Journal of Machine Learning Research 27 (2026) 1-50 Submitted 5/22; Revised 2/26; Published 2/26

The surrogate Gibbs-posterior of a corrected stochastic MALA: Towards uncertainty quantification for neural networks

Sebastian Bieringer SEBASTIAN.GUIDO.BIERINGER@UNI-HAMBURG.DE
Universität Hamburg
Institute of Experimental Physics
Luruper Chaussee 149, D-22761 Hamburg, Germany

Gregor Kasieczka GREGOR.KASIECZKA@UNI-HAMBURG.DE
Universität Hamburg
Institute of Experimental Physics
Luruper Chaussee 149, D-22761 Hamburg, Germany

Maximilian F. Steffen MAXIMILIAN.STEFFEN@KIT.EDU
Karlsruhe Institute of Technology
Institute of Stochastics
Englerstr. 2, D-76131 Karlsruhe, Germany

Mathias Trabs MATHIAS.TRABS@KIT.EDU
Karlsruhe Institute of Technology
Institute of Stochastics
Englerstr. 2, D-76131 Karlsruhe, Germany

Editor: Pierre Alquier

Abstract

MALA is a popular gradient-based Markov chain Monte Carlo method to access the Gibbs-posterior distribution. Stochastic MALA (sMALA) scales to large data sets, but changes the target distribution from the Gibbs-posterior to a surrogate posterior which only exploits a reduced sample size. We introduce a corrected stochastic MALA (csMALA) with a simple correction term for which distance between the resulting surrogate posterior and the original Gibbs-posterior decreases in the full sample size while retaining scalability. In a nonparametric regression model, we prove a PAC-Bayes oracle inequality for the surrogate posterior. Uncertainties can be quantified by sampling from the surrogate posterior. Focusing on Bayesian neural networks, we analyze the diameter and coverage of credible balls for shallow neural networks and we show optimal contraction rates for deep neural networks. Our credibility result is independent of the correction and can also be applied to the standard Gibbs-posterior. A simulation study in a high-dimensional parameter space demonstrates that an estimator drawn from csMALA based on its surrogate Gibbs-posterior indeed exhibits these advantages in practice.

Keywords: Gibbs-posterior, Stochastic neural network, optimal contraction rate, credible sets, oracle inequality

1. Introduction

An essential feature in modern data science, especially in machine learning as well as high-dimensional statistics, are large sample sizes and large parameter space dimensions. As a consequence, the design of methods for uncertainty quantification is characterized by a tension between numerically feasible and efficient algorithms and approaches which satisfy theoretically justified statistical properties. Motivated by this tension, we introduce a simple correction to stochastic MALA to achieve both: the method is scalable, *i.e.*, it 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. While the focus of this work is on the statistical behavior of the posterior distribution of the corrected stochastic MALA (csMALA), our simulation study demonstrates the practical benefits of our algorithm.

Bayesian methods enjoy high popularity for quantifying uncertainties in complex models. The classical approach to sample from the posterior distribution are Markov Chain Monte Carlo methods (MCMC). For large parameter spaces gradient-based Monte Carlo methods are particularly useful with, *e.g.*, Langevin dynamics serving as a prototypical example. State-of-the-art methods such as Metropolis adjusted Langevin (MALA) Besag (1994); Roberts and Tweedie (1996b) and Hamiltonian Monte Carlo Duane et al. (1987); Neal (2011) equip 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 costs 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 standard in empirical risk minimization and has been successfully applied for Langevin dynamics as well (Alexos et al., 2022; Li et al., 2016; Patterson and Teh, 2013; Welling and Teh, 2011). In this case, the MH steps remain as a computational bottleneck: Since the target distribution depends on the full dataset, 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 of the resulting *stochastic MALA* (sMALA) considerably, see Wu et al. (2022).

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 Metropolis-Hastings perspective by Andrieu and Roberts (2009) and Maclaurin and 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 to 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. In combination with the MALA methodology and stochastic gradients in the proposal distribution we obtain our *corrected stochastic MALA* (csMALA).

In a nonparametric regression problem under a quadratic loss and under classical assumptions, we investigate the distance of the surrogate posterior associated to 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 and Szabó (2022), we will instead investigate the surrogate posteriors directly which will allow for sharp rates. Still, our bounds for the Kullback-Leibler divergences indicate that our correction of the surrogate posterior is beneficial since our corrected surrogate posterior is closer to the classical Gibbs-posterior than the surrogate posterior without the correction.

We prove oracle inequalities for the surrogate posteriors of the stochastic MH method and its corrected modification. Based on that we can conclude contraction rates as well as rates of convergence for the surrogate posterior mean. These findings reveal that indeed the surrogate posterior of csMALA has a high concentration around the true regression function compared to the surrogate posterior of sMALA. Moreover, we investigate the size and coverage of credible balls from the surrogate posterior. These results apply also to the original Gibbs-posterior as a special case, which might be of interest independent of the discussion of the surrogate posterior.

We apply the oracle inequality and the credibility theorem in the context of shallow as well as deep neural networks. For shallow neural networks the oracle inequality yields optimal convergence rates and credible ball diameters for Hölder-regular functions up to a logarithmic factor. Due to the complex and non-invertible relationship between a Bayesian neural network and its parameters, establishing coverage guarantees remains a longstanding open problem. Towards this aim, we show that credibility can be verified when the critical value is computed in the parameter space rather than in the prediction space. For deep neural networks we show that the contraction rate of the corrected stochastic MH procedure coincides with the minimax rate by Schmidt-Hieber (2020) (up to a logarithmic factor) for Hölder-regular hierarchical regression functions. While the latter paper has analyzed sparse deep neural networks with ReLU activation function, similar results for fully connected networks are given by Kohler and Langer (2021) and we exploit their main approximation theorem. A mixing approach in the prior distribution, compatible with, *e.g.*, Alquier and Biau (2013) and Guedj and Alquier (2013), leads to a fully adaptive method.

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 approximate samples from the surrogate posterior of csMALA, as well as their mean, show a significant improvement in terms of the empirical prediction risk and size of credible balls over those taken from the surrogate posterior of the naive sMALA. The correction term cancels the bias on the size of accepted batches introduced by the stochastic setting. The Python code of the numerical example is available on GitHub.[1]

1. `https://github.com/sbieringer/csMALA.git`

Related literature. In view of possibly better scaling properties, variational Bayes methods have been studied intensively in recent years. Instead of sampling from the posterior distribution itself, variational Bayes methods 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 and Zhou (2020); Zhang and Gao (2020); Ray and Szabó (2022) and references therein.

Our oracle inequalities rely on PAC-Bayes theory which provides *probably approximately correct* error bounds and goes back to Shawe-Taylor and Williamson (1997) and McAllester (1999a,b). We refer to the review papers by Guedj (2019) and Alquier (2024), the monograph by Hellström et al. (2025) and the pioneering works by Catoni (2004, 2007). PAC-Bayes bounds in a regression setting have been studied, see, *e.g.*, Audibert (2004, 2009); Audibert and Catoni (2011) and the references therein. Note that applying PAC-Bayes techniques is not straightforward, as the underlying loss of the surrogate posterior differs from the standard L^2-loss with respect to which our oracle inequality is formulated.

While we focus on oracle inequalities, PAC-Bayes bounds could be used to derive numerical risk certificates and to derive cost functions for training, as demonstrated, *e.g.*, by Dziugaite and Roy (2017), Zhou et al. (2019), Pérez-Ortiz et al. (2021), Biggs and Guedj (2021, 2022, 2023) in the contexts of deep and shallow learning.

Our analysis of the Bayesian procedure from a frequentist point of view embeds into the nonparametric Bayesian inference, see Ghosal and van der Vaart (2017). Coverage of credible sets has been studied, for instance, by Szabó et al. (2015) and Rousseau and Szabó (2020) and based on the Bernstein-von Mises theorem in Castillo and Nickl (2014) among others. We would like to point out that our proof strategy to verify credibility with PAC-Bayes techniques seems novel and provides an alternative and possibly simpler way compared to the previously mentioned results.

While contraction rates for Bayes neural networks have been studied by Polson and Ročková (2018), Chérief-Abdellatif (2020) and Castillo and Egels (2025), the theoretical properties of credible sets are not well understood so far. Franssen and 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.

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 and 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 been given by Schmidt-Hieber (2020) as well as Bauer and 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.

Tailoring Markov chains to the needs of current neural network applications is a field of ongoing investigation. Different efforts to improve efficiency by improve mixing, that is transitioning between modes of the posterior landscape, exist. 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 Deng et al. (2020b, 2022) uses 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 and Jalaian (2021) introduced a splitting scheme for Hamiltonian Monte Carlo maintaining the full Hamiltonian.

Organization. The paper is organized as follows: In Section 2, we derive the stochastic MH procedure, introduce the stochastic MH correction and study the Kullback-Leibler divergences of the corresponding surrogate posteriors from the Gibbs posterior. In Section 3, we state the oracle inequality and the resulting contraction rates in a general regression setting and we investigate credible sets. In Section 4 we apply the methodology to the setting of Bayesian neural networks. The numerical performance of the method is studied in Section 5 and our key takeaways are summarized in Section 6. All proofs have been postponed to Section 7 which starts with an overview of how the main results and the auxiliary results relate to each other in their proofs.

2. Surrogate Gibbs-posteriors for regression

The aim is to estimate a regression function $f\colon \mathbb{R}^d \to \mathbb{R}$, $d \in \mathbb{N}$, under a quadratic loss and under classical assumptions based on a training sample $\mathcal{D}_n \coloneqq (\mathbf{X}_i, Y_i)_{i=1,\dots,n} \subseteq \mathbb{R}^d \times \mathbb{R}$ given by $n \in \mathbb{N}$ i.i.d. copies of generic random variables $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $Y = f(\mathbf{X}) + \varepsilon$ and observation error 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 (conditional) 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 $\mathbf{X}$. A common alternative from the aforementioned literature on risk certificates would be to consider the generalization gap which allows for the use of PAC-Bayes generalization bounds as training objectives, see, *e.g.*, Guedj (2019), Alquier (2024), Hellström et al. (2025).

We consider a parametric class of potential estimators $\mathcal{F} = \{f_\vartheta : \vartheta \in \Theta\}$, where the finite dimensional parameter space is fixed as $\Theta = [-B, B]^Q$ for simplicity with some $B \geqslant 1$ and a potentially large parameter dimension $Q \in \mathbb{N}$. For $f_\vartheta \in \mathcal{F}$ we abbreviate $R(\vartheta) = R(f_\vartheta)$ and

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

Throughout, $|x|_p$ denotes the ℓ^p-norm of a vector $x \in \mathbb{R}^d$, $p \in [1, \infty]$. For brevity, $|\cdot| := |\cdot|_2$ is the Euclidean norm. We write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$. The identity matrix in $\mathbb{R}^{d \times d}$ is denoted by I_d and $\mathcal{O}_{\mathbb{P}}$ is the stochastic big O Landau notation.

2.1 Prior and posterior distribution

As prior on the parameter set of the class $\mathcal{F}$ we choose a uniform distribution $\Pi = \mathcal{U}([-B, B]^Q)$. The corresponding *Gibbs posterior* $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$ is defined as the solution to the minimization problem

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

where the infimum is taken over all probability distributions ν on $\mathbb{R}^Q$. 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 (7.2) for a definition) to the prior distribution Π and weighted via the *inverse temperature parameter* $\lambda > 0$. Owing to the change of measure inequality, or, the Legendre transform of the Kullback-Leibler divergence, this optimization problem has a unique solution given by

$$\Pi_\lambda(\mathrm{d}\vartheta \mid \mathcal{D}_n) \propto \exp\big(- \lambda R_n(\vartheta) \big) \Pi(\mathrm{d}\vartheta), \tag{2.1}$$

see Csiszar (1975), Donsker and Varadhan (1976), Catoni (2004, 2007) or Lemma 18 below. While (2.1) coincides with the classical Bayesian posterior distribution if $Y_i = f_\vartheta(\mathbf{X}_i) + \varepsilon_i$ with i.i.d. $\varepsilon_i \sim \mathcal{N}(0, \frac{n}{2\lambda})$, the so-called tempered likelihood, see, *e.g.*, Dalalyan and Tsybakov (2008), Alquier and Lounici (2011), Bissiri et al. (2016), Guedj and Robbiano (2018), Guedj (2019), $\exp(-\lambda R_n(\vartheta))$ 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

$$\widehat{f}_\lambda := f_{\widehat{\vartheta}_\lambda} \qquad \text{for} \qquad \widehat{\vartheta}_\lambda \mid \mathcal{D}_n \sim \Pi_\lambda(\cdot \mid \mathcal{D}_n), \tag{2.2}$$

or via the posterior mean

$$\bar{f}_\lambda := \mathbb{E}\big[f_{\widehat{\vartheta}_\lambda} \,\big|\, \mathcal{D}_n \big] = \int f_\vartheta \, \Pi_\lambda(\mathrm{d}\vartheta \mid \mathcal{D}_n). \tag{2.3}$$

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

To apply the estimators $\widehat{f}_\lambda$ and $\bar{f}_\lambda$ in practice, we need to sample from the Gibbs posterior. 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 and Casella (2004). In particular, the *Langevin* MCMC sampler is given by

$$\vartheta^{(k+1)} = \vartheta^{(k)} - \gamma \nabla_\vartheta R_n(\vartheta^{(k)}) + s W_k, \tag{2.4}$$

where $\nabla_\vartheta R_n(\vartheta)$ denotes the gradient of $R_n(\vartheta)$ with respect to ϑ, $\gamma > 0$ is the learning rate and $sW_k \sim \mathcal{N}(0, s^2 I_Q)$ 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 and Wang (2022); Cheng and Bartlett (2018). Only in special cases, the convergence rates are faster, see, *e.g.*, Freund et al. (2022) for an overview and Dalalyan and 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 and Tweedie (1996b).

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 $\vartheta^{(0)} \in \mathbb{R}^Q$, 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

$$\alpha(\vartheta' \mid \vartheta) = \exp\big(- \lambda R_n(\vartheta') + \lambda R_n(\vartheta)\big) \mathbb{1}_{[-B,B]^Q}(\vartheta') \frac{q(\vartheta \mid \vartheta')}{q(\vartheta' \mid \vartheta)} \wedge 1. \tag{2.5}$$

In view of (2.4) the probability density q of the proposal distribution is given by

$$q(\vartheta' \mid \vartheta) = \frac{1}{(2\pi s^2)^{Q/2}} \exp\Big(- \frac{1}{2s^2} \big|\vartheta' - \vartheta + \gamma \nabla_\vartheta R_n(\vartheta)\big|^2 \Big). \tag{2.6}$$

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 and Tweedie (1996a, Theorem 2.2) with geometric rate.

To calculate the estimators $\widehat{f}_\lambda$ and $\bar{f}_\lambda$ from (2.2) and (2.3), respectively, 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 approximates

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

In practice, b can be calibrated by plotting the training loss against the number of iterations and looking for the point at which the training loss stabilizes.

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.

2.2 Surrogate posterior of 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 datasets 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 $\mathcal{B} \subseteq \{1,\dots,n\}$ of size $m<n$. While we could replace $\nabla_\vartheta R_n(\vartheta)$ in (2.6) 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 (2.5).

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 and Roberts (2009); Maclaurin and Adams (2014); Bardenet et al. (2017); Wu et al. (2022).

