# Estimating optimal PAC-Bayes bounds with Hamiltonian Monte Carlo

**Szilvia Ujváry**[*]
University of Cambridge
sru23@cam.ac.uk

**Gergely Flamich**
University of Cambridge
gf332@cam.ac.uk

**Vincent Fortuin**
Helmholtz AI, TU Munich
vincent.fortuin@helmholtz-munich.de

**José Miguel Hernández Lobato**
University of Cambridge
jmh233@cam.ac.uk

## Abstract

An important yet underexplored question in the PAC-Bayes literature is how much tightness we lose by restricting the posterior family to factorized Gaussian distributions when optimizing a PAC-Bayes bound. We investigate this issue by estimating data-independent PAC-Bayes bounds using the optimal posteriors, comparing them to bounds obtained using MFVI. Concretely, we (1) sample from the optimal Gibbs posterior using Hamiltonian Monte Carlo, (2) estimate its KL divergence from the prior with thermodynamic integration, and (3) propose three methods to obtain high-probability bounds under different assumptions. Our experiments on the MNIST dataset reveal significant tightness gaps, as much as 5-6% in some cases.

## 1 Introduction

PAC-Bayes is a tool to give high-probability generalization bounds for (generalized[2]) Bayesian learning algorithms. The main goal is to produce empirical bounds, also called risk certificates (RC), that are nonvacuous even in complex settings such as neural networks (NN). Current methods for NNs train a Bayesian Neural Network (BNN) use mean-field variational inference (MFVI) with an objective derived from a PAC-Bayes bound [7, 21] to obtain nonvacuous bounds. However, the resulting bounds are only tight for relatively simple datasets (e.g. MNIST, CIFAR-10) and require data-dependent priors [7, 21, 17]. MFVI has been widely promoted in the PAC-Bayes literature as an efficient and accurate alternative to MCMC methods [2, 13]. In the Bayesian learning community, there is more controversy about the expressivity of MFVI [11, 9]. Figure 1 illustrates the trade-offs between MFVI and MCMC methods on the 2 dimensional Rosenbrock function [23].

We contribute to this debate by empirically estimating how tight PAC-Bayes bounds can be, compared to those obtained by MFVI. We consider data-independent bounds since these are always loose, hence the effects of using better weight distributions are more visible. For the bounds in our interest, the optimal (minimizing) probability measure on the weights has the form of a Gibbs distribution [1]. We use Hamiltonian Monte Carlo (HMC) [6, 5] to approximately sample from this distribution and then estimate the bound. We find that, especially for small datasets, our RCs can improve on MFVI significantly, such as an 5-6% improvement over MFVI, resulting in a 10.8% RC on Binary MNIST. This demonstrates the potential to significantly tighten PAC-Bayes bounds by considering more complicated distributions than factorized Gaussians.

---

[*]Corresponding author

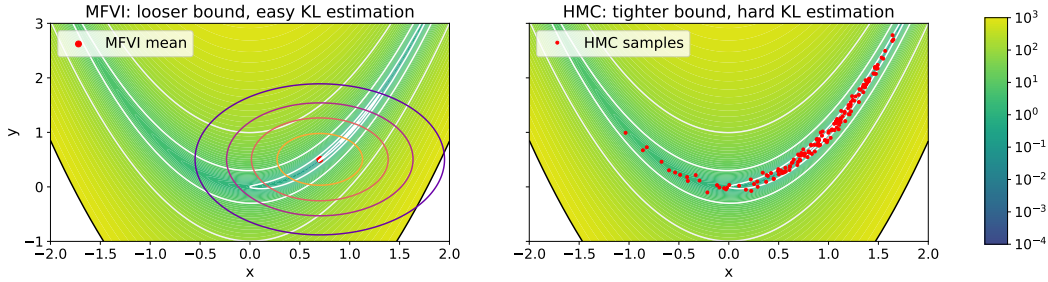[2]generalized Bayesian learning extends Bayesian learning by allowing more general losses and priors [13].

Figure 1: Illustration of the trade-offs of using MFVI and MCMC methods (e.g. HMC) in PAC-Bayes estimation. Our method solves the 'hard KL estimation' problem. The 2-dimensional Rosenbrock function is used as loss function [23]. The prior is centered at (0, 1.5) and has unit variance. Both methods are initialized at (0, 3).

Concretely, our contributions are as follows:

1. We estimate data-independent PAC-Bayes bounds via (1) approximately sampling from the optimal Gibbs posterior, (2) computing the KL divergence with thermodynamic integration and (3) ensuring a high-probability bound.

2. We thereby demonstrate on versions of MNIST that MFVI bounds can be tightened significantly.

3. We justify our approach by extensive diagnostic analysis on our HMC samples.

## 2  Background

We consider supervised classification tasks. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a measurable space, where $\mathcal{X} \subset \mathbb{R}^d$ is the space of data samples, and $\mathcal{Y} \subset \mathbb{R}$ is the space of $K$ labels. Let $\mathcal{M}_1(\mathcal{Z})$ denote the set of probability measures on $\mathcal{Z}$. Let $D$ be an unknown distribution $D \in \mathcal{M}_1(\mathcal{Z})$. A learning algorithm receives a set of $n$ i.i.d. samples $S = (z_1, ..., z_n)$, i.e. $S \sim D^n$. Further, we fix a weight space $\mathcal{W} \subseteq \mathbb{R}^p$ containing all possible weights. Each weight vector $\mathbf{w}$ maps to a *predictor function* $h_{\mathbf{w}} : \mathcal{X} \rightarrow \mathcal{Y}$ that assigns a label $y \in \mathcal{Y}$ to any input $x \in \mathcal{X}$. We seek a predictor function that minimizes the *risk* (expected loss) $L(\mathbf{w}) := \mathbb{E}_{z \sim D}[l(\mathbf{w}, z)]$, where $l : \mathcal{W} \times \mathcal{Z} \rightarrow [0, \infty)$ is a measurable loss function. Let $l^{\text{ce}}$ and $l^{0-1}$ denote the cross-entropy and $0-1$ loss (error) functions, respectively. The data-generating distribution being unknown, $L(\mathbf{w})$ is not observable, hence in practice, we compute the *empirical risk*, which depends on our sample set $S$ and is defined as $\hat{L}_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n l(\mathbf{w}, z_i)$. PAC-Bayes bounds are stated for *randomized predictors*. Given a data sample $x$, a randomized predictor makes a prediction at its label using a random sample of weights $\mathbf{w} \sim Q$, where $Q \in \mathcal{M}_1(\mathcal{W})$. We may identify the randomized predictor with its distribution $Q$. The risk of $Q$ is defined as $L(Q) := \mathbb{E}_{\mathbf{w} \sim Q}[\mathbb{E}_{z \sim D}[l(\mathbf{w}, z)]]$. The empirical risk of $Q$ is given by $\hat{L}_S(Q) := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{w} \sim Q}[l(\mathbf{w}, z_i)]]$. We use notations $L^{\text{ce}}(Q), \hat{L}_S^{\text{ce}}(Q), L^{0-1}(Q)$ and $\hat{L}_S^{0-1}(Q)$ for our risk functionals.

