \varphi_\delta(u)}{\varphi_\delta(u)}\Big) \\
&\eqqcolon f(u) - \operatorname{tr}(\Sigma) + \frac{1}{\delta}\Delta\Big(\frac{\widehat{\varphi}_{\delta,n}(u) - \varphi_\delta(u)}{\varphi_\delta(u)}\Big),
\end{aligned}$$

where we linearized the logarithm as in (5.9). This formula can be interpreted as a regression model with the following differences to the nonparametric regression (2.1) considered in this thesis: First, instead of observing an i.i.d. sample of size n, we have observations for every $u \in \mathbb{R}^p$. Second, there is an additional nuisance term $-\operatorname{tr}(\Sigma)$. Third, the estimation problem is ill-posed because the stochastic error will explode for large frequencies owing to $\varphi_\delta(u) \to 0$ for $|u| \to \infty$. Since the linearization and the stochastic error are as in Chapter 5, we can use our insights from the application of the spectral method.

In order to apply our regression techniques, we have to treat the nuisance term $-\operatorname{tr}(\Sigma)$, which can be done as proposed in (5.7). With this in mind, we now omit the nuisance term, i.e., we consider a pure jump Lévy process for clarity in the following arguments.

To estimate f, we introduce a class $\mathcal{F} = \{f_\vartheta : \vartheta \in \Theta\}$ of parameterized functions with parameter set Θ. Under the additional assumption $\int |x|^{2+s} \, \nu_L(\mathrm{d}x) < \infty$ for some $s > 0$, the function $f = -\mathscr{F}[|\cdot|^2\nu_L]$ is s-Hölder regular. Hence, we could choose $\mathcal{F}$ as the class of ReLU neural

networks from Chapter 3 and again use approximation results by Schmidt-Hieber (2020) or Kohler & Langer (2021). Imposing a prior Π on Θ, we would then consider the Gibbs posterior distribution with Π-density

$$\frac{\mathrm{d}\Pi_\lambda(\vartheta \mid \mathcal{D}_n)}{\mathrm{d}\Pi} \propto \exp\big(-\lambda R_n(\vartheta)\big),$$

where the empirical risk is defined as $R_n(\vartheta) \coloneqq \|\widehat{\Delta\psi_n} - f_\vartheta\|_{L^2(U)}$ with a compact set $U \subseteq \mathbb{R}^d$. The set U can be chosen to cut off the problematic large frequencies in a way which depends on the sample size, similarly to a bandwidth limited kernel. This choice of the risk is similar to the empirical study by Xu & Darve (2020) in the one-dimensional case. The differences are that they estimate the characteristic exponent ψ instead of its Laplacian, and they evaluate the risk on random frequencies uniformly drawn from some bounded interval.

A main challenge in applying the methodology from Section 2.1 will be the derivation of a concentration-type inequality since the empirical risk is no longer an average of i.i.d. random variables. Based on the resulting estimator

$$\widehat{f}_\lambda = f_{\widehat{\vartheta}_\lambda}, \qquad \text{where} \qquad \widehat{\vartheta}_\lambda \mid \mathcal{D}_n \sim \Pi_\lambda(\cdot \mid \mathcal{D}_n)$$

for f, we would then need to recover the corresponding estimator of the Lévy measure ν_L, which can be done similarly to Chapter 5. Cutting off the high-frequencies, we propose to estimate ν_L by

$$\widehat{\nu}_L(x) = |x|^{-2}\mathscr{F}^{-1}\big[\widehat{f}_\lambda \mathbb{1}_U\big](x), \qquad x \in \mathbb{R}^p \setminus \{0\}.$$