We augment our target distribution by a set of auxiliary random variables $Z_1,\dots,Z_n \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(\rho)$ 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

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

The marginal distribution in ϑ is then given by

$$\bar\Pi_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) = \sum_{z\in\{0,1\}^n} \bar\Pi_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n) \propto \prod_{i=1}^n \Big(\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\Big)\Pi(\mathrm{d}\vartheta). \tag{2.7}$$

As proposal for the MH algorithm we use

$$\begin{aligned}\bar q(\vartheta', z' \mid \vartheta, z) &= q_{\mathrm{s}}(\vartheta' \mid \vartheta, z)\prod_{i=1}^n \rho^{z_i'}(1-\rho)^{1-z_i'} \qquad \text{with} \\ q_{\mathrm{s}}(\vartheta' \mid \vartheta, z) &= \frac{1}{(2\pi s^2)^{Q/2}}\exp\Big(-\frac{1}{2s^2}\big|\vartheta' - \vartheta + \gamma\nabla_\vartheta R_n(\vartheta,z)\big|^2\Big).\end{aligned} \tag{2.8}$$

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 (2.6) with a stochastic gradient. The resulting acceptance probabilities are given by

$$\alpha(\vartheta', z' \mid \vartheta, z) = \frac{\bar q(\vartheta, z \mid \vartheta', z')\bar\Pi_{\lambda,\rho}(\vartheta', z' \mid \mathcal{D}_n)}{\bar q(\vartheta', z' \mid \vartheta, z)\bar\Pi_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n)} \wedge 1$$

$$= \frac{q_{\mathrm{s}}(\vartheta \mid \vartheta', z')}{q_{\mathrm{s}}(\vartheta' \mid \vartheta, z)} \mathbb{1}_{[-B,B]^Q}(\vartheta') \mathrm{e}^{-\lambda R_n(\vartheta', z') + \lambda R_n(\vartheta, z)} \wedge 1.$$

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 (2.2), we define the stochastic MH estimator

$$\widehat{f}_{\lambda,\rho} \coloneqq f_{\widehat{\vartheta}_{\lambda,\rho}} \qquad \text{for} \qquad \widehat{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n \sim \bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n). \tag{2.9}$$

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 (2.7) instead of the Gibbs posterior from (2.1). Writing

$$\bar{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) \propto \exp\big(-\lambda \bar{R}_{n,\rho}(\vartheta)\big)\Pi(\mathrm{d}\vartheta) \quad \text{with} \quad \bar{R}_{n,\rho}(\vartheta) \coloneqq -\frac{1}{\lambda}\sum_{i=1}^n \log\big(\rho \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} + 1 - \rho\big), \tag{2.10}$$

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 1. *If f and all f_ϑ are bounded by some constant $C_f > 0$, then we have*

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

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

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

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

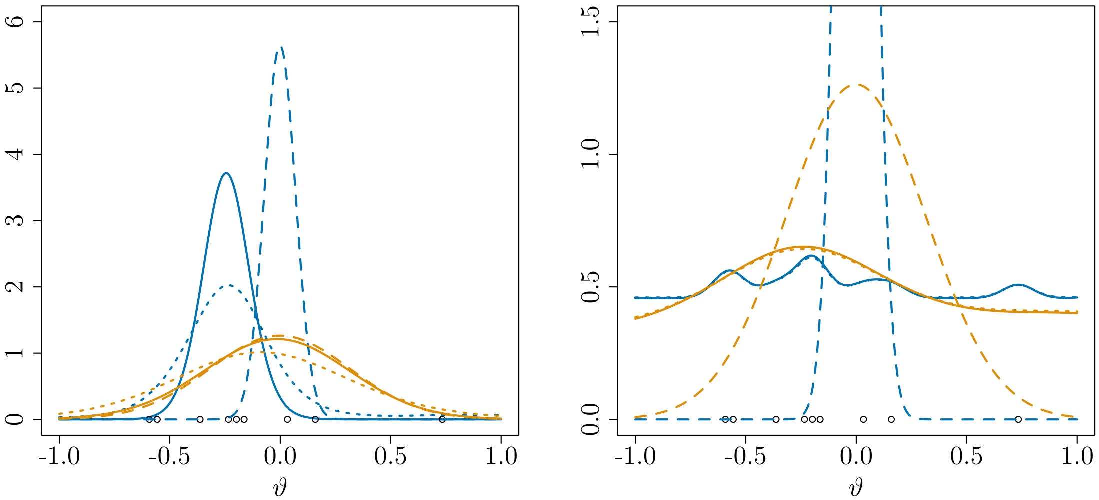

Figure 1: *Points*: $Y_1, \dots, Y_n \sim \mathcal{N}(0, 0.5)$ for $n = 10$. *Solid lines*: densities of $\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ with $\lambda = 10n$ (blue) and $\lambda = n/2$ (orange) and $\rho = 0.9$ (left) and $\rho = 0.1$ (right). *Dashed lines*: corresponding densities of $\Pi_\lambda(\cdot \mid \mathcal{D}_n)$. Dotted lines: corresponding densities of $\varpi_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$.

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 1 in a simple setting where $Y_i = \mathcal{N}(0, 0.5)$ and $f_\vartheta(x) = \vartheta$ 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. 1 for different choices of λ and ρ. Fig. 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 1 is quite natural in this setting. In particular, applying an approximation result from the variational Bayes literature by Ray and 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

$$\mathbb{E}\big[\bar{\Pi}_{\lambda,\rho}(\Theta_n \mid \mathcal{D}_n)\big] \leqslant \frac{2}{n\rho}\mathbb{E}\big[\,\mathrm{KL}\,\big(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)\,\big|\,\mathbb{Q}\big)\big] + C\mathrm{e}^{-n\rho/2}. \tag{2.11}$$

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 $\frac{1}{n\rho}\mathrm{KL}\,\big(\bar{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)\,\big|\,\Pi_\lambda(\cdot \mid \mathcal{D}_n)\big)$. 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 $n\rho$.

2.3 Surrogate posterior of 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 worse convergence rate of an estimator based on its surrogate posterior, as we will demonstrate with Theorem 6 in Section 3.1. This motivates the introduction of a corrected stochastic MALA (csMALA) and the analysis of its surrogate posterior.

To remedy the loss of information while retaining scalability, we define another joint target distribution as

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

with marginal distribution in ϑ given by

$$\begin{aligned}\widetilde{\Pi}_{\lambda,\rho}(\vartheta \mid \mathcal{D}_n) = \sum_{z\in\{0,1\}^n} \widetilde{\Pi}_{\lambda,\rho}(\vartheta, z \mid \mathcal{D}_n) &\propto \prod_{i=1}^{n}\Big(\rho\frac{\mathrm{e}^{-\frac{\lambda}{n}\ell_i(\vartheta)}}{\rho} + 1 - \rho\Big)\Pi(\mathrm{d}\vartheta) \\ &= \exp\big(-\lambda \widetilde{R}_{n,\rho}(\vartheta)\big)\Pi(\mathrm{d}\vartheta)\end{aligned}$$

with

$$\widetilde{R}_{n,\rho}(\vartheta) := -\frac{1}{\lambda}\sum_{i=1}^{n}\log\big(\mathrm{e}^{-\frac{\lambda}{n}\ell_i(\vartheta)} + 1 - \rho\big). \tag{2.12}$$

Remark 2. As with sMALA, we recover MALA for $\rho = 1$. Therefore, all of our upcoming results, especially those regarding uncertainty quantification, also hold for MALA.

Compared to $\bar{R}_{n,\rho}$ from (2.10) there is no ρ in the first term in the logarithm. In line with (2.2) and (2.3), we obtain the estimators

$$\widetilde{f}_{\lambda,\rho} := f_{\widetilde{\vartheta}_{\lambda,\rho}} \qquad \text{for} \qquad \widetilde{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n \sim \widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n) \tag{2.13}$$

and

$$\bar{f}_{\lambda,\rho} := \mathbb{E}\big[f_{\widetilde{\vartheta}_{\lambda,\rho}} \,\big|\, \mathcal{D}_n\big] = \int f_\vartheta \, \widetilde{\Pi}_{\lambda,\rho}(\mathrm{d}\vartheta \mid \mathcal{D}_n). \tag{2.14}$$

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 (2.8) leads to the acceptance probabilities

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

To take the randomized batches into account, we thus introduce a small *correction term* $\frac{\log\rho}{\lambda}|Z|_1 = \mathcal{O}_{\mathbb{P}}(\frac{n}{\lambda}\rho\log\rho)$ 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. *If f and all f_ϑ are bounded by some constant $C_f > 0$, then we have*

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

Compared to Lemma 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 surrogate posterior of sMALA. The only price to pay is a reduction of the inverse temperature parameter λ by the factor $(2-\rho)^{-1} \in [\frac{1}{2}, 1]$. As already mentioned in (2.11), we can conclude contraction and coverage results for $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ by combining Ray and Szabó (2022, Theorem 5) with Lemma 3 if $\lambda/n \to 0$. In a classification setting, this is often the case with $\lambda = \mathcal{O}(\sqrt{n})$, see (Alquier et al., 2016, Corollary 5.2), Catoni (2007). However, in our regression setting, the optimal choice is $\lambda = \mathcal{O}(n)$. Therefore, we instead directly analyze $\widetilde{\Pi}_{\lambda,\rho}(\cdot \mid \mathcal{D}_n)$ leading to a rate as good as the Gibbs measure itself could achieve.

The corrected stochastic MALA (csMALA) is summarized in Algorithm 1. The implementation omits the restriction of the proposed network weights to $[-B, B]^Q$ 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. The convergence of the algorithm for general $\zeta > 0$ and its dependence on ρ is an open research question which could be studied using state-of-the-art Markov chain Monte Carlo theory, see, *e.g.*, Chewi (2025).

3. Oracle inequality and uncertainty quantification

In this section we state the statistical guarantees for the estimators defined in terms of the surrogate posterior distributions. It is worth noting that our analysis is independent of the choice of the proposal distribution. We will derive oracle inequalities for the estimators $\widehat{f}_{\lambda,\rho}$ (Theorem 6) and $\widetilde{f}_{\lambda,\rho}$ (Theorem 4) and as a consequence an analogous oracle inequality for $\bar{f}_{\lambda,\rho}$ (Corollary 7), which verify that these estimators are not much worse than the optimal choice for ϑ. Afterwards, we discuss the properties of credible balls with respect to the posterior distributions.

3.1 Oracle inequality

Our first main result compares the performance of the estimator $\widetilde{f}_{\lambda,\rho}$ from (2.13) to the best possible estimator $f_{\vartheta^*} \in \mathcal{F} = \{f_\vartheta : \vartheta \in [-B, B]^Q\}$ for the *oracle choice*

$$\vartheta^* \in \operatorname*{arg\,min}_{\vartheta \in [-B,B]^Q} R(\vartheta) = \operatorname*{arg\,min}_{\vartheta \in [-B,B]^Q} \mathcal{E}(\vartheta). \tag{3.1}$$

The oracle is not accessible to the practitioner because $R(\vartheta)$ depends on the unknown distribution of $(\mathbf{X}, Y)$. Instead, the oracle serves as a benchmark against which the performance of the estimators can be assessed. A solution to the minimization problem in (3.1) always

Algorithm 1 csMALA - corrected stochastic Metropolis adjusted Langevin Algorithm

Input: inverse temperature $\lambda > 0$, learning rate $\gamma > 0$, standard deviation $s > 0$, correction parameter $\zeta \geqslant 0$, average batch proportion $\rho \in (0,1]$ of data used, burn-in $b \in \mathbb{N}$, gap length $c \in \mathbb{N}$, number of draws $N \in \mathbb{N}$.

1. Initialize $\vartheta^{(0)} \in \mathbb{R}^Q$ and $Z^{(0)} \sim \mathrm{Ber}(\rho)^{\otimes n}$.
2. Calculate $R_n^{(0)} = \frac{1}{n}\sum_{i=1}^n Z_i^{(0)}\ell_i(\vartheta^{(0)}) + \zeta\frac{\log\rho}{\lambda}|Z^{(0)}|_1$ and $\nabla R_n^{(0)} = \nabla_\vartheta R_n(\vartheta^{(0)}, Z^{(0)})$.
3. For $k = 0, \dots, b + cN$ do:
   (a) Draw $Z' \sim \mathrm{Ber}(\rho)^{\otimes n}$.
   (b) Draw $\vartheta' \sim \mathcal{N}(\vartheta^{(k)} - \gamma\nabla R_n^{(k)}, s^2 I_Q)$ and calculate $R'_n = \frac{1}{n}\sum_{i=1}^n Z'_i\ell_i(\vartheta') + \zeta\frac{\log\rho}{\lambda}|Z'|_1$ 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: $\widetilde{f}_{\lambda,\rho} = f_{\vartheta^{(b)}}$, $\bar{f}_{\lambda,\rho} = \frac{1}{N}\sum_{k=1}^N f_{\vartheta^{(b+ck)}}$

exists since $[-B,B]^Q$ is compact and we will assume $\vartheta \mapsto R(\vartheta)$ to be continuous. Throughout we assume w.l.o.g. that $\vartheta^* \in (-B,B)^Q$. If there is more than one solution, we choose one of them. We need some mild assumption on the regression model.

Assumption A.

1. ***Bounded regression function:*** *For some* $C_f \geqslant 1$ *we have* $\|f\|_\infty \leqslant C_f$.
2. ***Conditional sub-Gaussianity of observation noise:*** *There are constants* $\sigma, C_\varepsilon > 0$ *such that*
$$\mathbb{E}[|\varepsilon|^k \mid \mathbf{X}] \leqslant \frac{k!}{2}\sigma^2 C_\varepsilon^{k-2} \text{ a.s.}, \qquad \text{for all } k \geqslant 2.$$
3. ***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. For the function class we require the following:

Assumption B. *Let* $\mathcal{F} = \{f_\vartheta : \vartheta \in \Theta\}$ *with* $\Theta = [-B,B]^Q$ *satisfy:*

1. ***Bounded functions:*** *There is some* $C_f \geqslant 1$ *such that* $\|f_\vartheta\|_\infty \leqslant C_f$ *for all* $f_\vartheta \in \mathcal{F}$.
2. ***Lipschitz dependence on the parameters:*** *There is a norm* $|\cdot|_\Theta$ *on* Θ *and some constant* $\Delta > 0$ *such that*
$$\|f_\vartheta - f_{\widetilde{\vartheta}}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \Delta|\vartheta - \widetilde{\vartheta}|_\Theta \qquad \text{for all} \qquad \vartheta, \widetilde{\vartheta} \in \Theta.$$

We obtain the following non-asymptotic oracle inequality for our estimator $\widetilde{f}_{\lambda,\rho}$ from (2.13):

Theorem 4 (PAC-Bayes oracle inequality for the surrogate posterior of csMALA)**.** *Under Assumption A and Assumption B there are constants* $K_0, K_1 > 0$ *depending only on* $C_f, C_\varepsilon, \sigma$ *such that for* $\lambda = n/K_0$ *and sufficiently large* n *we have for all* $\delta \in (0,1)$ *with probability at least* $1-\delta$ *that*

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{K_1}{n}\big(Q\log(B\Delta n) - \log \mathrm{vol}(\mathcal{B}_1) + \log(2/\delta)\big), \tag{3.2}$$

where $\mathrm{vol}(\mathcal{B}_1)$ *denotes the volume of the* $|\cdot|_\Theta$*-unit ball in* $\mathbb{R}^Q$.

Remark 5. For $\rho = 1$ we do not need the conditional symmetry condition in Assumption A. The remaining parts of Assumption A are in line with the literature, see, *e.g.*, Alquier and Biau (2013); Guedj and Alquier (2013). An explicit admissible choice for λ is $\lambda = n/\big(2^5 C_f(C_\varepsilon \vee (2C_f)) + 2^7(C_f^2 + \sigma^2) + 2^3(\sigma C_f + \sigma^2)\big)$. The dependence of K_1 on $C_f, C_\varepsilon, \sigma$ is at most quadratic. While the oracle inequality is not sharp (the leading constant 12 in front of $\mathcal{E}(f_{\vartheta^*})$ is larger than 1), it leads to optimal contraction rates (up to logarithms) in terms of the sample size as we will demonstrate in the upcoming Proposition 13.

The term $\log \mathrm{vol}(\mathcal{B}_1)$ depends on the geometry induced by the $|\cdot|_\Theta$-norm. If $|\cdot|_\Theta = |\cdot|_p$ for some $p \in [1,\infty]$ and with the gamma function Γ, we have

$$-\log \mathrm{vol}(\mathcal{B}_1) = \log(\Gamma(1+Q/p)) - Q\log\big(2\Gamma(1+1/p)\big),$$

which is $-Q\log 2$ for $p=\infty$ and of order $Q\log Q$ for $p\in[1,\infty)$. Hence, as long as Q does not grow faster than polynomially in n, the stochastic error is of the order $Q\log(B\Delta n)/n$ and $Q\ll n/\log(n)$ is necessary to achieve convergence. This is especially the case in our application to Bayesian neural networks, see Section 4, where we obtain optimal convergence rates based on the above oracle inequality.