**PAC-Bayes bounds**   We consider the following PAC-Bayes bounds. For proofs, see [19], [25], [1].

**Theorem 2.1.** *Fix a loss function $l : \mathcal{W} \times \mathcal{Z} \rightarrow [0, 1]$ and fix arbitrary $\delta \in (0, 1)$. For any data-free distribution $P$ over $\mathcal{W}$, simultaneously for all distributions $Q$ over $\mathcal{W}$, with probability at least $1 - \delta$, we have*

$$\textbf{kl bound:}\quad \text{kl}(\hat{L}_S(Q)||L(Q)) \leq \frac{\text{KL}(Q||P) + \log(\frac{2\sqrt{n}}{\delta})}{n}. \tag{1}$$

*Further, simultaneously for all $Q \in \mathcal{W}$ and $\lambda \in (0, 2)$, we have the below relaxation of the kl bound*

$$\boldsymbol{\lambda} \textbf{ bound:}\quad L(Q) \leq \frac{\hat{L}_S(Q)}{1 - \frac{\lambda}{2}} + \frac{\text{KL}(Q||P) + \log(\frac{2\sqrt{n}}{\delta})}{n\lambda(1 - \frac{\lambda}{2})}. \tag{2}$$

**Theorem 2.2.** *For fixed $\lambda \in (0, 2)$, the optimal stochastic predictor corresponding to the $\lambda$ bound with prior $P$ that has density $p(\mathbf{w})$ is a Gibbs measure $Q_{\lambda}^*$ with density $q_{\lambda}^*(\mathbf{w}|z) = e^{-n\lambda\hat{L}_S(\mathbf{w})}p(\mathbf{w})/\mathbb{E}_{\mathbf{w} \sim P}\left[e^{-n\lambda\hat{L}_S(\mathbf{w})}\right]$.*

**Algorithm 1** PAC-Bayes estimation with HMC

---
Sample $\mathbf{w} \sim Q^*$ with HMC,
Estimate $\hat{L}_S^{0-1}(Q^*)$ from $\mathbf{w}$ using a high-probability bound,                    ▷ Suppl. M. 6.2.5
**for** $\beta$ in [0, 1] discretization **do**
    Sample $\mathbf{w}_\beta \sim Q^{*\beta} \propto e^{-\beta \widetilde{L}_S^{ce}(\mathbf{w})} p(\mathbf{w})$ with HMC,
    Estimate $\widetilde{L}_S^{ce}(Q^{*\beta})$ from $\mathbf{w}_\beta$ using a high-probability bound        ▷ Suppl. M. 6.2.5
**end for**
Estimate $\log p(z)$ with thermodynamic integration,                              ▷ Eqn. 4
Compute $\mathrm{KL}(Q^*||P)$,                                                     ▷ Prop 2.3
Compute bounds in Eqs. 1, 2 using $\hat{L}_S^{0-1}(Q^*)$ and $\mathrm{KL}(Q^*||P)$ .

---

For simplicity, we use the $\lambda$ bound with $\lambda = 1$ fixed. Even this simple setting improves over MFVI risk certificates (RCs) .

**Empirical estimation of PAC-Bayes bounds**  RCs are most often computed using MFVI on a BNN with an objective function derived from a PAC-Bayes bound. For posterior sampling, a bounded version of the cross-entropy loss $\widetilde{l^{ce}}$ (defined in Supplementary Material 6.3.1) is used. We use notation $\widetilde{L}_S^{ce}(Q)$ for when we use the compute empirical risks with $\widetilde{l^{ce}}$. The final RCs can be computed in terms of the $0 - 1$ loss too. The KL divergence of the final approximate posterior with respect to the prior is available analytically. The risk $\hat{L}_S(Q)$ is approximated as a Monte Carlo average over samples $\mathbf{w} \sim Q$. The following concentration inequality is then used to ensure an upper-bound on $L(Q)$ with pre-specified probability $1 - \delta$ [16].

**Theorem 2.3.** *Suppose* $W_1, W_2, ..., W_m \sim Q$ *are i.i.d., and* $\hat{Q}_m = \sum_{j=1}^m \delta_{W_j}$ *is their empirical distribution. Then for any* $\delta' \in (0,1)$ *with probability* $1 - \delta'$, *we have that* $\hat{L}_S(Q) \leq \mathrm{kl}^{-1}\left(\hat{L}_S(\hat{Q}_m), \frac{1}{m}\log(\frac{2}{\delta'})\right)$ .

# 3 Method

Our goal is to adapt the above pipeline to more general posterior approximations, specifically HMC samples with a Gibbs target density. Our method is summarized in Algorithm 1. The key challenges are estimating $\mathrm{KL}(Q^*||P)$ and the fact that MCMC samples are not independent.

**Approximate sampling from the Gibbs posterior**  As HMC requires gradients of the unnormalized target distribution, the $\widetilde{l^{ce}}$ loss is used to generate samples with the numerator of the Gibbs density formula in Theorem 2.2 as the target density.

**Estimating the KL divergence**  Having obtained approximate samples from $Q^*$ as described above, we now estimate $\mathrm{KL}(Q^*||P)$ by reducing the problem to (generalised) marginal likelihood estimation via the following fact.