To adapt the overall approach to high-dimensional Lévy processes, we could introduce a notion of sparsity into the model by assuming that $L_t = \Lambda^\top Z_t$ with an $\mathbb{R}^d$-valued pure jump Lévy process $(Z_t)_{t\geqslant 0}$, for d substantially smaller than p, and a sparse matrix $\Lambda \in \mathbb{R}^{d\times p}$. Denote the Lévy measure of Z by ν. Using the explicit relationship between ν and ν_L, see Cont & Tankov (2004, Theorem 4.1), we recover a multi-index structure

$$f(u) = -\int_{\mathbb{R}^p} |x|^2 \mathrm{e}^{\mathrm{i}\langle u,x\rangle}\, \nu_L(\mathrm{d}x) = -\int_{\mathbb{R}^d} |x|^2 \mathrm{e}^{\mathrm{i}\langle \Lambda u,x\rangle}\, \nu(\mathrm{d}x) = g(\Lambda u), \qquad u \in \mathbb{R}^p$$

with dimension reduction matrix Λ and link function $g(u) = -\int_{\mathbb{R}^d} |x|^2 \mathrm{e}^{\mathrm{i}\langle u,x\rangle}\, \nu(\mathrm{d}x)$, $u \in \mathbb{R}^d$. Hence, it is evident that we can exploit the reduced effective dimension of the model by estimating this multi-index model, either directly, or, in view of Corollary 4.6, with a sparse stochastic neural network using the methodology developed in this thesis. Overall, the combination of our regression techniques with our spectral based estimator for a multivariate Lévy density looks like a promising approach to the estimation of a high-dimensional Lévy density.

Bibliography

Aït-Sahalia, Y. & Jacod, J. (2012). Analyzing the spectrum of asset returns: Jump and volatility components in high frequency data. *Journal of Economic Literature*, 50(4), 1007–50.

Alexos, A., Boyd, A. J., & Mandt, S. (2022). Structured stochastic gradient MCMC. In *International Conference on Machine Learning* (pp. 414–434).

Alquier, P. (2021). User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*.

Alquier, P. & Biau, G. (2013). Sparse single-index model. *Journal of Machine Learning Research*, 14, 243–280.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2), 697–725.

Anthony, M. & Bartlett, P. L. (1999). *Neural network learning: Theoretical foundations*. Cambridge University Press.

Audibert, J.-Y. (2004). Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, 40(6), 685–736.

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4), 1591–1646.

Audibert, J.-Y. & Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, 39(5), 2766–2794.

Bardenet, R., Doucet, A., & Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47), 1–43.

Bauer, B. & Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4), 2261–2285.

Belomestny, D. (2010). Spectral estimation of the fractional order of a Lévy process. *The Annals of Statistics*, 38(1), 317–351.

Belomestny, D. (2011). Statistical inference for time-changed Lévy processes via composite characteristic function estimation. *The Annals of Statistics*, 39(4), 2205–2242.

Belomestny, D., Gugushvili, S., Schauer, M., & Spreij, P. (2022). Nonparametric Bayesian volatility estimation for gamma-driven stochastic differential equations. *Bernoulli*, 28(4), 2151–2180.

Belomestny, D. & Reiß, M. (2006). Spectral calibration of exponential Lévy models. *Finance and Stochastics*, 10(4), 449–474.

Belomestny, D. & Reiß, M. (2015). Estimation and calibration of Lévy models via Fourier methods, In: *Lévy Matters IV - Estimation for Discretely Observed Lévy Processes.* (pp. 1–76).

Belomestny, D. & Trabs, M. (2018). Low-rank diffusion matrix estimation for high-dimensional time-changed Lévy processes. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 54(3), 1583–1621.

Belomestny, D., Trabs, M., & Tsybakov, A. B. (2019). Sparse covariance matrix estimation in high-dimensional deconvolution. *Bernoulli*, 25(3), 1901–1938.

Besag, J. (1994). Comments on "Representations of knowledge in complex systems" by U. Grenander and M.I. Miller. *Journal of the Royal Statistical Society. Series B. Methodological*, 56(4), 549–581.

Bieringer, S., Kasieczka, G., Steffen, M. F., & Trabs, M. (2023). Statistical guarantees for stochastic Metropolis-Hastings. *arXiv preprint arXiv:2310.09335.*

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 78(5), 1103–1130.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In *International Conference on Machine Learning* (pp. 1613–1622).

Bücher, A. & Vetter, M. (2013). Nonparametric inference on Lévy measures and copulas. *The Annals of Statistics*, 41(3), 1485–1515.

Castillo, I. & Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics*, 42(5), 1941–1969.

Castillo, I., Schmidt-Hieber, J., & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5), 1986–2018.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization.* Springer.

Catoni, O. (2007). *PAC-Bayesian supervised classification: The thermodynamics of statistical learning*, volume 56 of *Lecture Notes-Monograph Series.* Institute of Mathematical Statistics.