The right-hand side of (3.2) can be interpreted similarly to the classical bias-variance decomposition in nonparametric statistics. The first term $\mathcal{E}(f_{\vartheta^*})=\mathbb{E}[(f_{\vartheta^*}(\mathbf{X})-f(\mathbf{X}))^2]$ quantifies the approximation error while the second term is an upper bound for the stochastic error. A key implication is that the excess risk is small if the methodology is applied to a setting where an oracle choice of ϑ balances the order of corresponding the oracle risk and the model complexity. For this, we will later consider neural networks, see Section 4. Theorem 4 is in line with classical PAC-Bayes oracle inequalities, see Bissiri et al. (2016), Guedj and Alquier (2013), Zhang (2006). 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 4 is to verify the compatibility between the risk $\widetilde{R}_{n,\rho}$ from (2.12) and the empirical risk R_n as we will establish in Proposition 16.

We obtain a similar result for $\widehat{f}_{\lambda,\rho}$ from (2.9). The key difference is that the stochastic error term is of order $\mathcal{O}(\frac{Q}{n\rho})$ instead of $\mathcal{O}(\frac{Q}{n})$ as in Theorem 4 (up to the logarithm). Consequently, an estimator based on the surrogate Gibbs-posterior of csMALA benefits from the full sample size thanks to the correction terms, whereas without the correction, it does not.

Theorem 6 (PAC-Bayes oracle inequality for the surrogate posterior of sMALA)**.** *Under Assumption A and Assumption B there are constants $K_0',K_1'>0$ depending only on C_f,C_ε,σ such that for $\lambda=n\rho/K_0'$ and sufficiently large n we have for all $\delta\in(0,1)$ with probability at least $1-\delta$ that*

$$\mathcal{E}(\widehat{f}_{\lambda,\rho})\leqslant 4\mathcal{E}(f_{\vartheta^*})+\frac{K_1'}{n\rho}\big(Q\log(B\Delta n)-\log\operatorname{vol}(\mathcal{B}_1)+\log(2/\delta)\big).$$

In view of Theorem 6 the following results are also true for the stochastic MH estimator if n is replaced by $n\rho$. However, we subsequently focus only on the analysis of $\widetilde{\Pi}_{\lambda,\rho}(\cdot\mid\mathcal{D}_n)$ for the sake of clarity.

The $1-\delta$ probability in Theorem 4 takes into account the randomness of the data and of the estimate. Denoting

$$\varrho_n^2 \coloneqq 12\|f_{\vartheta^*}-f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})}+\frac{K_1}{n}\big(Q\log(B\Delta n)-\log\operatorname{vol}(\mathcal{B}_1)\big), \tag{3.3}$$

we can rewrite (3.2) as

$$\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}\big(\|f_{\widetilde{\vartheta}_{\lambda,\rho}}-f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})}>\varrho_n^2+t^2\,\big|\,\mathcal{D}_n\big)\big]\leqslant 2\mathrm{e}^{-nt^2/K_1},\qquad t>0,$$

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 (2.14).

Corollary 7 (Posterior mean). *Under the conditions of Theorem 4 we have with probability at least* $1-\delta$ *that*

$$\mathcal{E}(\bar{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{K_2}{n}\big(Q\log(B\Delta n) - \log\mathrm{vol}(\mathcal{B}_1) + \log(2/\delta)\big)$$

with a constant K_2 *only depending on* $C_f, C_\varepsilon, \sigma$ *from Assumption A.*

3.2 Credible sets

In addition to the contraction rates, the Bayesian approach offers a gateway to uncertainty quantification. To this end, we will study *credible balls* of the form

$$\widehat{C}(\tau_\alpha) := \{h \in L^2 : \|h - \bar{f}_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \tau_\alpha\}, \qquad \alpha \in (0,1), \tag{3.4}$$

with critical values

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

By construction $\widehat{C}(\tau_\alpha)$ is the smallest L^2-ball around $\bar{f}_{\lambda,\rho}$ that contains $1-\alpha$ mass of the surrogate posterior measure. While we assume that the distribution $\mathbb{P}^{\mathbf{X}}$ of $\mathbf{X}$ is known, to calculate this ball in practice, the distribution of $\mathbf{X}$ could be replaced by its empirical analog.

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 and 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, aiming for a neural network setting causes an additional problem.

In order to prove coverage, *i.e.*, to prove that the credible ball contains the true regression function with high probability, 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 $\|f_\vartheta - f_{\vartheta'}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \Delta|\vartheta - \vartheta'|_\Theta$ by Assumption B, the converse direction does not hold in general. For instance, even locally around an oracle choice ϑ^* we cannot hope to control $|\vartheta|_p$ via $\|f_\vartheta\|_{L^2(\mathbb{P}^{\mathbf{X}})}$ for any $1 \leqslant p \leqslant \infty$ in view of the ambiguous network parametrization. As a consequence, we define another critical value at the level of the parameter space

$$\tau_\alpha^{\vartheta_0} := \arg\min_{\tau>0}\big\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : |\vartheta - \vartheta_0|_\Theta \leqslant \Delta^{-1}\tau \mid \mathcal{D}_n) > 1-\alpha\big\}$$

for some center $\vartheta_0 \in (-B,B)^Q$. Both critical values measure the fluctuation of the posterior. The theoretical properties of the credible ball are summarized in the following theorem:

Theorem 8 (Credible balls)**.** *Under Assumption A, Assumption B and with constants $K_0, K_1, K_2 > 0$ from above we have for $\lambda = n/(pK_0)$ with $p \geqslant 2$, ϱ_n^2 from (3.3) and sufficiently large n that*

$$\mathbb{P}\Big(\operatorname{diam}\big(\widehat{C}(\tau_\alpha)\big) \leqslant 4\sqrt{p\varrho_n^2 + \frac{2p(K_1 \vee K_2)}{n}\log\frac{2}{\alpha}}\Big) \geqslant 1-\alpha.$$

If $\mathcal{E}(f_{\vartheta^*}) = \mathcal{O}(Q/\lambda)$, then we have for sufficiently large $n \in \mathbb{N}$

$$\mathbb{P}\big(f \in \widehat{C}(\xi\tau_\alpha^{\vartheta_0})\big) \geqslant 1-\alpha,$$

where $\xi = C_\tau\sqrt{\log(B\Delta n)}$ with some sufficiently large $C_\tau > 0$ if ϑ_0 does not depend on $\mathcal{D}_n$ and $\xi = (\varrho_n/\Delta)^{-2/p}$ otherwise.

Therefore, the order of the diameter of $\widehat{C}(\tau_\alpha)$ coincides with the contraction rate deduced from Theorem 4. On the other hand, the larger credible set $\widehat{C}(\xi\tau_\alpha^{\vartheta})$ defines an honest confidence set for a fixed class of the regression functions if ξ is chosen sufficiently large. That is, f is contained in $\widehat{C}(\xi\tau_\alpha^{\vartheta_0})$ with probability at least $1-\alpha$. If B and Δ grow at most polynomially in n and ϑ_0 is data independent, we can conclude from Theorem 8 that

$$\mathbb{P}\big(f \in \widehat{C}(\tau_\alpha^{\vartheta_0}\log n)\big) \geqslant 1-\alpha \qquad \text{for sufficiently large } n,$$

which is in line with the inflation factor by Rousseau and Szabó (2020).

The condition $\mathcal{E}(f_{\vartheta^*}) = \mathcal{O}(Q/\lambda) = o(Q\log(B\Delta n)/\lambda)$ for the coverage result means that the rate is dominated by the stochastic error term, cf. (3.2). This can be achieved with a slightly larger parameter class than an optimal choice which balances the approximation error $\mathcal{E}(f_{\vartheta^*})$ and the stochastic error. This guarantees that the posterior is not underfitting and that the posterior's bias is covered by its dispersal. The necessity of under-smoothing is well known for statistical inference.

Example 1 (Linear model)**.** *We illustrate the above credibility statement for a linear model $f_\vartheta(\mathbf{x}) = \mathbf{x}^\top\vartheta$ with $\mathbf{x} \in [0,1]^d, \vartheta \in [-B,B]^d$ and true regression function $f = f_{\vartheta^*}$. Hence, the L^2-loss is*

$$\mathcal{E}(f_\vartheta) = \mathbb{E}[(\mathbf{X}^\top\vartheta - \mathbf{X}^\top\vartheta^*)^2] = (\vartheta-\vartheta^*)^\top\mathbb{E}[\mathbf{X}\mathbf{X}^\top](\vartheta-\vartheta^*).$$

In particular, Assumption B is satisfied for $|\cdot|_\Theta = |\cdot|_2$ and $\Delta^2 = \lambda_{\max}(\Sigma)$ being the largest eigenvalue of the design matrix $\Sigma = \mathbb{E}[\mathbf{X}\mathbf{X}^\top]$. If Σ is positive definite with smallest eigenvalue $\lambda_{\min}(\Sigma) > 0$, the norms $\|f_\vartheta\|_{L^2(\mathbb{P}^{\mathbf{X}})}$ and $|\vartheta|_2$ are equivalent. Theorem 4 then yields $\mathcal{E}(\widetilde{f}_{\lambda,\rho}) = \mathcal{O}(\frac{d}{n}\log n)$ with high probability.

Due to the linearity of $\vartheta \mapsto f_\vartheta$, the mean predictor from (2.14) *satisfies $\bar{f}_{\lambda,\rho}(\mathbf{X}) = \mathbf{X}\bar{\vartheta}_{\lambda,\rho}$ with $\bar{\vartheta} := \bar{\vartheta}_{\lambda,\rho} := \mathbb{E}[\widetilde{\vartheta}_{\lambda,\rho} \mid \mathcal{D}_n]$. Therefore,*

$$\begin{aligned}
\tau_\alpha &= \arg\min_{\tau>0}\big\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : \|\mathbf{X}(\vartheta - \bar{\vartheta}_{\lambda,\rho})\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \tau \mid \mathcal{D}_n) > 1-\alpha\big\}\\
&\leqslant \arg\min_{\tau>0}\big\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : |\vartheta - \bar{\vartheta}_{\lambda,\rho}| \leqslant \Delta^{-1}\tau \mid \mathcal{D}_n) > 1-\alpha\big\}\\
&= \tau_\alpha^{\bar{\vartheta}_{\lambda,\rho}} \leqslant \arg\min_{\tau>0}\Big\{\widetilde{\Pi}_{\lambda,\rho}\Big(\vartheta : \|\mathbf{X}(\vartheta - \bar{\vartheta}_{\lambda,\rho})\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant \frac{\lambda_{\min}(\Sigma)}{\Delta}\tau \,\Big|\, \mathcal{D}_n\Big) > 1-\alpha\Big\} = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\tau_\alpha.
\end{aligned}$$

Theorem 8 thus yields $\mathbb{P}\big(f \in \widehat{C}\big(\big(\frac{d}{n}\big)^{\kappa}\tau_\alpha\big)\big) \geqslant 1-\alpha$ *for an arbitrary small* $\kappa > 0$.

To illustrate the uncertainty quantification on both the function level and the parameter level, we consider a small simulation example with i.i.d. data $\mathcal{D}_n = (\mathbf{x}_i, y_i)_{1\leqslant i\leqslant n}$ *for* $n = 200, d = 2$ *drawn from* $Y = \mathbf{X}^\top\vartheta^* + \varepsilon$ *with* $\vartheta^* = (0.5, 0.5)^\top$, $\mathbf{X} = (1, U), U \sim \mathcal{U}(0,1), \varepsilon \sim \mathcal{N}(0, 0.1^2)$. *We applied MALA with* $\lambda = n, \gamma = 0.01$, *sMALA with* $\rho = 0.1, \lambda = n$ *and csMALA with* $\rho = 0.1, \lambda = n/\rho, \gamma = 0.01/\rho, \zeta = 0.2$. *We used a standard deviation of the proposals of* $s = 0.1$, *a burn-in time of* $b = 1000$ *steps and a gap length of* $c = 100$ *to draw* $N = 100$ *samples from each algorithm and calculate the critical values. Fig. 2 illustrates a typical run. All credible sets achieved a coverage of* 100% *on both the function level and the parameter level, even without an inflating factor* ξ *needed in theory. The radii of the credible sets generated by MALA and csMALA are comparable to each other and much smaller than that of sMALA, which suggests that the algorithm benefits from the correction term.*

This impression is confirmed by a Monte Carlo simulation with 100 *iterations to approximate the coverage and radii of the credible sets. There, all credible sets achieve a coverage of* 100% *and the average radii on the function level (parameter level) are* 0.1176 (0.2867) *for MALA,* 0.1548 (0.3802) *for sMALA and* 0.1167 (0.2838) *for csMALA.*

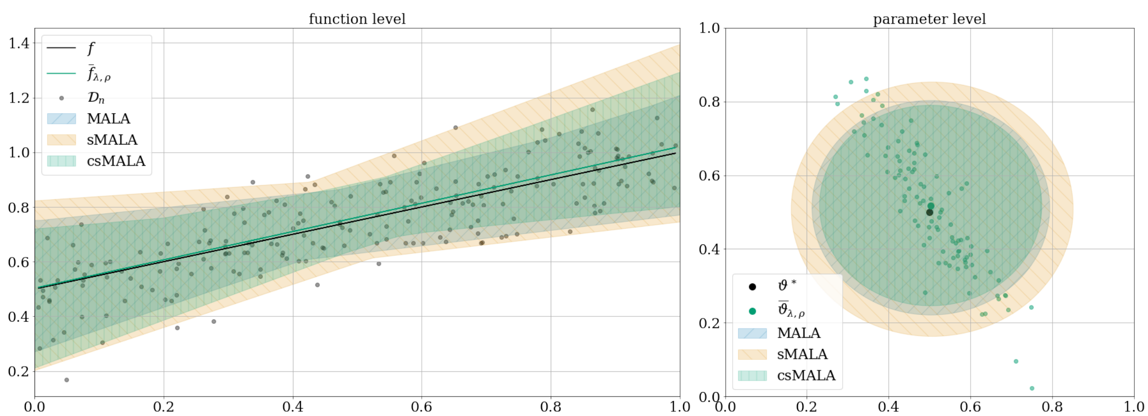

Figure 2: Illustration of credible balls in a linear model. The left hand side displays the data, the true regression function and the (approximate) posterior mean of csMALA. The hatched areas are bounded by the pointwise minimum and maximum sampled prediction from MALA, sMALA and csMALA, respectively. The right hand side shows the credible balls on the parameter level, centered at their respective posterior means. The faint dots indicate 100 parameters sampled using csMALA.

4. Application to Bayesian neural networks

In the sequel the estimator $\widehat{f}$ is chosen as a neural network such that we provide statistical guarantees for stochastic neural networks. More precisely, we consider a *feedforward multilayer perceptron* with $d \in \mathbb{N}$ 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) $\phi(x) := \max\{x, 0\}, x \in \mathbb{R}$, is used as activation function. We write $\phi_v x := \big(\phi(x_i + v_i)\big)_{i=1,\dots,d}$

for vectors $x, v \in \mathbb{R}^d$. With this notation we can represent such neural networks as

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

where the parameter vector ϑ contains all entries of the weight matrices $W^{(1)} \in \mathbb{R}^{r\times d}$, $W^{(2)}, \dots, W^{(L)} \in \mathbb{R}^{r\times r}, W^{(L+1)} \in \mathbb{R}^{1\times r}$ and the shift ('bias') vectors $v^{(1)}, \dots, v^{(L)} \in \mathbb{R}^r$, $v^{(L+1)} \in \mathbb{R}$. We thus set throughout

$$\Theta = [-B, B]^Q \qquad \text{with} \qquad Q \coloneqq (d+1)r + (L-1)(r+1)r + r + 1.$$

A layer-wise representation of g_ϑ is given by

$$\begin{aligned} \mathbf{x}^{(0)} &\coloneqq \mathbf{x} \in \mathbb{R}^d, \\ \mathbf{x}^{(l)} &\coloneqq \phi(W^{(l)}\mathbf{x}^{(l-1)} + v^{(l)}),\ l = 1, \dots, L, \\ g_\vartheta(\mathbf{x}) \coloneqq \mathbf{x}^{(L+1)} &\coloneqq W^{(L+1)}\mathbf{x}^{(L)} + v^{(L+1)}, \end{aligned} \tag{4.1}$$

where the activation function is applied coordinate-wise. We denote the class of all such functions g_ϑ by $\mathcal{G}(d, L, r)$. For some $C_f \geqslant 1$, we also introduce the class of clipped networks

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

4.1 Shallow networks

We start with shallow neural networks where $L = 1$ and thus $Q = (d+2)r + 1$. The study of the approximation properties of shallow networks has a long history going back to Barron (1993) and is closely related to their variational spaces Bach (2017). Building on this branch of literature, the following result is a corollary of (Yang and Zhou, 2024, Theorem 2), where we formulate the approximation properties of shallow networks with uniformly bounded weights. Let $\mathcal{C}_p^\beta([0,1]^d, C_0)$ denote classical Hölder balls of functions $f\colon [0,1]^d \to \mathbb{R}$ with Hölder regularity $\beta > 0$ and Hölder norms bounded by $C_0 > 0$.