**Proposition 3.1.** *Fix a prior measure* $P$ *with density* $p(\mathbf{w})$ *and define a (possibly generalized) likelihood* $p(z|\mathbf{w})$. *For any distribution* $Q$ *that is dependent on the data* $z$, *with density* $q(\mathbf{w}|z)$ *the following holds*

$$\mathrm{KL}(q(\mathbf{w}|z)||p(\mathbf{w})) = \mathbb{E}_q\left[\log(p(z|\mathbf{w}))\right] - \log(Z) + \mathrm{KL}(q(\mathbf{w}|z)||p(\mathbf{w}|z)). \tag{3}$$

*where* $Z = p(z)$, *the (generalised) marginal likelihood.*

For $Q = Q^*$ the true posterior, the last term is $0$ and the term of most interest becomes the generalised log marginal likelihood $\log(Z)$. We estimate this term using thermodynamic integration [18] as

$$-\log(Z) = \int_0^1 \mathbb{E}_{\mathbf{w} \sim \pi_\beta}\left[n\hat{L}_S(\mathbf{w})\right] \mathrm{d}\beta, \text{ where } \pi_\beta \propto e^{-\beta\hat{L}_S(\mathbf{w})}p(\mathbf{w}). \tag{4}$$

Notice that $\pi_\beta$ is itself a Gibbs distribution, hence computing the thermodynamic integral reduces to sampling from Gibbs distributions for values of $\beta \in [0, 1]$. The details of this derivation are given in Supplementary Material 6.2.3. When computing the integral, to control the tail of our bound, it is necessary to ensure that we overestimate its true value with high probability. We use the trapezium rule, which we show in Supplementary Material 6.2.4 to upper-bound the integral.
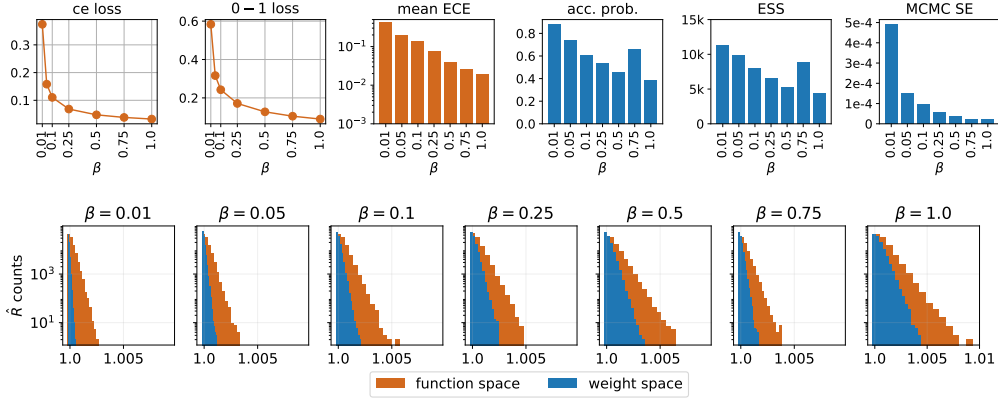
Figure 2: HMC diagnostics for MNIST - half, over 4 chains of length 5000, for each value of $\beta$.

**High-probability upper bounds**    The last step is to control the overall probability of our bound using concentration inequalities on the tails of all of our estimators. Note that Theorem 2.3 used with MFVI is valid only for independent samples. We propose the following three options for the general setting. Each has different assumptions and their comparison serves as a sanity check on the magnitude of our estimates. Derivations are given in Supplementary Material 6.2.5.

- We use Theorem 2.3 after thinning our samples to reduce autocorrelations.
  *This is the simplest option, but we can only guarantee uncorrelatedness.*

- An asymptotic, probability $(1 - \alpha/2)$ confidence interval of the form $[0, c)$, where

$$c = \frac{(1 + \epsilon)\hat{\sigma}_m}{\sqrt{2\alpha m}} + \frac{1}{m}\sum_{i=1}^{m}\hat{L}_S(\mathbf{w}_i). \tag{5}$$

  This bound requires assumptions on the order of the bias and variance of our estimator $\frac{1}{m}\sum_{i=1}^{m}\hat{L}_S(\mathbf{w}_i)$, where $\mathbf{w}_i \sim Q$, and a consistent estimator of its asymptotic variance [24].
  *This gives the tightest RCs, and has reasonable assumptions but they are hard to check.*

- Using a sanity-check bound $\mathrm{KL}(Q||P) \leq n\mathbb{E}_G[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})] + \mathrm{KL}(G||P)$. In Prop. 6.4, we prove this result for any $G$ satisfying $\mathrm{KL}(Q||Q^*) \leq \mathrm{KL}(G||Q^*) + n\mathbb{E}_Q\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right]$, where $Q^*$ is the Gibbs posterior. Taking $G$ to be the Gaussian from MFVI, the assumption only requires that $Q$ is not much further from the true posterior and $G$. If $\mathbb{E}_Q\left[\widetilde{L}_S^{0-1}(\mathbf{w})\right] < \mathbb{E}_G\left[\widetilde{L}_S^{0-1}(\mathbf{w})\right]$, this bound can be tighter than MFVI.
  *This has the mildest assumptions and hence is easiest to estimate, but is the most conservative.*

## 4   Experiments and results

We apply our method to three versions of the MNIST train dataset: Binary MNIST, where labels are mapped $y \mapsto y \pmod 2$, MNIST reduced to $14 \times 14$ pixels and full MNIST. We also compare results when using only half of the 60,000 datapoints. We use a data-independent factorized Gaussian prior and small MLPs. For each configuration, we generate 4 independent HMC chains of 5000 samples each. We compare results to estimating the same RCs with MFVI. All RCs are valid with probability at least 0.95. Further details are given in Supplementary Material 6.3.4.

**HMC diagnostics**    Figure 2 shows HMC diagnostics for the MNIST-half dataset. The rest of the diagnostic results are supplied in Supplementary Material 6.4. We report expected calibration error, HMC acceptance probability, effective sample size, MCMC standard error and $\hat{R}$ statistics along with cross-entropy and $0-1$ losses for all datasets and $\beta$ values. Unsurprisingly, performance increases as $\beta$ increases, since our samples get more similar to the true posterior. The $\hat{R}$ statistic is very close to 1.0 both in weight and function space, indicating approximate convergence in our chains.