Cheng, X. & Bartlett, P. (2018). Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83 (pp. 186–211).

Chérief-Abdellatif, B.-E. (2020). Convergence rates of variational inference in sparse deep learning. In *International Conference on Machine Learning* (pp. 1831–1842).

Cobb, A. D. & Jalaian, B. (2021). Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting. *Uncertainty in Artificial Intelligence.*

Cont, R. & Tankov, P. (2004). *Financial modelling with jump processes.* Chapman & Hall/CRC.

Dalalyan, A. S., Juditsky, A., & Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9, 1647–1678.

Dalalyan, A. S. & Riou-Durand, L. (2020). On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3), 1956–1988.

Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Deng, W., Feng, Q., Gao, L., Liang, F., & Lin, G. (2020a). Non-convex learning via replica exchange stochastic gradient MCMC. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research* (pp. 2474–2483).

Deng, W., Liang, S., Hao, B., Lin, G., & Liang, F. (2022). Interacting contour stochastic gradient Langevin dynamics. In *International Conference on Learning Representations.*

Deng, W., Lin, G., & Liang, F. (2020b). A contour stochastic gradient Langevin dynamics algorithm for simulations of multi-modal distributions. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems.*

DeVore, R., Hanin, B., & Petrova, G. (2021). Neural network approximation. *Acta Numerica*, 30, 327–444.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics letters B*, 195(2), 216–222.

Duval, C. & Mariucci, E. (2021). Spectral-free estimation of Lévy densities in high-frequency regime. *Bernoulli*, 27(4), 2649–2674.

Dziugaite, G. K. & Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3), 1257–1272.

Figueroa-López, J. E. (2008). Small-time moment asymptotics for Lévy processes. *Statistics & Probability Letters*, 78(18), 3355–3365.

Figueroa-López, J. E. (2011). Sieve-based confidence intervals and bands for Lévy densities. *Bernoulli*, 17(2), 643–670.

Franssen, S. & Szabó, B. (2022). Uncertainty quantification for nonparametric regression using empirical Bayesian neural networks. *arXiv preprint arXiv:2204.12735*.

Freund, Y., Ma, Y.-A., & Zhang, T. (2022). When is the convergence time of Langevin algorithms dimension independent? A composite optimization viewpoint. *Journal of Machine Learning Research*, 23, 1–32.

Gegler, A. & Stadtmüller, U. (2010). Estimation of the characteristics of a Lévy process. *Journal of Statistical Planning and Inference*, 140(6), 1481–1496.

Ghosal, S. & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Giné, E. & Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Number 40 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graves, A. (2011). Practical variational inference for neural networks. *Advances in neural information processing systems*, 24.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*.

Guedj, B. & Alquier, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7, 264–291.

Gugushvili, S. (2012). Nonparametric inference for discretely sampled Lévy processes. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 48(1), 282–307.

Hoffmann, M. & Nickl, R. (2011). On adaptive inference and confidence bands. *The Annals of Statistics*, 39(5), 2383–2409.

Hristache, M., Juditsky, A., Polzehl, J., & Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6), 1537–1566.

Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31.

Kappus, J. (2012). *Nonparametric adaptive estimation for discretely observed Lévy processes*. Dissertation, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II.

Kappus, J. & Reiß, M. (2010). Estimation of the characteristics of a Lévy process observed at arbitrary frequency. *Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research*, 64(3), 314–328.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klock, T., Lanteri, A., & Vigogna, S. (2021). Estimating multi-index models with response-conditional least squares. *Electronic Journal of Statistics*, 15(1), 589–629.

Knapik, B. T., van der Vaart, A. W., & van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5), 2626–2657.

Kohler, M. & Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4), 2231–2249.

Li, C., Chen, C., Carlson, D. E., & Carin, L. (2016). Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 1788–1794).

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 316–327.

Maclaurin, D. & Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer.

McAllester, D. A. (1999a). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory* (pp. 164–170).

McAllester, D. A. (1999b). Some PAC-Bayesian theorems. *Machine Learning*, 37(3), 355–363.

McMurry, T. L. & Politis, D. N. (2004). Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics*, 16(3-4), 549–562.