Lemma 9. *For any $0 < \beta < (d+3)/2, C_0 > 0$ there is a constant $C_{\text{shallow}} = C_{\text{shallow}}(\beta, C_0) > 0$ such that for any $f \in \mathcal{C}^\beta([0,1]^d, C_0)$ there is a shallow network $g_\vartheta \in \mathcal{G}(d, 1, r)$ with weights uniformly bounded by $B = 1 \vee r^{(3-d-2\beta)/(2d)}$ such that*

$$\|g_\vartheta - f\|_{L^\infty([0,1]^d)} \leqslant C_{\text{shallow}} r^{-\beta/d}.$$

If $d \geqslant 3$ or $\beta > \frac{3-d}{2}$, all network weights can thus be uniformly bounded by one. In combination with a Lipschitz bound of the neural networks with respect to the parameters (see Lemma 21 below), Theorem 4 then yields the following result:

Theorem 10 (PAC-Bayes oracle inequality for shallow neural networks)**.** *Let $\mathbb{E}[|\mathbf{X}|^2] < \infty$. Under Assumption A there are constants $K_0, K_1 > 0$ depending only on $C_f, C_\varepsilon, \sigma$ such that for $\lambda = n/K_0$ and sufficiently large n we have for all $\delta \in (0,1)$ with probability at least $1-\delta$ that*

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{K_1}{n}\big(Q\log(rBn) + \log(2/\delta)\big).$$

If $\mathbf{X} \in [0,1]^d$ *and* $f \in \mathcal{C}^\beta([0,1]^d, C_0)$ *for some* $0 < \beta < (d+3)/2, C_0 > 0$ *we obtain for some* $K_2 > 0$

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant K_2\big(r^{-2\beta/d} + \frac{r}{n}\log(rBn) + \log(2/\delta)\big).$$

A shallow network with number of neurons r *of order* $n^{d/(2\beta+d)}$ *thus yields* $\mathcal{E}(\widetilde{f}_{\lambda,\rho}) = \mathcal{O}_{\mathbb{P}}(n^{-2\beta/(2\beta+d)}\log n)$*. All bounds hold true for* $\bar{f}_{\lambda,\rho}$*, too.*

Therefore, the shallow Bayesian neural network achieves the minimax optimal rate of convergence up to the factor $\log n$. The result is in line with Tinsi and Dalalyan (2022) who use generally aggregated shallow networks but with Gaussian priors. For empirical risk minimizers the rate has been verified in Yang and Zhou (2024). Towards uncertainty quantification, we can deduce the following statement from Theorem 8.

Proposition 11 (Credible balls for shallow neural networks)**.** *Let* $\mathbf{X} \in [0,1]^d$*. Under Assumption A let* $f \in \mathcal{C}^\beta([0,1]^d, C_0)$ *for some* $0 < \beta < (d+3)/2, C_0 > 0$*. Then there are constants* $K_0, K_1, K_2 > 0$ *such that for* $\lambda = n/K_0$*,* $r = K_1 n^{d/(2\beta+d)}$ *and any* $\alpha \in (0,1)$ *the credible set from* (3.4) *satisfies for sufficiently large* n *that*

$$\mathbb{P}\big(\operatorname{diam}\big(\widehat{C}(\tau_\alpha)\big) \leqslant K_2 n^{-\beta/(2\beta+d)}\big) \geqslant 1-\alpha.$$

For $\tau_\alpha^{\vartheta_0} = \arg\min_{\tau>0}\big\{\widetilde{\Pi}_{\lambda,\rho}(\vartheta : |\vartheta-\vartheta_0|_1 \leqslant \tau/B \mid \mathcal{D}_n) > 1-\alpha\big\}$ *for any (non-random)* $\vartheta_0 \in (-B,B)^Q$ *we have for* $n \to \infty$

$$\mathbb{P}\big(f \in \widehat{C}\big(\log(n)\tau_\alpha^{\vartheta_0}\big)\big) \geqslant 1-\alpha.$$

4.2 Deep neural networks

For deep neural networks with $L \geqslant 2$ we obtain the following corollary of Theorem 4:

Theorem 12 (PAC-Bayes oracle inequality for deep neural networks)**.** *Let* $\mathbb{E}[|\mathbf{X}|^2] < \infty$*. Under Assumption A there are constants* $K_0, K_1 > 0$ *depending only on* $C_f, C_\varepsilon, \sigma$ *such that for* $\lambda = n/K_0$ *and sufficiently large* n *we have for all* $\delta \in (0,1)$ *with probability at least* $1-\delta$ *that*

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{K_1}{n}\big(QL\log(rBn) + \log(2/\delta)\big).$$

Moreover, we have with probability at least $1-\delta$ *that*

$$\mathcal{E}(\bar{f}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta^*}) + \frac{K_2}{n}\big(QL\log(rBn) + \log(2/\delta)\big)$$

with a constant K_2 *only depending on* $C_f, C_\varepsilon, \sigma$ *from Assumption A.*

Note that consistency requires $Q(\log n)/n \to 0$ which excludes the application to over-parametrized neural networks. Under additional structural assumptions such as sparsity a similar PAC-Bayes oracle inequality can be derived as demonstrated by Steffen and Trabs (2026) for the standard Gibbs-posterior.

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

$$\mathcal{H}(p,\mathbf{d},\mathbf{t},\beta,C_0) \coloneqq \Big\{g_p \circ \cdots \circ g_0 \colon [0,1]^d \to \mathbb{R} \,\Big|\, g_i = (g_{ij})_j^\top : [a_i,b_i]^{d_i} \to [a_{i+1},b_{i+1}]^{d_{i+1}},$$

$$g_{ij} \text{ depends on at most } t_i \text{ arguments}, g_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([a_i,b_i]^{t_i}, C_0), \text{ for some } |a_i|,|b_i| \leqslant C_0\Big\},$$

where $\mathbf{d} := (d, d_1, \ldots, d_p, 1) \in \mathbb{N}^{p+2}, \mathbf{t} := (t_0, \ldots, t_p) \in \mathbb{N}^{p+1}, \beta := (\beta_0, ..., \beta_q) \in (0,\infty)^{p+1}$. For a detailed discussion of $\mathcal{H}(p, \mathbf{d}, \mathbf{t}, \beta, C_0)$, see Schmidt-Hieber (2020). Theorem 12 reveals the following convergence rate which is in line with the upper bounds by Schmidt-Hieber (2020) and Kohler and Langer (2021):

Proposition 13 (Rates of convergence)**.** *Let $\mathbf{X} \in [0,1]^d$. In the situation of Theorem 12, 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 $p, |\mathbf{d}|_\infty, |\beta|_\infty, C_0$ such that the estimators $\widetilde{f}_{\lambda,\rho}$ and $\bar{f}_{\lambda,\rho}$ with $B = C_3 n$ for some $C_3 > 0$ satisfy for sufficiently large n uniformly over all hierarchical functions $f \in \mathcal{H}(p, \mathbf{d}, \mathbf{t}, \beta, C_0)$*

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant K_3 \Big(\frac{(\log n)^3}{n}\Big)^{2\beta^*/(2\beta^*+t^*)} + K_3 \frac{\log(2/\delta)}{n} \qquad \textit{and}$$
$$\mathcal{E}(\bar{f}_{\lambda,\rho}) \leqslant K_4 \Big(\frac{(\log n)^3}{n}\Big)^{2\beta^*/(2\beta^*+t^*)} + K_4 \frac{\log(2/\delta)}{n}$$

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

$$\beta^* := \beta^*_{i^*}, \qquad t^* := t^*_{i^*} \qquad \textit{for} \qquad i^* \in \underset{i=0,\ldots,p}{\arg\min} \frac{2\beta_i^*}{2\beta_i^* + t_i^*} \qquad \textit{and} \qquad \beta_i^* := \beta_i \prod_{l=i+1}^{p} (\beta_l \wedge 1).$$

The constants K_3 and K_4 only depend on upper bounds for $q, \mathbf{d}, \beta$ and C_0 as well as the constants from Assumption A.

It has been proved by Schmidt-Hieber (2020) that this is the minimax optimal rate of convergence for the nonparametric estimation of f up to logarithmic factors. Studying the special case of classical Hölder balls $\mathcal{C}_d^\beta([0,1]^d, C_0)$, a contraction rate of order $n^{-2\beta/(2\beta+d)}$ has been derived for deep neural networks by Polson and Ročková (2018), Chérief-Abdellatif (2020). Hierarchical regression functions have also been studied by Castillo and Egels (2025).

4.3 Adaptive choice of the network width

To balance the approximation error term and the stochastic error term in (3.3), 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. Similarly, in a high-dimensional setting where the network size Q increases rapidly with the sample size, overparametrized sparse neural networks can be considered, see Steffen and Trabs (2026). While these methods have interesting theoretical properties, an efficient implementation is challenging and left for future research. Still, they can be seen as a link to the literature on model selection through hierarchical priors, see, *e.g.*, Guedj and Alquier (2013) for additive models.

We set

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

where $\Pi_r = \mathcal{U}([-B,B]^{Q_r})$ with

$$Q_r \coloneqq (d+1)r + (L-1)(r+1)r + r + 1.$$

The basis 2 of the geometric weights is arbitrary and 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 $\check{f}_{\lambda,\rho}$ by drawing a parameter ϑ from the surrogate-posterior distribution with respect to this prior, *i.e.*,

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

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

$$\vartheta_r^* \in \operatorname*{arg\,min}_{\vartheta \in [-B,B]^{Q_r}} R(\vartheta) \tag{4.2}$$

given any width r. We obtain the following adaptive version of Theorem 4:

Theorem 14 (Width-adaptive oracle inequality)**.** *Under Assumption A and if* $\mathbb{E}[|\mathbf{X}|^2] < \infty$*, there is a constant* $K_1 > 0$ *depending only on* $C_f, C_\varepsilon, \sigma$ *such that for* $\lambda = n/K_0$ *(with* K_0 *from Theorem 4) and sufficiently large* n *we have for all* $\delta \in (0,1)$ *with probability at least* $1-\delta$ *that*

$$\mathcal{E}(\check{f}_{\lambda,\rho}) \leqslant \min_{r=1,\dots,n} \Big(12\mathcal{E}(f_{\vartheta_r^*}) + \frac{K_1}{n}\big(Q_r L \log(rBn) + \log(2/\delta)\big)\Big).$$

Since the modified estimator mimics the performance of the optimal network choice regardless of width, we obtain the following width-adaptive version of Proposition 13 with no additional loss in the convergence rate:

Corollary 15 (Width-adaptive rates of convergence)**.** *Let* $\mathbf{X} \in [0,1]^d$*. In the situation of Theorem 14, there exists a network depth* $L = C_4 \log n$ *with* $C_4 > 0$ *only depending on upper bounds for* $p, |\mathbf{d}|_\infty, |\beta|_\infty, C_0$ *such that the estimator* $\check{f}_{\lambda,\rho}$ *satisfies for sufficiently large* n *and* $B = C_5 n$ *uniformly over all hierarchical functions* $f \in \mathcal{H}(p, \mathbf{d}, \mathbf{t}, \beta, C_0)$

$$\mathcal{E}(\check{f}_{\lambda,\rho}) \leqslant K_2\Big(\frac{(\log n)^3}{n}\Big)^{2\beta^*/(2\beta^*+t^*)} + K_2\frac{\log(2/\delta)}{n}$$

with probability at least $1-\delta$*, where* β^* *and* t^* *are as in Proposition 13. The constant* K_2 *only depends on upper bounds for* $p, |\mathbf{d}|_\infty, |\beta|_\infty$ *and* C_0 *as well as the constants* $C_f, C_\varepsilon, \sigma$ *from Assumption A.*

For sparse neural networks, contraction rates for hierarchical Bayes procedures have been analyzed by Polson and Ročková (2018) and Steffen and Trabs (2026). Adaptivity with heavy-tailed weights has been achieved by Castillo and Egels (2025). It should be noted that we cannot hope to construct credible sets with coverage as in Theorem 8 based on the adaptive posterior distribution. It is well known that adaptive honest confidence sets are only possible under additional assumptions, *e.g.*, self-similarity or polished tail conditions, on the regularity of the regression function, see Hoffmann and Nickl (2011) and we remark that such conditions with respect to the network parametrization seem infeasible.

5. Numerical examples

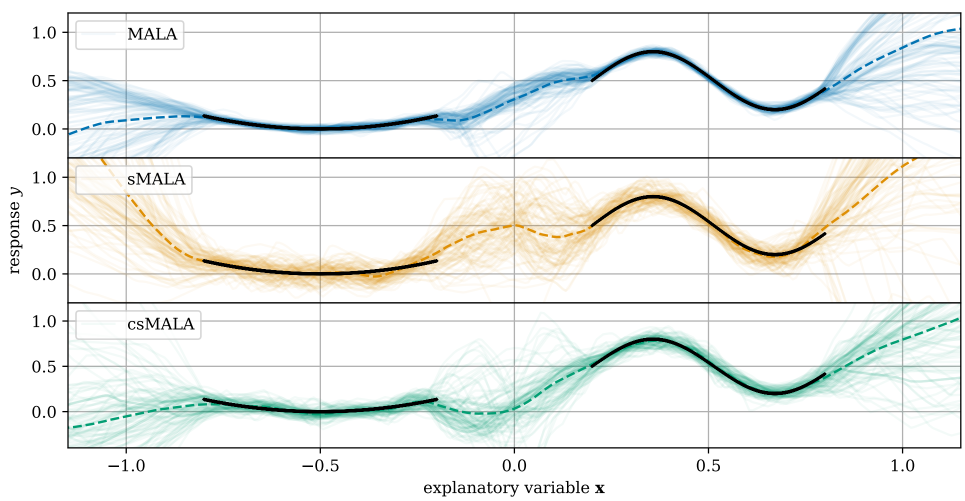

Figure 3: 100 samples drawn from the different MALA chains, given a training sample (black markers) of 10000 points. Random variables are drawn for $\rho = 0.1$. The dashed line shows the corresponding posterior mean $\bar{f}_\lambda$.

Section 2.3 introduces a correction to the batch-wise approximation of the empirical risk when calculating the MH step. In the following, we will show the merit of this correction for learning a one-dimensional regression function using a feed-forward neural network of $L = 2$ layers of $r = 100$ nodes each and ReLU activation. 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 $\mathbf{X}_i \sim \mathcal{U}([-0.8, -0.2] \cup [0.2, 0.8])$ and true regression function

$$f(x) = \begin{cases} 1.5(x+0.5)^2 & \text{for } x < 0 \\ 0.3\sin(10x-2) + 0.5 & \text{for } x \geqslant 0. \end{cases}$$

We generate the training sample following $Y = f(\mathbf{X}) + \varepsilon$ 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

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

Table 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.