4

Table 1: Training and test metrics and RC estimates using our three bounds. **Bold** numbers indicate the tighter certificate out of the Gibbs and MFVI ones for the same dataset.

| Setup | | | Train/test stats | | | 0-1 RC with kl bound | | | 0-1 RC with $\lambda$ bound | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Dataset | Model | Train 0-1 | Test 0-1 | KL/n | kl inverse | asympt | naive | kl inverse | asympt | naive |
| MFVI | Binary Half | 1L | 0.0924 | 0.0982 | 0.0106 | 0.1600 | 0.1426 | 0.1600 | 0.2385 | 0.2090 | 0.2385 |
| Gibbs p. | Binary Half | 1L | 0.0562 | 0.0561 | 0.0205 | **0.1342** | **0.1065** | **0.1417** | **0.1820** | **0.1428** | **0.1877** |
| MFVI | Binary | 1L | 0.0960 | 0.0928 | 0.0105 | 0.1640 | 0.1452 | 0.1640 | 0.2452 | 0.2136 | 0.2452 |
| Gibbs p. | Binary | 1L | 0.0404 | 0.0415 | 0.0195 | **0.1080** | **0.0702** | **0.1184** | **0.1435** | **0.0969** | **0.1566** |
| MFVI | $14 \times 14$ Half | 2L | 0.1348 | 0.1319 | 0.0148 | 0.2449 | 0.2059 | 0.2449 | 0.3595 | 0.3070 | 0.3595 |
| Gibbs p. | $14 \times 14$ Half | 2L | 0.0954 | 0.1010 | 0.0477 | **0.1888** | **0.1805** | **0.2324** | **0.3273** | **0.2481** | **0.3180** |
| MFVI | $14 \times 14$ | 2L | 0.1389 | 0.1313 | 0.0140 | 0.2379 | 0.1991 | 0.2379 | 0.3595 | 0.2930 | 0.3595 |
| Gibbs p. | $14 \times 14$ | 2L | 0.0695 | 0.0723 | 0.0381 | **0.1855** | **0.1335** | **0.1920** | **0.2507** | **0.1810** | **0.2600** |
| MFVI | MNIST Half | 2L | 0.1256 | 0.1264 | 0.0199 | 0.2302 | 0.2025 | 0.2302 | 0.3387 | 0.2911 | 0.3387 |
| Gibbs p. | MNIST Half | 2L | 0.0898 | 0.0970 | 0.0428 | **0.2248** | **0.1740** | **0.2247** | **0.3091** | **0.2377** | **0.3068** |
| MFVI | MNIST | 2L | 0.1236 | 0.1200 | 0.0196 | 0.2070 | 0.1987 | 0.2070 | 0.2977 | 0.1714 | 0.2977 |
| Gibbs p. | MNIST | 2L | 0.0653 | 0.0691 | 0.0334 | **0.1759** | **0.1269** | **0.1880** | **0.2381** | 0.2822 | 0.2553 |

**Bound calculation results**   Table 1 shows the results of our RC estimation, compared to MFVI. Our method always improves $0 - 1$ RCs, most significantly for Binary MNIST, where the gap can reach 10% in some configurations. The gap increases when more data is used. More analysis is given in Supplementary Material 6.5.

## 5   Discussion

**Related work**   Our work opposes the view that MFVI does not considerably loosen PAC-Bayes bounds. The theoretical paper Alquier et al. [2] argues for this view by showing that Gibbs posteriors concentrate around the global minimum of the loss at the same rate as their best Gaussian approximations. However, their result is proven for binary linear classification only, and the best possible Gaussian approximation may not be attainable in practice. Many other works criticize MFVI especially for small neural nets, including Foong et al. [11] and Farquhar et al. [9] in Bayesian learning. In the PAC-Bayes setting, Pitas [22] experimentally demonstrates the limitations of MFVI in PAC-Bayes and improves RCs using a simplified KFAC Laplace approximation. To our knowledge, we are the first to tightly estimate the optimal value of PAC-Bayes bounds (i.e. in the Gibbs posterior), with the closest work being Dziugaite and Roy [8], who attempt this task in the context of data-dependent priors. They find that their bound is very loose in practice, which may be due to not using tempering in the KL estimation. The work Foong et al. [10] evaluates the tightness limits of the whole PAC-Bayes framework (rather than specific bounds) and restricts to very small datasets (30-60 datapoints). Our work builds on Izmailov et al. [14] and Wenzel et al. [26] who scaled HMC up to neural networks.

**Limitations**   The chief limitation of our method is that PAC-Bayes bound estimates require assumptions that are impossible to completely verify for MCMC samples. What is possible are empirical estimates accompanied by robust diagnostics. Hence we make claims under sets of assumptions of varying strength. Even our most conservative estimates demonstrate the inaccuracy of MFVI.

**Conclusion**   We have demonstrated that using MFVI in PAC-Bayes estimation leads to data-independent bounds that are much looser than necessary. We provided estimates of optimal bounds on MNIST versions, using HMC and thermodynamic integration. Our estimates have plausible magnitude and are supported by extensive diagnostic analysis. The improvement over MFVI is largest for small models and adding more data tightens our estimates more than their MFVI approximations. Our results demonstrate the need for better posterior approximations for tight bounds.

## Acknowledgments and Disclosure of Funding

# References

[1] Pierre Alquier. User-friendly introduction to PAC-Bayes bounds, 2023. URL https://arxiv.org/pdf/2110.11216.pdf.

[2] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016. URL http://jmlr.org/papers/v17/15-290.html.

[3] Alexandros Beskos, Natesh S. Pillai, Gareth O. Roberts, Jesus M. Sanz-Serna, and Andrew M. Stuart. Optimal tuning of the Hybrid Monte-Carlo algorithm, 2010. URL https://arxiv.org/pdf/1001.4460.pdf.

[4] Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv: Methodology*, 2017. URL https://api.semanticscholar.org/CorpusID:88514713.

[5] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC, may 2011. doi: 10.1201/b10905. URL https://doi.org/10.1201%2Fb10905.

[6] Simon Duane, A.D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(87)91197-X. URL https://www.sciencedirect.com/science/article/pii/037026938791197X.

[7] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR*, abs/1703.11008, 2017. URL https://arxiv.org/abs/1703.11008.

[8] Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/9a0ee0a9e7a42d2d69b8f86b3a0756b1-Paper.pdf.

[9] Sebastian Farquhar, Lewis Smith, and Yarin Gal. Liberty or depth: Deep Bayesian neural nets do not need complex weight posterior approximations. In *Advances in Neural Information Processing Systems*, volume 33, pages 4346–4357, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/2dfe1946b3003933b7f8ddd71f24dbb1-Paper.pdf.