Mies, F. (2020). Rate-optimal estimation of the Blumenthal-Getoor index of a Lévy process. *Electronic Journal of Statistics*, 14(2), 4165–4206.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics, In: *Handbook of Markov chain Monte Carlo*. (pp. 113–163).

Neumann, M. H. & Reiß, M. (2009). Nonparametric estimation for Lévy processes from low-frequency observations. *Bernoulli*, 15(1), 223–248.

Nickl, R. & Reiß, M. (2012). A Donsker theorem for Lévy measures. *Journal of Functional Analysis*, 263(10), 3306–3332.

Nickl, R., Reiß, M., Söhl, J., & Trabs, M. (2016). High-frequency Donsker theorems for Lévy measures. *Probability Theory and Related Fields*, 164(1-2), 61–108.

Nickl, R. & Wang, S. (2022). On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *Journal of the European Mathematical Society*.

Papagiannouli, K. (2020). Minimax rates for the covariance estimation of multi-dimensional Lévy processes with high-frequency data. *Electronic Journal of Statistics*, 14(2), 3525–3562.

Patterson, S. & Teh, Y. W. (2013). Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems 26* (pp. 3102–3110).

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., & Szepesvári, C. (2021). Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 1–40.

Polson, N. G. & Ročková, V. (2018). Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, 31, 938–949.

Ray, K. & Szabó, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539), 1270–1281.

Reiß, M. (2013). Testing the characteristics of a Lévy process. *Stochastic Processes and their Applications*, 123(7), 2808–2828.

Robert, C. P. & Casella, G. (2004). *Monte Carlo statistical methods*. Springer, second edition.

Roberts, G. O. & Tweedie, R. L. (1996a). Exponential convergence of of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4), 341–363.

Roberts, G. O. & Tweedie, R. L. (1996b). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1), 95–110.

Rousseau, J. & Szabó, B. (2020). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *The Annals of Statistics*, 48(4), 2155–2179.

Sato, K.-I. (1999). *Lévy processses and infinitely divisible distributions*. Cambridge University Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85–117.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 1875–1897.

Steffen, M. F. (2023a). Estimating a multivariate Lévy density based on discrete observations. *arXiv preprint arXiv:2305.14315*.

Steffen, M. F. (2023b). PAC-Bayes bounds for high-dimensional multi-index models with unknown active dimension. *arXiv preprint arXiv:2303.13474*.

Steffen, M. F. & Trabs, M. (2023). A PAC-Bayes oracle inequality for sparse neural networks. *arXiv preprint arXiv:2204.12392*.

Szabó, B., van der Vaart, A. W., & van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4), 1391–1428.

Taheri, M., Xie, F., & Lederer, J. (2021). Statistical guarantees for regularized neural networks. *Neural Networks*, 142, 148–161.

Trabs, M. (2015). Quantile estimation for Lévy measures. *Stochastic Processes and their Applications*, 125(9), 3484–3521.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van de Geer, S., Bühlmann, P., & Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5, 688–749.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Welling, M. & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning* (pp. 681–688).

Woyczyński, W. A. (2001). *Lévy processes in the physical sciences*. Springer.

Wu, T.-Y., Rachel Wang, Y. X., & Wong, W. H. (2022). Mini-batch Metropolis-Hastings with reversible SGLD proposal. *Journal of the American Statistical Association*, 117(537), 386–394.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35(6), 2654–2690.

Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484), 1631–1640.

Xia, Y., Tong, H., Li, W. K., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 64(3), 363–410.

Xu, K. & Darve, E. (2020). Calibrating multivariate Lévy processes with neural networks. volume 107 of *Proceedings of Machine Learning Research* (pp. 207–220).

Yang, Z., Balasubramanian, K., Wang, Z., & Liu, H. (2017). Learning non-Gaussian multi-index model via second-order Stein's method. *Advances in Neural Information Processing Systems*, 30, 6097–6106.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 103–114.

Zhang, A. Y. & Zhou, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics*, 48(5), 2575–2598.

Zhang, F. & Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4), 2180–2207.

Zhang, R., Li, C., Zhang, J., Chen, C., & Wilson, A. G. (2020). Cyclical stochastic gradient MCMC for Bayesian deep learning. In *8th International Conference on Learning Representations*.

Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474), 630–643.