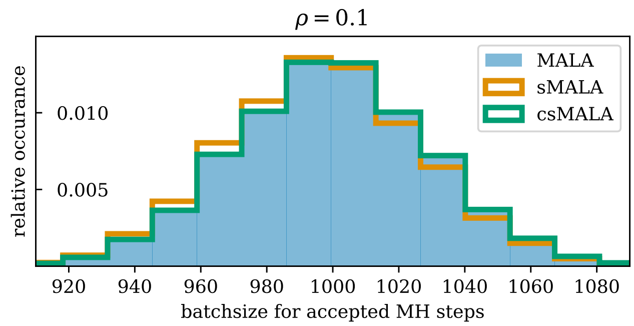

Figure 4: Histogram of the summed auxiliary variables, that is the number of training samples contributing to the stochastic risk, for all accepted steps. For MALA the MH acceptance step is calculated on the full sample and the distribution of the samples contribution to the risk gradients is thus unbiased by the batch size.

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 depicts exactly this behaviour, 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 the subsequent chains are listed in Table 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 2.3. 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 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

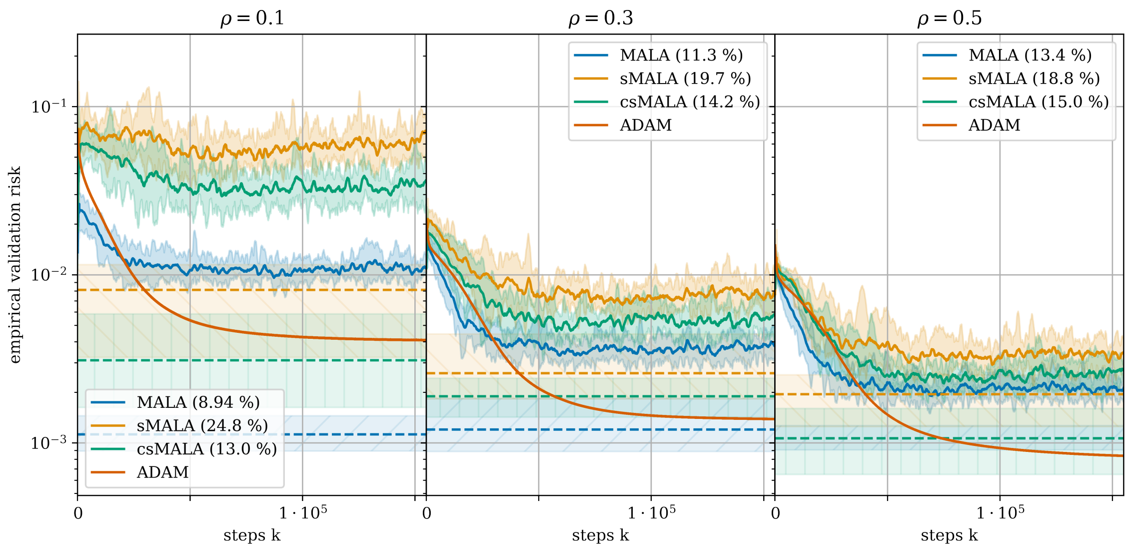

Figure 5: Average empirical risk on a validation set of 10000 points during running of the MALA chains. We show different batch probabilities ρ, as well as the values of the posterior mean (dashed lines). Uncertainties correspond to the minimum and maximum values of 10 identical chains. For clarity, a the 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`. For a fair comparison, we ran all algorithms, including MALA, with stochastic gradients in the proposals.

distributed auxiliary variables. Conversely, for higher values updates are preferably accepted for low amounts of points in the risk approximation. This bias to 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. 4 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 thus falls like the empirical risk with $\zeta \ll 1$ due to the proportionality to n^{-1}.

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

$$R_n(\widehat{f}) = \frac{1}{n_{\mathrm{val}}} \sum_{i=1}^{n_{\mathrm{val}}} \big(Y^{\mathrm{val}}_i - \widehat{f}(\mathbf{X}^{\mathrm{val}}_i)\big)^2$$

during running of the chain. Fig. 5 illustrates the behaviour of the empirical validation risk for the different MALA algorithms, as well as for a simple inference fit using `ADAM` Kingma and 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 see, 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 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

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

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 2. While the uncertainty estimates of all algorithms remain conservative, we find the correction term leads to considerably more precise credible sets.

To illustrate Theorem 4 and Theorem 6, 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 fall with growing n, while sMALA should not decay due to the constant $n\rho$. The numerical simulation of Fig. 6 coincides with the theoretical expectations. For our corrected algorithm, we regain the scaling behavior of MALA as expected.

6. Conclusion

Motivated by MALAs lack of scalability, we considered a stochastic variant, sMALA. It turned out that the corresponding surrogate Gibbs-posterior does not benefit from the full sample size, which is in line with the literature. To remedy this drawback, we introduced a simple correction to sMALA, namely csMALA. Subsequently, we studied its surrogate Gibbs-posterior and verified that it does indeed take advantage of the full sample. Thereby, we refute the conjecture in the literature that a stochastic Metropolis-Hastings step necessarily reduces the effective sample size.

We quantified this phenomenon in terms of the distance to MALAs posterior as well as with oracle inequalities in a nonparametric regression under classical assumptions. Further, we investigated credible sets based on the surrogate Gibbs-posterior which are theoretically valid even without the correction term since MALA is a special case of csMALA. Overall, the quality of the uncertainty quantification depends on the correspondence between the

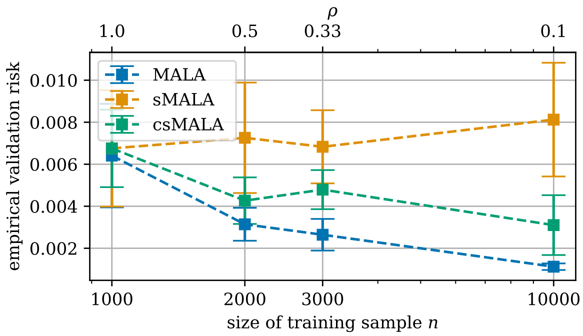

Figure 6: Scaling of the empirical risk of the posterior mean $\bar{f}$ on a 10000 point validation set with the size of the training sample. We scale ρ to keep the average batch size $n\rho = 1000$ constant. Errorbars report the standard deviation of 10 identical chains.

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 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%.

parameters of the network and its output. In a simple linear model, the size of our credible sets depends on the eigenvalues of the design matrix. With a focus on Bayesian neural networks, we derived credible sets for shallow neural networks and optimal contraction rates for deep neural networks. In our simulation study, we demonstrated that the theoretically desirable properties of the surrogate Gibbs-posterior of csMALA carry over to an estimator directly drawn from csMALA.

7. Proofs

Before proving our results, we provide an overview of their interplay with respect to the proofs. The proofs of our main results use our PAC-Bayes oracle inequality for the surrogate Gibbs-posterior of csMALA (Theorem 4) as a starting point. To prove Theorem 4, we first need to ensure compatibility between the corrected risk $\widetilde{R}_{n,\rho}$ and the true excess risk, see Proposition 16 in Section 7.1. Together with the Legendre transform of the Kullback-Leibler divergence (Lemma 18), this leads to a PAC-Bayes bound, see Proposition 19 in Section 7.2. The proof of Theorem 4 is then completed in Section 7.3 by balancing the terms from this PAC-Bayes bound. A simplified version of this overall proof strategy leads to the PAC-Bayes oracle inequality for the surrogate Gibbs-posterior of sMALA (Theorem 6) as we sketch in Section 7.4.

With Theorem 4 at hand, we can verify the remaining main results. As previously mentioned, our analogous results to those following Theorem 4 could also be proved for the original Gibbs-posterior of MALA as well as the surrogate Gibbs-posterior of sMALA, but we focus on the surrogate Gibbs-posterior of csMALA for clarity. A first application of Theorem 4 is its extension to the posterior mean, see Corollary 7. Together with this corollary, we are ready to prove our first contribution towards uncertainty quantification

(Theorem 8) in Section 7.5. We postpone the proof of Corollary 7 as well as the remaining results from Section 3 to Section 7.7.

In Section 7.6, we apply Theorem 4 and Theorem 8 to Bayesian neural networks by verifying Assumption B 2 (Lemma 21) and exploiting the approximation properties of ReLU nets. The remaining proofs of auxiliary lemmas used along the way are postponed to Section 7.8.

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

The first step in our analysis is to verify that the corrected empirical risk $\widetilde{R}_{n,\rho}$ which arises from the stochastic MH step is compatible with the excess risk $\mathcal{E}(\vartheta)=\mathbb{E}\big[\big(f(\mathbf{X}_1)-f_\vartheta(\mathbf{X}_1)\big)^2\big]$. More precisely, we require the following concentration inequality. A concentration inequality for the empirical risk $R_n(\vartheta)-R_n(f)$ follows as the special case where $\rho=1$.

Proposition 16. *Grant Assumption A and Assumption B 1. Define*

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

and set $C_{n,\lambda}:=\frac{\lambda}{n}\frac{8(C_f^2+\sigma^2)}{1-w\lambda/n}$, $w:=16C_f(C_\varepsilon\vee 2C_f)$. *Then for all* $\lambda\in[0,n/w)\cap\big[0,\frac{n\log 2}{8(C_f^2+\sigma^2)}\big]$, $\rho\in(0,1]$ *and* $n\in\mathbb{N}$ *we have*

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

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

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

We have

$$\widetilde{\mathcal{E}}_n(\vartheta)=\frac{1}{n}\sum_{i=1}^n\big(\ell_i(\vartheta)-\ell_i(f)\big)\psi_\rho'\big(\xi_i\tfrac{\lambda}{n}\ell_i(\vartheta)+(1-\xi_i)\tfrac{\lambda}{n}\ell_i(f)\big)\tag{7.1}$$

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 (7.1):

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

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 $\frac{\lambda}{n}\ell_1(\cdot) \leqslant \frac{\lambda}{n}8(C_f^2+\sigma^2) \leqslant \log 2$ for $\frac{\lambda}{n} \leqslant \frac{\log 2}{8(C_f^2+\sigma^2)}$. Hence,

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

where we used Chebyshev's inequality in the last estimate. 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 $\xi'_1 \in [0,1]$

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

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,

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

In combination with the bounds for E_1 we obtain

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

Define $Z_i(\vartheta) := \frac{n}{\lambda}\big(\psi_\rho\big(\frac{\lambda}{n}\ell_i(\vartheta)\big) - \psi_\rho\big(\frac{\lambda}{n}\ell_i(f)\big)\big)$ 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

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

and

$$\mathbb{E}\big[\exp\big(-\lambda\widetilde{\mathcal{E}}_n(\vartheta) + \lambda\mathcal{E}(\vartheta)\big)\big] = \mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n(-Z_i(\vartheta)-\mathbb{E}[-Z_i(\vartheta)])}\big]\mathrm{e}^{\lambda(\mathcal{E}(\vartheta)-\mathbb{E}[\widetilde{\mathcal{E}}_n(\vartheta)])}$$

$$\leqslant \mathbb{E}\big[\mathrm{e}^{\frac{\lambda}{n}\sum_{i=1}^n(-Z_i(\vartheta)-\mathbb{E}[-Z_i(\vartheta)])}\big]\mathrm{e}^{(\frac{3\lambda}{4}+\frac{\lambda^2}{n}(\sigma C_f+\sigma^2))\mathcal{E}(\vartheta)}.$$

To bound the centered exponential moments, we use Bernstein's inequality. The required bounds for the regular moments are obtained as in Alquier and Biau (2013); Guedj and Alquier (2013): The second moments are bounded by

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

Moreover, we have for $k\geqslant 3$

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

Hence, Bernstein's inequality (Massart, 2007, inequality (2.21)) yields

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

for $C_{n,\lambda}$ as defined in Proposition 16. The same bound remains true if we replace Z_i by $-Z_i$. We conclude

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

■

Remark 17. 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 A that $\bar{\mathcal{E}}_n(\vartheta):=\bar R_{n,\rho}(\vartheta)-\bar R_{n,\rho}(f)$ with $\bar R_{n,\rho}$ from (2.10) satisfies for all $\lambda\in[0,n/w)\cap\big[0,\frac{n\log 3}{8(C_f^2+\sigma^2)}\big]$, $\rho\in(0,1]$ and $n\in\mathbb{N}$:

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

7.2 A 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

$$\mathrm{KL}(\mu \mid \nu) := \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu, & \text{if } \mu \ll \nu \\ \infty, & \text{otherwise} \end{cases}. \tag{7.2}$$

The aforementioned Legendre transform of the Kullback-Leibler divergence is a key ingredient for PAC-Bayes bounds, cf. Csiszar (1975), Donsker and Varadhan (1976), Catoni (2004, 2007). We include the short proof for the sake of completeness.

Lemma 18. *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*

$$\log\Big(\int \exp \circ h \,\mathrm{d}\mu\Big) = \sup_{\nu} \Big(\int h \,\mathrm{d}\nu - \mathrm{KL}(\nu \mid \mu)\Big), \tag{7.3}$$

where the supremum is taken over all probability measures ν on $(E, \mathscr{A})$. If additionally, h is bounded from above on the support of μ, then the supremum in (7.3) *is attained for $\nu = g$ with the Gibbs distribution g,* i.e., $\frac{\mathrm{d}g}{\mathrm{d}\mu} :\propto \exp \circ h$.

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

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

■

Note that no generality is lost by considering only those probability measures ν on $(E, \mathscr{A})$ such that $\nu \ll \mu$ and thus

$$\log\Big(\int \exp \circ h \,\mathrm{d}\mu\Big) = -\inf_{\nu \ll \mu} \Big(\mathrm{KL}(\nu \mid \mu) - \int h \,\mathrm{d}\nu\Big).$$

In combination with Proposition 16 we can verify a PAC-Bayes bound for the excess risk. The basic proof strategy is in line with the PAC-Bayes literature, see e.g. Alquier and Biau (2013).

Proposition 19 (PAC-Bayes bound)**.** *Grant Assumption A and Assumption B 1. For any sample-dependent (in a measurable way) probability measure $\varrho \ll \Pi$ and any $\lambda \in (0, n/w)$ and $\rho \in (0, 1]$ such that $C_{n,\lambda} + \frac{\lambda}{n}(\sigma C_f + \sigma^2) \leqslant \frac{1}{8}$, we have*

$$\mathcal{E}(\widetilde{\vartheta}_{\lambda,\rho}) \leqslant 9 \int \mathcal{E} \,\mathrm{d}\varrho + \frac{16}{\lambda}\big(\mathrm{KL}(\varrho \mid \Pi) + \log(2/\delta)\big) \tag{7.4}$$

with probability at least $1 - \delta$.