[10] Andrew Foong, Wessel Bruinsma, David Burt, and Richard Turner. How tight can PAC-Bayes be in the small data regime? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4093–4105. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/214cfbe603b7f9f9bc005d5f53f7a1d3-Paper.pdf.

[11] Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. On the expressiveness of approximate inference in Bayesian neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 2020. ISBN 9781713829546.

[12] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 08834237. URL http://www.jstor.org/stable/2246093.

[13] Benjamin Guedj. A primer on PAC-Bayesian learning, 2019. URL https://arxiv.org/pdf/2110.11216.pdf.

[14] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are Bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/izmailov21a.html.

[15] Galin L. Jones and Qian Qin. Markov chain Monte Carlo in practice. *Annual Review of Statistics and Its Application*, 9(1):557–578, 2022. doi: 10.1146/annurev-statistics-040220-090158. URL https://doi.org/10.1146/annurev-statistics-040220-090158.

[16] John Langford and Rich Caruana. (Not) bounding the true error. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14.

MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/98c7242894844ecd6ec94af67ac8247d-Paper.pdf.

[17] Sanae Lotfi, Marc Anton Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=o8nYuR8ekFm.

[18] Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/618faa1728eb2ef6e3733645273ab145-Paper.pdf.

[19] Andreas Maurer. A note on the PAC Bayesian theorem, 2004. URL https://arxiv.org/pdf/cs/0411099.pdf.

[20] Mahdi Naeini, Pakdaman, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

[21] Maria Perez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 22:227:1–227:40, 2020. URL https://api.semanticscholar.org/CorpusID:220793611.

[22] Konstantinos Pitas. Dissecting non-vacuous generalization bounds based on the mean-field approximation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7739–7749. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/pitas20a.html.

[23] H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184, 01 1960. ISSN 0010-4620. doi: 10.1093/comjnl/3.3.175. URL https://doi.org/10.1093/comjnl/3.3.175.

[24] Jeffrey S. Rosenthal. Simple confidence intervals for MCMC without CLTs. *Electronic Journal of Statistics*, 11(1):211 – 214, 2017. doi: 10.1214/17-EJS1224. URL https://doi.org/10.1214/17-EJS1224.

[25] Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In Steve Hanneke and Lev Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 466–492. PMLR, 15–17 Oct 2017. URL https://proceedings.mlr.press/v76/thiemann17a.html.

[26] Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

# 6 Supplementary Material

## 6.1 Overview of contents

The supplementary material is structured as follows. Section 6.2.1 proves the simple result needed to deduce KL estimation to marginal likelihood estimation. Sections 6.2.2, 6.2.3 and 6.2.4 are centered around thermodynamic integration. We introduce the method, apply it in our setting and justify why the trapezium rule gives an upper bound on our thermodynamic integral. Section 6.2.5 details our high-probability bounds used to control the probability of our PAC-Bayes estimates, including their assumptions. Then, Section 6.3 details technical aspects of the implementation and hyperparameters used. Finally, Sections 6.4 and 6.5 supply further diagnostic and RC estimation results and analysis.

**A note on notation** Throughout the supplementary material, for the sake of generality, we state our results for general $\lambda$ values. However, the reader should keep in mind that we fix $\lambda = 1$ in our experiments.

## 6.2 Details on the method

In this section, we provide further details, derivations, and proofs necessary for our method.

### 6.2.1 Proof of Proposition 3.1

*Proof.* we can write the ELBO in two ways:

$$\mathbb{E}_q \left[ \log \frac{p(z, \mathbf{w})}{q(\mathbf{w}|z)} \right] = \mathbb{E}_q \left[ \log \frac{p(\mathbf{w})p(z|\mathbf{w})}{q(\mathbf{w}|z)} \right] = \mathbb{E}_q \left[ \log \frac{p(z)p(\mathbf{w}|z)}{q(\mathbf{w}|z)} \right] \tag{6}$$

We can expand both terms as:

$$\mathbb{E}_q \left[ \log p(z|\mathbf{w}) \right] + \mathbb{E}_q \left[ \log \frac{p(\mathbf{w})}{q(\mathbf{w}|z)} \right] = \mathbb{E}_q \left[ \log p(z) \right] + \mathbb{E}_q \left[ \log \frac{p(\mathbf{w}|z)}{q(\mathbf{w}|z)} \right], \tag{7}$$

which we can rewrite as

$$\mathbb{E}_q \left[ \log p(z|\mathbf{w}) \right] - \mathbb{E}_q \left[ \log \frac{q(\mathbf{w}|z)}{p(\mathbf{w})} \right] = \mathbb{E}_q \left[ \log p(z) \right] - \mathbb{E}_q \left[ \log \frac{q(\mathbf{w}|z)}{p(\mathbf{w}|z)} \right]. \tag{8}$$

We can now rewrite this as KL divergences and notice that $p(z)$ is independent of $\mathbf{w}$, hence

$$\mathbb{E}_q \left[ \log p(z|\mathbf{w}) \right] - \text{KL}[q(\mathbf{w}|z)||p(\mathbf{w})] = \log p(z) - \text{KL}[q(\mathbf{w}|z)||p(\mathbf{w}|z)]. \tag{9}$$

We can reorder this as:

$$\text{KL}[q(\mathbf{w}|z)||p(\mathbf{w})] = \mathbb{E}_q \left[ \log p(z|\mathbf{w}) \right] - \log p(z) + \text{KL}[q(\mathbf{w}|z)||p(\mathbf{w}|z)]. \tag{10}$$

$\square$

### 6.2.2 Introduction to Thermodynamic Integration

Thermodynamic integration (TI) is a physics-inspired method that allows us to approximate intractable normalizing constants of high dimensional distributions [18]. The main insight is to transform the problem into estimating the difference of two log normalizing constants. Since we are required to estimate $\log(Z)$, this framework suits our purposes.