Proof Proposition 16 yields

$$\mathbb{E}\big[\exp\big(\lambda \widetilde{\mathcal{E}}_n(\vartheta) - \big(1 + C_{n,\lambda} + \tfrac{\lambda}{n}(\sigma C_f + \sigma^2)\big)\lambda \mathcal{E}(\vartheta) - \log \delta^{-1}\big)\big] \leqslant \delta \qquad \text{and}$$

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

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

$$\mathbb{E}\Big[\int\exp\big(\lambda\widetilde{\mathcal{E}}_n(\vartheta)-\big(1+C_{n,\lambda}+\tfrac{\lambda}{n}(\sigma C_f+\sigma^2)\big)\lambda\mathcal{E}(\vartheta)-\log\delta^{-1}\big)\,\mathrm{d}\Pi(\vartheta)\Big]\leqslant\delta \text{ and } \quad (7.5)$$
$$\mathbb{E}\Big[\int\exp\big(\lambda\big(\tfrac{1}{4}-C_{n,\lambda}-\tfrac{\lambda}{n}(\sigma C_f+\sigma^2)\big)\mathcal{E}(\vartheta)-\lambda\widetilde{\mathcal{E}}_n(\vartheta)-\log\delta^{-1}\big)\,\mathrm{d}\Pi(\vartheta)\Big]\leqslant\delta.$$

The Radon-Nikodym density 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

$$\widetilde{D}_\lambda:=\int\mathrm{e}^{-\lambda\widetilde{R}_{n,\rho}(\vartheta)}\,\Pi(\mathrm{d}\vartheta)=\int\exp\Big(-\sum_{i=1}^n\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big)\Big)\,\Pi(\mathrm{d}\vartheta). \quad (7.6)$$

We obtain

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

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

$$\big(\tfrac{1}{4}-C_{n,\lambda}-\tfrac{\lambda}{n}(\sigma C_f+\sigma^2)\big)\mathcal{E}(\widetilde{\vartheta})-\widetilde{\mathcal{E}}_n(\widetilde{\vartheta})+\frac{1}{\lambda}\sum_{i=1}^n\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\widetilde{\vartheta})\big)-\frac{1}{\lambda}\big(\log\delta^{-1}-\log\widetilde{D}_\lambda\big)\geqslant0.$$

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

$$\begin{aligned}\mathcal{E}(\widetilde{\vartheta})&\leqslant8\Big(\widetilde{\mathcal{E}}_n(\widetilde{\vartheta})-\frac{1}{\lambda}\sum_{i=1}^n\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\widetilde{\vartheta})\big)+\frac{1}{\lambda}\big(\log\delta^{-1}-\log\widetilde{D}_\lambda\big)\Big)\\&\leqslant8\Big(-\frac{1}{\lambda}\sum_{i=1}^n\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(f)\big)+\frac{1}{\lambda}\big(\log\delta^{-1}-\log\widetilde{D}_\lambda\big)\Big)\end{aligned}$$

Lemma 18 with $h=-\sum_{i=1}^n\psi_\rho(\frac{\lambda}{n}\ell_i(\vartheta))$ yields

$$\log\widetilde{D}_\lambda=-\inf_{\varrho\ll\Pi}\Big(\mathrm{KL}(\varrho\mid\Pi)+\int\sum_{i=1}^n\psi_\rho\big(\tfrac{\lambda}{n}\ell_i(\vartheta)\big)\,\mathrm{d}\varrho(\vartheta)\Big). \quad (7.7)$$

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

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

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 $C_{n,\lambda}+\frac{\lambda}{n}(\sigma C_f+\sigma^2) \leqslant \frac{1}{8}$, Jensen's inequality and (7.5) to obtain for any probability measure $\varrho \ll \Pi$ (which may depend on $\mathcal{D}_n$)

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

Using $\mathbb{1}_{[0,\infty)}(x) \leqslant \mathrm{e}^{\lambda x}$ again, we conclude with probability 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 at least $1-2\delta$

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

7.3 Proof of Theorem 4

We fix a radius $\eta \in (0,1]$ and apply Proposition 19 with $\varrho = \varrho_\eta$ defined via

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

with ϑ^* from (3.1). Note that indeed $C_{n,\lambda} + \frac{\lambda}{n}(\sigma C_f + \sigma^2) \leqslant \frac{1}{8}$ for n sufficiently large. In order to control the integral term, we decompose

$$\int \mathcal{E}\,\mathrm{d}\varrho_\eta = \mathcal{E}(\vartheta^*) + \int \mathbb{E}\big[(f_\vartheta(\mathbf{X}) - f(\mathbf{X}))^2 - (f_{\vartheta^*}(\mathbf{X}) - f(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta)$$
$$= \mathcal{E}(\vartheta^*) + \int \mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta)$$
$$+ 2\int \mathbb{E}\big[(f(\mathbf{X}) - f_{\vartheta^*}(\mathbf{X}))(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))\big]\,\mathrm{d}\varrho_\eta(\vartheta)$$
$$\leqslant \mathcal{E}(\vartheta^*) + \int \mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X}) - f_\vartheta(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta)$$

$$+ 2\int \mathbb{E}\big[(f(\mathbf{X})-f_{\vartheta^*}(\mathbf{X}))^2\big]^{1/2}\mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X})-f_{\vartheta}(\mathbf{X}))^2\big]^{1/2}\,\mathrm{d}\varrho_\eta(\vartheta)$$
$$\leqslant \frac{4}{3}\mathcal{E}(\vartheta^*) + 4\int \mathbb{E}\big[(f_{\vartheta^*}(\mathbf{X})-f_{\vartheta}(\mathbf{X}))^2\big]\,\mathrm{d}\varrho_\eta(\vartheta), \tag{7.8}$$

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 $\vartheta \mapsto f_\vartheta(\mathbf{x})$ from Assumption B 2. We obtain

$$\int \mathcal{E}\,\mathrm{d}\varrho_\eta \leqslant \frac{4}{3}\mathcal{E}(\vartheta^*) + \frac{1}{n^2} \qquad \text{for} \qquad \eta = \frac{1}{2\Delta\, n}. \tag{7.9}$$

It remains to bound the Kullback-Leibler term in (7.4) which can be done with the following lemma:

Lemma 20. *Suppose $\mathcal{B}_\eta(\vartheta^*) := \{\vartheta \in \mathbb{R}^Q : |\vartheta-\vartheta^*|_\Theta \leqslant \eta\} \subseteq \Theta$. Then the probability measure $\frac{\mathrm{d}\varrho_\eta}{\mathrm{d}\Pi}(\vartheta) \propto \mathbb{1}_{\mathcal{B}_\eta(\vartheta^*)}$ satisfies*

$$\mathrm{KL}(\varrho_\eta \mid \Pi) = Q\log(2B/\eta) - \log\mathrm{vol}(\mathcal{B}_1),$$

where $\mathrm{vol}(\mathcal{B}_1)$ *denotes the volume of the* $|\cdot|_\Theta$*-unit ball* $\mathcal{B}_1(0)$.

Indeed, $\mathcal{B}_\eta(\vartheta^*) \subseteq \Theta$ for sufficiently large n, so plugging (7.9) and the bound from Lemma 20 into the PAC-Bayes bound (7.4), we conclude

$$\mathcal{E}(\widetilde{\vartheta}_{\lambda,\rho}) \leqslant 12\mathcal{E}(\vartheta^*) + \frac{9}{n^2} + \frac{16}{\lambda}\big(Q\log(4B\Delta n) - \log\mathrm{vol}(\mathcal{B}_1) + \log(2/\delta)\big)$$
$$\leqslant 12\mathcal{E}(\vartheta^*) + \frac{K_1}{n}\big(Q\log(B\Delta n) - \log\mathrm{vol}(\mathcal{B}_1) + \log(2/\delta)\big).$$

for some constant K_1 only depending on $C_f, C_\varepsilon, \sigma$. ■

7.4 Proof of Theorem 6

Due to Remark 17 we can prove analogously to Proposition 19 the following PAC-Bayes bound under Assumption A: For any sample-dependent (in a measurable way) probability measure $\varrho \ll \Pi$ and any $\lambda \in (0, n/w)$ and $\rho \in (0,1]$ such that $C_{n,\lambda} + \frac{\lambda}{n\rho}4(\sigma C_f + \sigma^2) \leqslant \frac{1}{4}$, 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 at least $1-\delta$. From here we can continue as in Section 7.3. ■

7.5 Proof of Theorem 8

Choosing $\lambda = \frac{n}{pK_0}$, Theorem 4 and Corollary 7 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 $s_n^2 := p\varrho_n^2 + \frac{2p(K_1\vee K_2)}{n}\log\frac{2}{\alpha}$. We conclude

$$\mathbb{P}\big(\,\mathrm{diam}(\widehat{C}(\tau_\alpha)) \leqslant 4s_n\big)$$

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

The first statement in Theorem 8 is thus verified.
For the coverage statement, we denote $\bar\xi\coloneqq\xi\Delta$ and bound

$$
\begin{aligned}
\mathbb{P}\big(f\in\widehat{C}(\xi\tau_\alpha^{\vartheta_0})\big)&=\mathbb{P}\big(\|f-\bar f_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})}\leqslant\xi\tau_\alpha^{\vartheta_0}\big)\\
&\geqslant\mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\vartheta:|\vartheta-\vartheta_0|_\Theta\leqslant\bar\xi^{-1}\|f-\bar f_{\lambda,\rho}\|_{L^2(\mathbb{P}^{\mathbf{X}})}\mid\mathcal{D}_n)<1-\alpha\big)\\
&\geqslant\mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\vartheta:|\vartheta-\vartheta_0|_\Theta\leqslant\bar\xi^{-1}s_n\mid\mathcal{D}_n)<1-\alpha\big)-\alpha^2\\
&=1-\alpha^2-\mathbb{P}\big(\widetilde{\Pi}_{\lambda,\rho}(\vartheta:|\vartheta-\vartheta_0|_\Theta\leqslant\bar\xi^{-1}s_n\mid\mathcal{D}_n)\geqslant1-\alpha\big)\\
&\geqslant1-\alpha^2-(1-\alpha)^{-1}\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)\mid\mathcal{D}_n)\big]
\end{aligned}
$$

with balls

$$
\mathcal{B}_\kappa(\vartheta_0)=\big\{\vartheta\in\Theta:|\vartheta-\vartheta_0|_\Theta\leqslant\kappa\big\},\qquad\kappa>0.
$$

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 Hölder imply for $p\geqslant2$

$$
\begin{aligned}
&\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)\mid\mathcal{D}_n)\big]\\
&\quad=\mathbb{E}\Big[\widetilde{D}_\lambda^{-1}\int_{\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)}\mathrm{e}^{-\lambda\widetilde{R}_{n,\rho}(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big]\\
&\quad=\mathbb{E}\Big[\widetilde{D}_\lambda^{-1}\mathrm{e}^{-\lambda\widetilde{R}_{n,\rho}(f)}\int_{\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)}\mathrm{e}^{-\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big]\\
&\quad\leqslant\mathbb{E}\big[\widetilde{D}_\lambda^{-2}\mathrm{e}^{-2\lambda\widetilde{R}_{n,\rho}(f)}\big]^{1/2}\mathbb{E}\Big[\Big(\int_{\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)}\mathrm{e}^{-\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big)^2\Big]^{1/2}\\
&\quad\leqslant\mathbb{E}\big[\widetilde{D}_\lambda^{-2}\mathrm{e}^{-2\lambda\widetilde{R}_{n,\rho}(f)}\big]^{1/2}\mathbb{E}\Big[\Pi(\mathcal{B}_{s_n/\bar\xi}(\vartheta_0))^{2(1-1/p)}\Big(\int_{\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)}\mathrm{e}^{-p\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big)^{2/p}\Big]^{1/2}.
\end{aligned}
$$

Abbreviating $\mathcal{B}_1=\mathcal{B}_1(0)$ as above, we note that the uniform prior yields

$$
\Pi(\mathcal{B}_{s_n/\bar\xi}(\vartheta_0))\leqslant\Pi(\mathcal{B}_{s_n/\bar\xi}(0))=\big(\frac{s_n}{\bar\xi}\big)^Q\Pi(\mathcal{B}_1)
$$

which is not random. To bound the expectation of the integral we extend the integration domain to $D=\Theta$ if ϑ_0 is data dependent and keep $D=\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)$ otherwise. The smaller choice of $\lambda=n/(pK_0)$ instead of n/K_0 ensures $C_{n,p\lambda}+\frac{p\lambda}{n}(\sigma C_f+\sigma^2)\leqslant\frac18$ allowing us to apply Proposition 16 with $p\lambda$. The second factor can thus be bounded using

$$\begin{aligned}\mathbb{E}\Big[\int_D \mathrm{e}^{-p\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\,\Pi(\mathrm{d}\vartheta)\Big]&=\int_D\mathbb{E}\big[\mathrm{e}^{-p\lambda\widetilde{\mathcal{E}}_n(\vartheta)}\big]\,\Pi(\mathrm{d}\vartheta)\\&\leqslant\int_D\exp\big(\big(C_{n,p\lambda}+\tfrac34+\tfrac{p\lambda}{n}(\sigma C_f+\sigma^2)-1\big)p\lambda\mathcal{E}(\vartheta)\big)\,\Pi(\mathrm{d}\vartheta)\\&\leqslant\Pi(D).\end{aligned}$$

Based on (7.7), we conclude with $R_p=\big(\frac{s_n}{\bar\xi}\big)^{-Q/p}\Pi(\mathcal{B}_1)^{-1/p}$ for a data-dependent ϑ_0 and $R_p=1$ otherwise that

$$\begin{aligned}&\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)\mid\mathcal{D}_n)\big]\\&\quad\leqslant\Pi(\mathcal{B}_1)\big(\frac{s_n}{\bar\xi}\big)^QR_q\mathbb{E}\big[\widetilde{D}_\lambda^{-2}\mathrm{e}^{-2\lambda\widetilde{R}_{n,\rho}(f)}\big]^{1/2}\\&\quad=\Pi(\mathcal{B}_1)\big(\frac{s_n}{\bar\xi}\big)^QR_q\mathbb{E}\Big[\exp\Big(\inf_{\varrho\ll\Pi}\Big(2\,\mathrm{KL}(\varrho\mid\Pi)+2\int\lambda\widetilde{R}_{n,\rho}(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big)-2\lambda\widetilde{R}_{n,\rho}(f)\Big)\Big]^{1/2}\\&\quad=\Pi(\mathcal{B}_1)\big(\frac{s_n}{\bar\xi}\big)^QR_q\mathbb{E}\Big[\exp\Big(\inf_{\varrho\ll\Pi}\Big(2\,\mathrm{KL}(\varrho\mid\Pi)+\int2\lambda\widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big)\Big)\Big]^{1/2}.\end{aligned}$$

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

$$\frac{\mathrm{d}\varrho_{\eta'}}{\mathrm{d}\Pi}(\vartheta)\propto\mathbb{1}_{\{|\vartheta-\vartheta^*|_\Theta\leqslant\eta'\}},\qquad\eta'=\frac{s_n}{4\Delta\sqrt{\log(B\Delta n)}}.$$

we can moreover bound using (7.8), Assumption B and Lemma 20

$$\begin{aligned}&\inf_{\varrho\ll\Pi}\Big(\mathrm{KL}(\varrho\mid\Pi)+\int\lambda\widetilde{\mathcal{E}}_n(\vartheta)\,\mathrm{d}\varrho(\vartheta)\Big)\\&\qquad\leqslant\mathrm{KL}(\varrho_{\eta'}\mid\Pi)+\frac43\lambda\mathcal{E}(\vartheta^*)+4\lambda\int\mathbb{E}\big[\big(f_{\vartheta^*}(\mathbf{X})-f_\vartheta(\mathbf{X})\big)^2\big]\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\&\qquad\qquad+\lambda\int\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\\&\qquad\leqslant Q\log\frac{1}{\eta'}-\log\Pi(\mathcal{B}_1)+\frac43\lambda\mathcal{E}(\vartheta^*)+\frac{\lambda s_n^2}{\log(B\Delta n)}\\&\qquad\qquad+\lambda\int\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta).\end{aligned}$$

In the sequel $K_i>0,i=5,6,\dots$, are numerical constants which may depend on $C_f,C_\varepsilon,\sigma,d$ and α. Since $\log(B\Delta n)\mathcal{E}(\vartheta^*)\leqslant s_n^2\leqslant K_5Q\log(B\Delta n)/\lambda$ by assumption, we obtain

$$\begin{aligned}\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar\xi}(\vartheta_0)\mid\mathcal{D}_n)\big]&\leqslant R_q\exp\big(-Q\log\bar\xi+Q\log\big(4\Delta\sqrt{\log(B\Delta n)}\big)+3K_5Q\big)\\&\qquad\times\mathbb{E}\Big[\exp\Big(2\lambda\int\big(\widetilde{\mathcal{E}}_n(\vartheta)-\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\Big)\Big]^{1/2}\end{aligned}$$

$$\leqslant R_q \exp\big(-Q\log\xi + Q(K_6 + \log\sqrt{\log(B\Delta n)})\big)$$
$$\times\, \mathbb{E}\Big[\int \exp\big(2\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\Big]^{1/2}$$

applying Jensen's inequality in the last line. To bound the expectation in the previous line, Fubini's theorem, Proposition 16 with $C_{n,2\lambda} + \frac{2\lambda}{n}(\sigma C_f + \sigma^2) \leqslant \frac{1}{8}$ and Lemma 21 imply

$$\begin{aligned}
\mathbb{E}\Big[\int \exp\big(2\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta)\Big] &= \int \mathbb{E}\big[\exp\big(2\lambda\big(\widetilde{\mathcal{E}}_n(\vartheta) - \mathcal{E}(\vartheta)\big)\big)\big]\,\mathrm{d}\varrho_{\eta'}(\vartheta) \\
&\leqslant \int \exp\big(2\lambda\big(C_{n,2\lambda} + \tfrac{2\lambda}{n}(\sigma C_f + \sigma^2)\big)\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta) \\
&\leqslant \int \exp\big(\tfrac{1}{4}\lambda\mathcal{E}(\vartheta)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta) \\
&\leqslant \int \exp\big(\tfrac{1}{2}\lambda\big(\mathcal{E}(\vartheta^*) + \|f_\vartheta - f_{\vartheta^*}\|^2_{L^2(\mathbb{P}^{\mathbf{X}})}\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta) \\
&\leqslant \int \exp\big(\tfrac{1}{2}\lambda\big(\mathcal{E}(\vartheta^*) + s_n^2/(\log(B\Delta n)\big)\big)\,\mathrm{d}\varrho_{\eta'}(\vartheta) \\
&\leqslant \mathrm{e}^{K_5 Q}.
\end{aligned}$$