Consider two probability measures $\Pi_1, \Pi_2 \in \mathcal{M}_1(\mathcal{W})$ with corresponding densities $\pi_1(\mathbf{w}), \pi_2(\mathbf{w})$ and their unnormalized versions

$$\pi_i(\mathbf{w}) = \frac{\tilde{\pi}_i(\mathbf{w})}{W_i}, \qquad W_i = \int \tilde{\pi}(\mathbf{w}) \mathrm{d}\mathbf{w}, \quad i \in \{0, 1\}. \tag{11}$$

To apply TI, we form a geometric path between $\pi_0(\mathbf{w})$ and $\pi_1(\mathbf{w})$ via a scalar parameter $\beta \in [0, 1]$:

$$\pi_\beta(\mathbf{w}) = \frac{\tilde{\pi}_\beta(\mathbf{w})}{W_\beta} = \frac{\tilde{\pi}_1(\mathbf{w})^\beta \tilde{\pi}_0(\mathbf{w})^{1-\beta}}{W_\beta}, \quad W_\beta = \int \tilde{\pi}_\beta(\mathbf{w}) \mathrm{d}\mathbf{w}, \quad \beta \in [0, 1]. \tag{12}$$

The central identity of thermodynamic integration is as follows. The right-hand side of Equation 13 is referred to as the *thermodynamic integral*.

8

**Proposition 6.1.** *Define the potential as $U_\beta(\mathbf{w}) = \log \tilde{\pi}_\beta(\mathbf{w})$ and let $U'_\beta(\mathbf{w}) := \frac{\partial}{\partial \beta} U_\beta(\mathbf{w})$. Then,*

$$\log(W_1) - \log(W_0) = \int_0^1 \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ U'_\beta(\mathbf{w}) \right] d\beta. \tag{13}$$

*Proof.* Please refer to Appendix A in [18]. □

### 6.2.3 Using Thermodynamic Integration

As seen in Section 6.2.2, thermodynamic integration requires defining two probability measures with unnormalized densities, $\tilde{\pi}_i(\mathbf{w})$ and calculates the difference of the log normalizing constants $\log(W_1) - \log(W_0)$. If $\tilde{\pi}_0$ is defined such that $W_0 = 1$, then TI calculates the log normalizing constant of $\tilde{\pi}_1$. We thus define

$$\tilde{\pi}_0(\mathbf{w}) := p(\mathbf{w}), \quad \tilde{\pi}_1(\mathbf{w}) := p(z, \mathbf{w}), \tag{14}$$

thus $W_0 = 1$ and $W_1 = \int p(z, \mathbf{w}) d\mathbf{w} = Z$. In TI, we then define the geometric path for $\beta \in [0, 1]$:

$$\tilde{\pi}_\beta(\mathbf{w}) := p(z, \mathbf{w})^\beta p(\mathbf{w})^{1-\beta}. \tag{15}$$

Then we have that $U_\beta(\mathbf{w}) = \log \tilde{\pi}_\beta(\mathbf{w}) = \beta \log p(z, \mathbf{w}) + (1 - \beta) \log p(\mathbf{w})$ and thus

$$\frac{\partial}{\partial \beta} U_\beta(\mathbf{w}) = \log \frac{p(z, \mathbf{w})}{p(\mathbf{w})} = \log \frac{e^{-n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})} p(\mathbf{w})}{p(\mathbf{w})} = -n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}). \tag{16}$$

Then, the thermodynamic integration formula (Equation 13 yields the following form for the log normalizing constant

$$\log(Z) = \int_0^1 \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ -n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right] d\beta. \tag{17}$$

We will estimate $\mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ -n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right]$ with a Monte Carlo average, using samples $\mathbf{w} \sim \pi_\beta$. To sample from each $\pi_\beta$, we can use HMC again to draw from the log joint

$$\log \tilde{\pi}_\beta(\mathbf{w}) = \beta \log p(z, \mathbf{w}) + (1 - \beta) \log p(\mathbf{w}) \tag{18}$$

$$= -\beta \lambda n \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) + \beta \log p(\mathbf{w}) + (1 - \beta) \log p(\mathbf{w}) \tag{19}$$

$$= -\beta \lambda n \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) + \log p(\mathbf{w}). \tag{20}$$

### 6.2.4 The trapezium rule upper bounds the thermodynamic integral

In this section, we state our results for the *negative* log normalizing constant, $-\log(Z)$ for convenience. A naive strategy for calculating the integrand for $\log(Z)$ is to use a Monte Carlo average for the integral.

$$-\log(Z) = \int_0^1 \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right] d\beta \tag{21}$$

$$\approx \frac{1}{B} \sum_{b=0}^{B-1} \left[ \mathbb{E}_{\mathbf{w} \sim \pi_{\frac{b}{B}}} \left[ n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right] \right] \tag{22}$$

Notice that the right-hand-side in Equation 21 is a left Riemann sum (i.e. the integral is approximated at the left end of each subinterval of $[0, 1]$) on the function $\mathbf{w} \mapsto \mathbb{E}_{\mathbf{w} \sim \pi_{b/B}} \left[ n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right]$, where each subinterval has length $\frac{1}{B}$.

The fact that this quantity upper bounds the KL divergence is established in [18] by showing that the integrand is a decreasing function of $\beta$. Let $g(\beta) = \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ -U'_\beta(\mathbf{w}) \right] = \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right]$, our integrand. We notice that $U'_\beta(\mathbf{w}) = -n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})$ is independent of $\beta$, hence we may abandon the subscript and use notation $U'(\mathbf{w}) = -n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})$). [18] show that

**Proposition 6.2.** $\frac{\partial g(\beta)}{\partial \beta} = -\mathrm{Var}_{\mathbf{w} \sim \pi_\beta} \left[ U'(\mathbf{w}) \right] \le 0.$

9

*Proof.* Please refer to Appendix A in [18]. Note that they define $g(\beta)$ as the *negative* of our $g(\beta)$. $\square$

Hence $\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [-U'(\mathbf{w})]$ is a monotonically decreasing function, and this shows that calculating the left sums upper bounds our integral. In fact, we can show more. Below we show that $g(\beta)$ is convex, hence the trapezium rule can also be used to upper bound this integral.

**Proposition 6.3.** *Let* $g(\beta) = \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [-U'(\mathbf{w})] = \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}) \right]$. *Then we have that* $\frac{\partial^2 g(\beta)}{\partial \beta^2} \geq 0$, *hence* $g(\beta)$ *is convex.*

*Proof.* We make use of the fact that $\frac{\partial}{\partial \beta} \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [-U'(\mathbf{w})] = -\mathrm{Var}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]$, proved in [18] (note that they define $g(\beta)$ to be the negative of our $g(\beta)$, hence the two statements differ by a minus sign). Plugging this in and expanding the variance, we obtain

$$\frac{\partial^2 g(\beta)}{\partial \beta^2} = \frac{\partial^2}{\partial \beta^2} \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [-U'(\mathbf{w})] \tag{23}$$

$$= -\frac{\partial}{\partial \beta} \left[ \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^2 \right] - \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^2 \right] \tag{24}$$

$$= -\frac{\partial}{\partial \beta} \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^2 \right] + 2\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \frac{\partial}{\partial \beta} \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \tag{25}$$

$$= A + B. \tag{26}$$

For $B$ we can use the above, i.e.

$$B = 2\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \, \mathrm{Var}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \tag{27}$$

$$= 2\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \, \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^2 \right] - 2\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^3. \tag{28}$$

For $A$, note that $U'(\mathbf{w})$ is *independent of* $\beta$, hence we can directly plug into equation (30) in [18], noting that we have an extra $U'(\mathbf{w})$ term and we need the negative of their expression:

$$A = \int (U'(\mathbf{w}))^2 \pi_\beta(\mathbf{w}) \mathbb{E}_{\mathbf{w} \sim \pi_\beta(\mathbf{w})} [U'(\mathbf{w})] \, d\mathbf{w} - \int (U'(\mathbf{w}))^3 \pi_\beta(\mathbf{w}) d\mathbf{w} \tag{29}$$

$$= \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \, \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^2 \right] - \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^3 \right]. \tag{30}$$

Hence $A + B$ is

$$A + B = -2\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^3 - \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^3 \right] + 3\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \, \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^2 \right]. \tag{31}$$

Since $U'(\mathbf{w}) = -n\lambda \hat{L}_S(\mathbf{w})$, it is negative. The function $x \mapsto x^3$ is concave on $(-\infty, 0)$, while the function $x \mapsto x^2$ is convex. Hence Jensen's inequality gives:

$$\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^3 \geq \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^3 \right] \tag{32}$$

$$\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^2 \leq \mathbb{E}_{\mathbf{w} \sim \pi_\beta} \left[ (U'(\mathbf{w}))^2 \right]. \tag{33}$$

Thus,

$$A + B = \frac{\partial^2 g(\beta)}{\partial \beta^2} \geq 3 \left[ -\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^3 + \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})] \, \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^2 \right] \tag{34}$$

$$= 3 \left[ -\mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^3 + \mathbb{E}_{\mathbf{w} \sim \pi_\beta} [U'(\mathbf{w})]^3 \right] = 0, \tag{35}$$

thus $g(\beta)$ is convex. $\square$

### 6.2.5 High-probability bounds

**The kl inverse bound** To ensure approximate independence, the bounds were calculated on a thinned version of the Gibbs samples, which ensured that the ESS of the thinned sample is close to the remaining sample size. This resulted in retaining 1000-3000 samples out of 19800, depending on the dataset. For comparison, MFVI was also evaluated using the same amount of (exact) Gaussian samples.

**The asymptotic bound** The Markov chain Central Limit Theorem (MCCLT) provides a confidence interval, but the conditions of this theorem are very strong. Fortunately, [24] provide a similar, but weaker confidence interval that does not require the MCCLT. The assumptions on our estimator $\frac{1}{m}\sum_{i=1}^{m}\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w}_i)$, $\mathbf{w}_i \sim Q$ are as follows.

1. $O\left(\frac{1}{n}\right)$ variance of the estimator

2. Bias of order smaller than $O\left(\frac{1}{\sqrt{n}}\right)$

3. An estimator $\hat{\sigma}_m$ for the *asymptotic variance* $\sigma := \lim_{m\to\infty} m\mathrm{Var}_{\mathbf{w}\sim Q}\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right]$ that converges in probability.

Assumptions 1. and 2. are typical in MCMC applications [24]. For assumption 3, we estimate the asymptotic variance using the fact that the MCMC standard error converges to $\frac{\sigma}{\sqrt{m}}$, following one of the suggestions of [24]. The (one-sided) version on the confidence interval in [24] has form

$$I_{m,\epsilon} = \left[0, \frac{1}{m}\sum_{i=1}^{m}\hat{L}_S(\mathbf{w}_i) + m^{-\frac{1}{2}}\hat{\sigma}_m\left(2\alpha\right)^{-\frac{1}{2}}\left(1+\epsilon\right)\right) \quad \text{with prob. at least } 1-\alpha, \qquad (36)$$

where $m$ is the number of samples, $\hat{\sigma}_m$ is an estimate of the asymptotic variance, and $\epsilon$ appears in the proof when formalizing Assumption 3. We take $\epsilon = 0.01$. Fair comparison to MFVI demands that we use the classical CLT to obtain an asymptotic confidence interval in this case. This is valid as both the transformed cross-entropy and the $0-1$ losses are bounded. The CLT confidence interval has form $I_\alpha = \left[0, \frac{1}{m}\sum_{i=1}^{m}\hat{L}_S(\mathbf{w}_i) + q_\alpha\frac{\hat{\sigma}_m}{\sqrt{n}}\right)$, where $q_\alpha$ is the appropriate quantile of the standard Gaussian distribution[3].

**The naive bound** Given the difficulties in verifying assumptions for MCMC samples, the reader may wonder if we can give, perhaps looser, estimates for $C(Q_\lambda^*, \lambda)$ with milder assumptions. Let $Q$ denote the underlying distribution of our HMC samples, which may not be Gibbs posterior. The following simple proposition can be used to give a (loose) bound for $\mathrm{KL}(Q||P)$ only requiring that $Q$ is not much further from the Gibbs posterior than the mean-field approximation. We expect that $Q$ is much closer to the Gibbs posterior than MFVI, hence this is a very mild assumption.