We conclude

$$\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar{\xi}}(\vartheta_0) \mid \mathcal{D}_n)\big] \leqslant R_p \exp\big(-Q\big(\log\xi - K_6 - K_5 - \log\sqrt{\log(B\Delta n)}\big)\big).$$

If $R_p = 1$, we obtain for a sufficiently large $K_7 > 0$ and $\xi = K_7\sqrt{\log(B\Delta n)}$ that $\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar{\xi}}(\vartheta_0) \mid \mathcal{D}_n)\big] \leqslant \alpha(1-\alpha)^2$ and thus

$$\mathbb{P}\big(f \in \widehat{C}(\xi\tau_\alpha^{\vartheta_0})\big) \geqslant 1 - \alpha^2 - \alpha(1-\alpha) \geqslant 1 - \alpha.$$

If $R_p = \big(s_n/\bar{\xi}\big)^{-Q/p}\Pi(\mathcal{B}_1)^{-1/p}$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}(\mathcal{B}_{s_n/\bar{\xi}}(\vartheta_0) \mid \mathcal{D}_n)\big] \\
&\quad\leqslant \Pi(\mathcal{B}_1)^{-1/p}\exp\big(-Q\big(\log\xi - \log\big((\Delta\xi/s_n)^{1/p}\big) - K_6 - K_5 - \log\sqrt{\log(B\Delta n)}\big)\big)
\end{aligned}$$

such that we need $\xi = \big(\Delta/s_n\big)^{1/p'}$ for any $p' < p$ and sufficiently large n. The claim the follows by choosing $p' = p/2$. ■

7.6 Proofs for Section 4

To prove Theorem 10, Proposition 11 and Theorem 12, we need the following bound:

Lemma 21. *Let $\vartheta, \widetilde{\vartheta} \in [-B, B]^Q$. Then we have for $\mathbf{x} \in \mathbb{R}^d$ that*

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

If $L = 1$, we also have for all $\mathbf{x} \in \mathbb{R}^d$

$$|f_\vartheta(\mathbf{x}) - f_{\widetilde{\vartheta}}(\mathbf{x})| \leqslant B(|\mathbf{x}|_1 + 1)|\vartheta - \widetilde{\vartheta}|_1 \leqslant (d+3)rB(|\mathbf{x}|_1 + 1)|\vartheta - \widetilde{\vartheta}|_\infty.$$

7.6.1 Proof of Theorem 10 and Theorem 12

It remains to verify Assumption B for the considered class of neural networks where we choose $|\cdot|_\Theta = |\cdot|_\infty$. While the boundedness is ensured by construction of $\mathcal{F}(d, L, r, C_f)$, the Lipschitz continuity with respect to $|\cdot|_\Theta = |\cdot|_\infty$ is verified by Lemma 21. In particular,

$$\|f_\vartheta - f_{\widetilde{\vartheta}}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant 4(2rB)^L|\vartheta - \widetilde{\vartheta}|_\infty \mathbb{E}[|\mathbf{X}|_1^2 \vee 1]^{1/2} \leqslant 8\mathbb{E}[|\mathbf{X}|^2]d(2rB)^L|\vartheta - \widetilde{\vartheta}|_\infty$$

or $\|f_\vartheta - f_{\widetilde{\vartheta}}\|_{L^2(\mathbb{P}^{\mathbf{X}})} \leqslant (d+3)rB(\mathbb{E}[|\mathbf{X}|^2]+1)|\vartheta - \widetilde{\vartheta}|_\infty$ for $L = 1$. The statement is then an immediate consequence of Theorem 4. ■

7.6.2 Proof of Proposition 11

This is a straight-forward application of Theorem 8 in combination with Lemma 21, where we used $|\cdot|_\Theta = |\cdot|_1$ with $\Delta = B\mathbb{E}[(|\mathbf{X}|_1 + 1)^2]^{1/2}$ for the credibility statement. ■

7.6.3 Proof of Theorem 14

The outline of the proof Theorem 14 is similar to that of Theorem 4. To bound the Kullback-Leibler term in (7.10), we employ the following modification of Lemma 20:

Lemma 22. *We have* $\mathrm{KL}(\varrho_{r,\eta} \mid \check{\Pi}) \leqslant Q_r \log(2B/\eta) + r$.

Note that the only property of the prior that we used in the proof of Proposition 19 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 $\check{\Pi}$. We obtain with probability at least $1 - \delta$

$$\mathcal{E}(\check{\vartheta}_{\lambda,\rho}) \leqslant 9\int \mathcal{E}\,\mathrm{d}\varrho + \frac{16}{\lambda}\big(\mathrm{KL}(\varrho \mid \check{\Pi}) + \log(2/\delta)\big). \tag{7.10}$$

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 (4.2). Replacing ϑ^* with ϑ_r^* in the arguments from Section 7.3 and verifying Assumption B 2 as in Section 7.6.1, we find

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

Owing to Lemma 22, we thus have with probability $1 - \delta$

$$\mathcal{E}(\check{\vartheta}_{\lambda,\rho}) \leqslant 12\mathcal{E}(f_{\vartheta_r^*}) + \frac{K_1}{n}\big(Q_r L \log(rBn) + \log(2/\delta)\big),$$

for some $K_1 > 0$ only depending on $C_f, C_\varepsilon, \sigma$. Choosing r to minimize the upper bound in the last display yields the assertion. ■

7.6.4 Proof of Lemma 9

By Yang and Zhou (2024, Theorem 2) for any $f\in\mathcal{C}^\beta([0,1]^d,c_0)$ there is a shallow network $g_\vartheta(\mathbf{x})=\sum_{i=1}^r W_i^{(2)}\phi(W_{i,\cdot}^{(1)}\mathbf{x}+v_i^{(1)})$ with $|W_{i,\cdot}^{(1)}|^2+|v_i^{(1)}|^2=1$ for all $i=1,\dots,r$ ($v^{(2)}=0$) and $|W^{(2)}|_1\leqslant M$ such that

$$\|g_\vartheta-f\|_\infty\leqslant C_{\mathrm{shallow}}r^{-\beta/d}\vee M^{-2\beta/(d+3-2\beta)}.$$

Choosing $M=r^{(d+3-2\beta)/(2d)}$, we obtain $\|g_\vartheta-f\|_\infty\leqslant C_{\mathrm{shallow}}r^{-\beta/d}$. Note that the normalization of $W_{i,\cdot}^{(1)}$ and $v_i^{(1)}$ implies $|W_{ij}^{(1)}|\leqslant 1$ and $|v_i^{(1)}|\leqslant 1$ for all $i=1,\dots,r,j=1,\dots,d$.

To obtain a uniform bound on the entries of the vector $W^{(2)}\in\mathbb{R}^d$, we note that for each entry with $|W_i^{(2)}|\geqslant B$ we can reproduce $W_i^{(2)}\phi(W_{i,\cdot}^{(1)}\mathbf{x}+v_i^{(1)})$ by summing over $\lceil |W_i^{(2)}|/B\rceil$ copies of the neuron $\phi(W_{i,\cdot}^{(1)}\mathbf{x}+v_i^{(1)})$ weighted with weights bounded by B. As a result we can reproduce g_ϑ with a shallow network with at most $r+|W^{(2)}|_1/B\leqslant r+M/B$ neurons and weights uniformly bounded by B. Choosing $B=1\vee r^{(3-d-2\beta)/(2d)}$ ensures that $M/B\leqslant r$. ■

7.7 Remaining proofs for Section 3

7.7.1 Proof of Lemma 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

$$\frac1\lambda\sum_{i=1}^n\log\big(\rho\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}+1-\rho\big)\geqslant\frac1\lambda\sum_{i=1}^n\rho\log\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}+(1-\rho)\log 1=-\frac1n\sum_{i=1}^n\ell_i(\vartheta)=-R_n(\vartheta).$$

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

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

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

$$S_n(\vartheta)=\frac1\lambda\sum_{i=1}^n\Big(\log\big(\rho(\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}-1)+1\big)+\frac\lambda n\ell_i(\vartheta)\Big)$$

$$\leqslant \frac{\rho}{\lambda}\sum_{i=1}^n \big(\mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)} - 1 + \tfrac{\lambda}{n\rho}\ell_i(\vartheta)\big)$$
$$\leqslant \frac{\rho}{2\lambda}\sum_{i=1}^n \big(\tfrac{\lambda}{n\rho}\ell_i(\vartheta)\big)^2 \mathrm{e}^{-\frac{\lambda}{n\rho}\ell_i(\vartheta)}$$
$$\leqslant \frac{\lambda}{n\rho}\cdot\frac{1}{2n}\sum_{i=1}^n |\ell_i(\vartheta)|^2.$$

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

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

and thus

$$\frac{1}{\lambda}\,\mathrm{KL}\,\big(\bar\Pi_{\lambda,\rho}(\cdot\mid\mathcal{D}_n)\,\big|\,\Pi_\lambda(\cdot\mid\mathcal{D}_n)\big) \leqslant \frac{\lambda}{n\rho}\Big(64C_f^4 + \frac{4}{n}\sum_{i=1}^n \varepsilon_i^4\Big)\int \bar\Pi(\mathrm{d}\vartheta\mid\mathcal{D}_n)$$
$$= \frac{\lambda}{n\rho}\Big(64C_f^4 + \frac{4}{n}\sum_{i=1}^n \varepsilon_i^4\Big).$$

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

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

such that

$$\mathrm{KL}\,\big(\bar\Pi_{\lambda,\rho}(\cdot\mid\mathcal{D}_n)\,\big|\,\varpi_{\lambda,\rho}(\cdot\mid\mathcal{D}_n)\big) = \int \big(T_n(\vartheta) - \lambda\bar R_{n,\rho}(\vartheta)\big)\,\bar\Pi_{\lambda,\rho}(\mathrm{d}\vartheta\mid\mathcal{D}_n) + \log\frac{D_{\varpi,\lambda}}{\bar D_\lambda}.$$

We have

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

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 $\log \frac{D_{\varpi,\lambda}}{\bar{D}_\lambda} \leqslant -n\log(1-\rho)$ and thus

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

7.7.2 Proof of Lemma 3

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 (7.6) that

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

7.7.3 Proof of Corollary 7

Jensen's and Markov's inequality yield for ϱ_n^2 from (3.3) that

$$\begin{aligned}
&\mathbb{P}\Big(\mathcal{E}(\bar{f}_{\lambda,\rho}) > \varrho_n^2 + \frac{K_1}{n} + \frac{K_1}{n}\log(2/\delta)\Big) \\
&\qquad= \mathbb{P}\Big(\|\mathbb{E}[f_{\widetilde{\vartheta}_{\lambda,\rho}} \mid \mathcal{D}_n] - f\|_{L^2(\mathbb{P}^{\mathbf{X}})}^2 > \varrho_n^2 + \frac{K_1}{n} + \frac{K_1}{n}\log(2/\delta)\Big) \\
&\qquad\leqslant \mathbb{P}\Big(\mathbb{E}\big[\|f_{\widetilde{\vartheta}_{\lambda,\rho}} - f\|_{L^2(\mathbb{P}^{\mathbf{X}})}^2 \,\big|\, \mathcal{D}_n\big] > \varrho_n^2 + \frac{K_1}{n} + \frac{K_1}{n}\log(2/\delta)\Big)
\end{aligned}$$

$$= \mathbb{P}\Big(\int_{\frac{K_1}{n}\log(2/\delta)}^{\infty} \widetilde{\Pi}_{\lambda,\rho}\big(\|f_{\widetilde{\vartheta}_{\lambda,\rho}} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} > \varrho_n^2 + t \,\big|\, \mathcal{D}_n\big)\,\mathrm{d}t > \frac{K_1}{n}\Big)$$
$$\leqslant \frac{n}{K_1}\int_{\frac{K_1}{n}\log(2/\delta)}^{\infty} \mathbb{E}\big[\widetilde{\Pi}_{\lambda,\rho}\big(\|f_{\widetilde{\vartheta}_{\lambda}} - f\|^2_{L^2(\mathbb{P}^{\mathbf{X}})} > \varrho_n^2 + t \,\big|\, \mathcal{D}_n\big)\big]\,\mathrm{d}t.$$

Using Theorem 4, we thus obtain

$$\mathbb{P}\Big(\mathcal{E}(\bar{f}_{\lambda,\rho}) > \varrho_n^2 + \frac{K_1}{n} + \frac{K_1}{n}\log(2/\delta)\Big) \leqslant \frac{2n}{K_1}\int_{\frac{K_1}{n}\log(2/\delta)}^{\infty} \mathrm{e}^{-nt/K_1}\,\mathrm{d}t = \delta. \qquad \blacksquare$$

7.7.4 Proof of Proposition 13

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

$$f = f_p \circ \cdots \circ f_0 = h_p \circ \cdots \circ h_0$$

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 $i = 1, \dots, p-1$ and $\widetilde{h}_{pj} \in \mathcal{C}_{t_p}^{\beta_q}\big([0,1]^{t_p}, C_0(2C_0)^{\beta_p}\big)$ 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 $Q_{L_i,r_i} = c_i M^{t_i}$ parameters and

$$\|\widetilde{h}_{ij} - \widetilde{g}_{ij}\|_{L^\infty([0,1]^{t_i})} \leqslant C_i M_i^{-2\beta_i}. \tag{7.11}$$

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

1. $L_i \geqslant 5 + \lceil \log_4(M^{2\beta_i})\rceil\big(\lceil \log_2(\max\{\lfloor\beta_i\rfloor, t_i\} + 1\rceil + 1\big)$ and
2. $r_i \geqslant 2^{t_i+6}\binom{t_i+\lfloor\beta_i\rfloor}{t_i} t_i^2(\lfloor\beta_i\rfloor + 1)M_i^{t_i}$

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 (7.11). 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

$$\|h_{ij} - g_{ij}\|_{L^\infty([0,1]^{d_i})} \leqslant \|\widetilde{h}_{ij} - \widetilde{g}_{ij}\|_{L^\infty([0,1]^{t_i})} \leqslant C_i M_i^{-2\beta_i}. \tag{7.12}$$

Setting $g = g_p \circ \cdots \circ g_0$ with $g_i = (g_{ij})_j$ we obtain a neural network $g \in \mathcal{G}(d, L, r)$ with $r = \max_{i=0,\dots,p} r_i d_{i+1}$ and $L = \sum_{i=0}^p L_i$.
Counting the number of parameters of g and using $L_i = C_i M_i^{t_i}$, we get

$$Q_{L,r} \leqslant K_8 \sum_{i=0}^{p} L_i r_i^2$$

for some $K_8 > 0$.

It follows from Schmidt-Hieber (2020, Lemma 3) and (7.12) that

$$\|f-g\|_{L^\infty([0,1]^d)} \leqslant C_0 \prod_{l=0}^{p-1}(2C_0)^{\beta_{l+1}} \sum_{i=0}^{p} \big\||h_i-g_i|_\infty\big\|_{L^\infty([0,1]^{d_i})}^{\prod_{l=i+1}^{p}\beta_l\wedge 1} \leqslant K_9 \sum_{i=0}^{p} M_i^{-2\beta_i},$$

for some $K_9 > 0$.

Applying Theorem 4 together with $\mathcal{E}(f_{\vartheta^*}) \leqslant \|f-g\|^2_{L^\infty([0,1]^d)}$ we now obtain

$$\mathcal{E}(\widetilde{f}_{\lambda,\rho}) \leqslant K_{10} \sum_{i=0}^{p} M_i^{-4\beta_i} + \frac{K_{10}}{n}\sum_{i=0}^{p} M_i^{2t_i}(\log n)^3 + K_{10}\frac{\log(2/\delta)}{n} \tag{7.13}$$

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

$$M_i = \Big\lceil \Big(\frac{n}{(\log n)^3}\Big)^{1/(4\beta_i+2t_i)} \Big\rceil$$

ensures $L, r \leqslant n$ for sufficiently large n, balances the first two terms in the upper bound (7.13) and thus yields the asserted convergence rate for $\widetilde{f}_{\lambda,\rho}$.

The convergence rate for the posterior mean can be proved analogously using Corollary 7. ■

7.7.5 Proof of Corollary 15

The statement follows by choosing L in the upper bound from Theorem 14 as in the statement of Proposition 13 and then using the same approximation result to control excess-risk of the corresponding oracle choice ϑ^*_r. ■

7.8 Proofs of the auxiliary results

7.8.1 Proof of Lemma 20

Denoting $\mathcal{B}_\eta(\vartheta^*) = \{\vartheta \in \mathbb{R}^Q : |\vartheta - \vartheta^*|_\Theta \leqslant \eta\}$, we have $\frac{\mathrm{d}\varrho_\eta}{\mathrm{d}\Pi}(\vartheta) = \mathbb{1}_{\mathcal{B}_\eta(\vartheta^*)}/\Pi(\mathcal{B}_\eta(\vartheta^*))$. If $\mathcal{B}_\eta(\vartheta^*) \subseteq [-B,B]^Q$, the uniformity of Π yields

$$\begin{aligned}\mathrm{KL}(\varrho_\eta \mid \Pi) = \int \log\Big(\frac{\mathrm{d}\varrho_\eta}{\mathrm{d}\Pi}\Big)\mathrm{d}\varrho_\eta = -\log\big(\Pi(\mathcal{B}_\eta(\vartheta^*)\big) &= -\log\big(\Pi(\mathcal{B}_\eta(0)\big) \\ &= Q\log(2B/\eta) - \log \mathrm{vol}(\mathcal{B}_1(0)). \quad ■\end{aligned}$$

7.8.2 Proof of Lemma 21

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 $\widetilde{\mathbf{x}}^{(l)}$, $l = 0,\dots,L+1$, analogously to (4.1). We can recursively deduce from the Lipschitz-continuity of ϕ that for $l = 2,\dots,L$:

$$|\mathbf{x}^{(1)}|_1 \leqslant |W^{(1)}\mathbf{x}|_1 + |v^{(1)}|_1$$

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

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 $y \mapsto (-C_f) \vee (y \wedge C_f)$ has Lipschitz constant 1, we conclude

$$
\begin{aligned}
|f_\vartheta(\mathbf{x}) - f_{\widetilde{\vartheta}}(\mathbf{x})| &\leqslant |g_\vartheta(\mathbf{x}) - g_{\widetilde{\vartheta}}(\mathbf{x})|\\
&= |\mathbf{x}^{(L+1)} - \widetilde{\mathbf{x}}^{(L+1)}|\\
&= |W^{(L+1)}\mathbf{x}^{(L)} + v^{(L+1)} - \widetilde{W}^{(L+1)}\widetilde{\mathbf{x}}^{(L)} - \widetilde{v}^{(L+1)}|\\
&\leqslant |(W^{(L+1)} - \widetilde{W}^{(L+1)})\mathbf{x}^{(L)}| + |\widetilde{W}^{(L+1)}(\mathbf{x}^{(L)} - \widetilde{\mathbf{x}}^{(L)})| + |v^{(L+1)} - \widetilde{v}^{(L+1)}|\\
&\leqslant |W^{(L+1)} - \widetilde{W}^{(L+1)}|_\infty |\mathbf{x}^{(L)}|_1 + |\widetilde{W}^{(L+1)}|_\infty |\mathbf{x}^{(L)} - \widetilde{\mathbf{x}}^{(L)}|_1 + \eta\\
&\leqslant \eta(2rB)^L(|\mathbf{x}|_1 \vee 1) + \eta(rB)^L 2^{L+1}(|\mathbf{x}|_1 \vee 1) + \eta\\
&\leqslant \eta 4(2rB)^L(|\mathbf{x}|_1 \vee 1).
\end{aligned}
$$

For shallow neural networks we proceed slightly differently. We have

$$
\begin{aligned}
|\mathbf{x}^{(1)}|_\infty &\leqslant |W^{(1)}|_\infty |\mathbf{x}|_1 + |v^{(1)}|_\infty \leqslant B(|\mathbf{x}|_1 + 1),\\
|\mathbf{x}^{(1)} - \widetilde{\mathbf{x}}^{(1)}|_1 &\leqslant |W^{(1)} - \widetilde{W}^{(1)}|_1 |\mathbf{x}|_\infty + |v^{(1)} - \widetilde{v}^{(1)}|_1
\end{aligned}
$$

and thus

$$
\begin{aligned}
|f_\vartheta(\mathbf{x}) - f_{\widetilde{\vartheta}}(\mathbf{x})| &\leqslant |g_\vartheta(\mathbf{x}) - g_{\widetilde{\vartheta}}(\mathbf{x})|\\
&\leqslant |(W^{(2)} - \widetilde{W}^{(2)})\mathbf{x}^{(1)}| + |\widetilde{W}^{(2)}(\mathbf{x}^{(1)} - \widetilde{\mathbf{x}}^{(1)})| + |v^{(2)} - \widetilde{v}^{(2)}|\\
&\leqslant |W^{(2)} - \widetilde{W}^{(2)}|_1 |\mathbf{x}^{(1)}|_\infty + |\widetilde{W}^{(2)}|_\infty |\mathbf{x}^{(1)} - \widetilde{\mathbf{x}}^{(1)}|_1 + |v^{(2)} - \widetilde{v}^{(2)}|\\
&\leqslant B(|\mathbf{x}|_1 + 1)|\vartheta - \widetilde{\vartheta}|_1.
\end{aligned}
$$

■

7.8.3 PROOF OF LEMMA 22

We will show that

$$\frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\breve{\Pi}} = 2^r(1-2^{-n})\frac{\mathrm{d}\varrho_{r,\eta}}{\mathrm{d}\Pi_r}, \tag{7.14}$$

from which we can deduce

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

Since $\varrho_{r,\eta}$ and Π_r are product measures, their KL-divergence is equal to the sum of the KL-divergences in each of the Q_r factors. For each such factor, we are comparing

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

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

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

and the lemma follows.

For (7.14), note that $\varrho_{r,\eta}$ can only assign a positive probability to subsets $A \subseteq [-B, B]^{Q_r}$. Hence,

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

■

Acknowledgement

The authors would like to thank Botond Szabó for helpful comments. SB is supported by DASHH (Data Science in Hamburg - HELMHOLTZ Graduate School for the Structure of Matter) with the grant HIDSS-0002. SB and GK acknowledge support by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy - EXC 2121 Quantum Universe - 390833306. MS and MT acknowledge support by the DFG through project TR 1349/3-1. The empirical studies were enabled by the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany.

References

A. Alexos, A.J. Boyd, and S. Mandt. Structured stochastic gradient MCMC. In *ICML*, volume 162, pages 414–434, 2022.

P. Alquier. User-friendly introduction to PAC-Bayes bounds. *Found. Trends Mach. Learn.*, 17(2):174–303, 2024.

P. Alquier and G. Biau. Sparse single-index model. *J. Mach. Learn. Res.*, 14:243–280, 2013.

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145, 2011.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.*, 17(236):1–41, 2016.

C. Andrieu and G.O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.*, 37(2):697–725, 2009.

M. Anthony and P.L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge Univ. Press, 1999.

J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(6):685–736, 2004.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Stat.*, 37(4):1591–1646, 2009.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Stat.*, 39(5): 2766–2794, 2011.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.*, 18(19):1–53, 2017.

R. Bardenet, A. Doucet, and C.C. Holmes. On Markov chain Monte Carlo methods for tall data. *J. Mach. Lear. Res.*, 18(47):1515–1557, 2017.

A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.

B. Bauer and M. Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Stat.*, 47(4):2261–2285, 2019.

J. Besag. Comments on "Representations of knowledge in complex systems" by U. Grenander and M.I. Miller. *J. R. Stat. Soc. Ser. B*, 56(4):549–581, 1994.

F. Biggs and B. Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10):1280, 2021.

F. Biggs and B. Guedj. Non-vacuous generalisation bounds for shallow neural networks. In *ICML*, volume 162, pages 1963–1981, 2022.

F. Biggs and B. Guedj. Tighter PAC-Bayes generalisation bounds by leveraging example difficulty. In *AISTATS*, volume 206, pages 8165–8182, 2023.

P.G. Bissiri, C.C. Holmes, and S.G. Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B*, 78(5):1103–1130, 2016.

D.M. Blei, A. Kucukelbir, and J.D. McAuliffe. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877, 2017.

I. Castillo and P. Egels. Posterior and variational inference for deep neural networks with heavy-tailed weights. *J. Mach. Learn. Res.*, 26(122):1–58, 2025.

I. Castillo and R. Nickl. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Stat.*, 42(5):1941–1969, 2014.

O. Catoni. *Statistical learning theory and stochastic optimization.* Springer, 2004.

O. Catoni. *PAC-Bayesian supervised classification: The thermodynamics of statistical learning*, volume 56 of *Lecture Notes-Monograph Series.* IMS, 2007.

X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proc. Alg. Lear. Theory*, volume 83, pages 186–211, 2018.

B.-E. Chérief-Abdellatif. Convergence rates of variational inference in sparse deep learning. In *ICML*, volume 119, pages 1831–1842, 2020.

S. Chewi. *Log-Concave Sampling.* unfinished draft, `https://chewisinho.github.io/main.pdf`, 2025.

A.D. Cobb and B. Jalaian. Scaling Hamiltonian Monte Carlo inference for Bayesian neural networks with symmetric splitting. In *UAI*, volume 161, pages 675–685, 2021.

I. Csiszar. *I*-Divergence Geometry of Probability Distributions and Minimization Problems. *Ann. Probab.*, 3(1):146–158, 1975.

A.S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

A.S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1):39–61, 2008.

W. Deng, Q. Feng, L. Gao, F. Liang, and G. Lin. Non-convex learning via replica exchange stochastic gradient MCMC. In *ICML*, volume 119, pages 2474–2483, 2020a.

W. Deng, G. Lin, and F. Liang. A contour stochastic gradient Langevin dynamics algorithm for simulations of multi-modal distributions. In *Adv. Neural Inf. Proc. Syst.*, volume 33, pages 15725–15736, 2020b.

W. Deng, S. Liang, B. Hao, G. Lin, and F. Liang. Interacting contour stochastic gradient Langevin dynamics. In *ICLR*, 2022.

R. DeVore, B. Hanin, and G. Petrova. Neural network approximation. *Acta Numer.*, 30: 327–444, 2021.

M.D. Donsker and S.R.S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time–III. *Commun. Pure and Appl. Math.*, 29(4):389–461, 1976.

S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.

G.K. Dziugaite and D.M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017.

S. Franssen and B. Szabó. Uncertainty quantification for nonparametric regression using empirical Bayesian neural networks. *arXiv:2204.12735*, 2022.

Y. Freund, Y.-A. Ma, and T. Zhang. When is the convergence time of Langevin algorithms dimension independent? A composite optimization viewpoint. *J. Mach. Lear. Res.*, 23: 1–32, 2022.

S. Ghosal and A.W. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Stat. and Prob. Mathematics*. Cambridge Univ. Press, 2017.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

B. Guedj. A primer on PAC-Bayesian learning. In *Proc. 2nd SMF congress*, pages 391–414, 2019.

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.*, 7:264–291, 2013.

B. Guedj and S. Robbiano. PAC-Bayesian high dimensional bipartite ranking. *J. Stat. Plan. Inference*, 196:70–86, 2018.

F. Hellström, G. Durisi, B. Guedj, and M. Raginsky. Generalization bounds: Perspectives from information theory and PAC-Bayes. *Found. Trends Mach. Learn.*, 18(1):1–223, 2025.

M. Hoffmann and R. Nickl. On adaptive inference and confidence bands. *Ann. Stat.*, 39(5): 2383–2409, 2011.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

B.T. Knapik, A.W. van der Vaart, and J.H. van Zanten. Bayesian inverse problems with Gaussian priors. *Ann. Stat.*, 39(5):2626–2657, 2011.

M. Kohler and S. Langer. On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Stat.*, 49(4):2231–2249, 2021.

C. Li, C. Chen, D.E. Carlson, and L. Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *AAAI*, volume 30, pages 1788–1794, 2016.

D. Maclaurin and R.P. Adams. Firefly Monte Carlo: Exact MCMC with subsets of data. In *UAI*, pages 543–552, 2014.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

D.A. McAllester. PAC-Bayesian model averaging. In *COLT*, pages 164–170, 1999a.

D.A. McAllester. Some PAC-Bayesian theorems. *Mach. Learn.*, 37(3):355–363, 1999b.

R.M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, pages 113–163. Chapman and Hall/CRC, 2011.

R. Nickl and S. Wang. On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *J. Eur. Math. Soc.*, 26(3):1031–1112, 2022.

S. Patterson and Y.W. Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Adv. Neural Inf. Proc. Syst.*, volume 26, pages 3102–3110, 2013.

M. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 22:1–40, 2021.

N.G. Polson and V. Ročková. Posterior concentration for sparse deep learning. In *Adv. Neural Inf. Proc. Syst.*, volume 31, pages 938–949, 2018.

K. Ray and B. Szabó. Variational Bayes for high-dimensional linear regression with sparse priors. *J. Am. Stat. Assoc.*, 117(539):1270–1281, 2022.

C.P. Robert and G. Casella. *Monte Carlo statistical methods.* Springer, second edition, 2004.

G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996a.

G.O. Roberts and R.L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996b.

J. Rousseau and B. Szabó. Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Ann. Stat.*, 48(4):2155–2179, 2020.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Netw.*, 61:85–117, 2015.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Stat.*, 48(4):1875–1897, 2020.

J. Shawe-Taylor and R.C. Williamson. A PAC analysis of a Bayesian estimator. In *COLT*, pages 2–9, 1997.

M.F. Steffen and M. Trabs. A PAC-Bayes oracle inequality for sparse neural networks. In *MathSEE Symposium 2023*, volume 515 of *Springer Proc. Math. Stat.*, pages 131–151, 2026.

B. Szabó, A.W. van der Vaart, and J.H. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Stat.*, 43(4):1391–1428, 2015.

L. Tinsi and A. Dalalyan. Risk bounds for aggregated shallow neural networks using Gaussian priors. In *COLT*, volume 178, pages 227–253, 2022.

M. Welling and Y.W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pages 681–688, 2011.

T.-Y. Wu, Y.X. Rachel Wang, and W.H. Wong. Mini-batch Metropolis-Hastings with reversible SGLD proposal. *J. Am. Stat. Assoc.*, 117(537):386–394, 2022.

Y. Yang and D.-X. Zhou. Nonparametric regression using over-parameterized shallow ReLU neural networks. *J. Mach. Learn. Res.*, 25:1–35, 2024.

D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94: 103–114, 2017.

A.Y. Zhang and H.H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Stat.*, 48(5):2575–2598, 2020.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *Ann. Stat.*, 48(4):2180–2207, 2020.

R. Zhang, C. Li, J. Zhang, C. Chen, and A.G. Wilson. Cyclical stochastic gradient MCMC for Bayesian deep learning. In *ICLR*, 2020.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inf. Theory*, 52(4):1307–1321, 2006.

W. Zhou, V. Veitch, M. Austern, R.P. Adams, and P. Orbanz. Non-vacuous generalization bounds at the ImageNet scale: A PAC-Bayesian compression approach. In *ICLR*, 2019.