**Proposition 6.4.** *Let $P$ be the prior, and let $Q_\lambda^*$ denote the corresponding (Gibbs) posterior $Q_\lambda^* \propto e^{-n\lambda\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})}p(\mathbf{w})$. Suppose that we are able to simulate from a distribution Q. Let G be another distribution (in our case, a Gaussian) such that $\mathrm{KL}(Q||Q_\lambda^*) \leq \mathrm{KL}(G||Q_\lambda^*) + n\mathbb{E}_Q\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right]$. Then,*

$$\mathrm{KL}(Q||P) \leq n\lambda\mathbb{E}_G\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right] + \mathrm{KL}(G||P). \qquad (37)$$

*Proof.* We defer the proof until the end of this section. $\qquad\square$

Let us denote our bound objectives as $C(Q, \lambda)$. We can use this estimate to obtain an upper bound on $C(Q, \lambda)$, which we denote by $C^{\mathrm{UB}}(Q, \lambda)$. For this, we need an estimate for $\widetilde{L}_S^{\mathrm{ce}}(Q)$, for which we use the kl inverse bound (Theorem 2.3). It can be shown easily that we have $C^{\mathrm{UB}}(Q, \lambda) \geq C(G, \lambda)$ if we use the $\lambda$ bound and the cross-entropy loss. However, we are most interested in risk certificates in terms of accuracy. For the $0-1$ loss, our upper bound on the $\lambda$ bound becomes

$$L^{0\text{-}1}(Q) \leq \frac{\hat{L}_S^{0\text{-}1}(Q)}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(G||P) + \log(2\sqrt{n}/\delta)}{n\lambda(1-\frac{\lambda}{2})} + \frac{\widetilde{L}_S^{\mathrm{ce}}(G)}{1-\frac{\lambda}{2}} =: C^{\mathrm{UB}}(Q, \lambda). \qquad (38)$$

This can result in $C^{\mathrm{UB}}(Q, \lambda) \leq C(G, \lambda)$ if $Q$ has lower $0-1$ loss than $G$. The difference between this and the true $\lambda$ bound on $L^{0\text{-}1}(Q)$ is precisely $\mathrm{KL}(G||Q_\lambda^*) - \mathrm{KL}(Q||Q_\lambda^*) + \widetilde{L}_S^{\mathrm{ce}}(Q)$. We will use this bound for a sanity check on our results. If our computed value for $C(Q, \lambda)$ is much higher than this bound, then we likely overestimated the true value, even if our bound is smaller than the one with MFVI. If, on the contrary, our estimate is much lower than this upper bound, that means that

---

[3]Although the population variance is unknown, we decided to use this instead of the Student-t distribution to ensure comparability to $I_{m,\epsilon}$, and because we have a large sample size, $m = 20,000$ across the four chains.

either $\mathrm{KL}(G||Q_\lambda^*) - \mathrm{KL}(Q||Q_\lambda^*) + \widetilde{L}_S^{\mathrm{ce}}(Q)$ is high or our chains did not even manage to achieve $\mathrm{KL}(Q||Q_\lambda^*) \leq \mathrm{KL}(G||Q_\lambda^*) + \widetilde{L}_S^{\mathrm{ce}}(Q)$.

*Proof of Proposition 6.4* We use Proposition 3.1 with $P$ having density $p(\mathbf{w})$, $G$ having density $q(\mathbf{w}|z)$ and $Q_\lambda^*$ having density $p(\mathbf{w}|z)$ and $Z = \mathbb{E}_P\left[e^{-n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})}\right]$ being the marginal likelihood. We get

$$\mathrm{KL}(G||P) = -n\lambda \mathbb{E}_G\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right] - \log(Z) + \mathrm{KL}(G||Q_\lambda^*). \tag{39}$$

Reordering this, we obtain an estimate for $-\log(Z)$:

$$-\log(Z) = \mathrm{KL}(G||P) - \mathrm{KL}(G||Q_\lambda^*) + n\lambda \mathbb{E}_G\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right]. \tag{40}$$

This estimate can be used to calculate $\mathrm{KL}(Q||P)$, invoking Proposition 3.1 again.

$$\begin{aligned}
\mathrm{KL}(Q||P) = {} & n\lambda \mathbb{E}_G\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right] + \mathrm{KL}(G||P) \\
& + \left(-n\lambda \mathbb{E}_Q\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right] + \mathrm{KL}(Q||Q_\lambda^*) - \mathrm{KL}(G||Q_\lambda^*)\right)
\end{aligned} \tag{41}$$

Since we assumed that $\mathrm{KL}(Q||Q_\lambda^*) \leq \mathrm{KL}(G||Q_\lambda^*) + n\lambda \mathbb{E}_Q\left[\widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})\right]$, the last term is negative. By ignoring it, we thus obtain an upper bound on $\mathrm{KL}(Q||P)$.

We note that for $G \equiv P$, Equation 39 becomes Jensen's inequality for the function $-\log$, i.e.,

$$-\log \mathbb{E}_G\left[e^{-n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})}\right] \leq -\mathbb{E}_G\left[\log\left(e^{-n\lambda \widetilde{L}_S^{\mathrm{ce}}(\mathbf{w})}\right)\right]. \tag{42}$$

## 6.3 Implementation details

### 6.3.1 Transforming the cross-entropy loss

PAC-Bayes bounds assume a loss function bounded in $[0, 1]$. However, the cross-entropy loss is unbounded from above. Hence we transform $l^{\mathrm{ce}}(\mathbf{w}, y) = \sum_{i=1}^k y_i \log(p_i)$ as follows.

$$\widetilde{l^{\mathrm{ce}}}(\mathbf{w}, y) := \frac{-\sum_{i=1}^k y_i \log(\max(p_i, p_{\min}))}{-\log(p_{\min})}, \tag{43}$$

with some $p_{\min} > 0$, which now falls into $[0, 1]$. We will take $p_{\min} = 10^{-4}$. The corresponding risk functionals will be denoted by $\widetilde{L}^{\mathrm{ce}}(Q)$ and $\widetilde{L}_S^{\mathrm{ce}}(Q)$.

### 6.3.2 Estimating the kl bound

After arriving at a posterior distribution $Q$ either by sampling from the Gibbs posterior or MFVI, we wish to compute a risk certificate on the *error* (0-1 loss) $L^{0-1}(Q)$ of the stochastic predictor given by $Q$. The transformed cross-entropy loss is only used for sampling. To compute the risk certificate, we use the kl bound (Equation 1) since it is the tightest. To invert the Bernoulli KL, we define

$$\mathrm{kl}^{-1}(x, b) := \sup\{y \in [x, 1] : \mathrm{kl}(x||y) \leq b\}. \tag{44}$$

This can be seen as a proper definition of the inverse Bernoulli KL. Our RC is then

$$L^{0-1}(Q) \leq \mathrm{kl}^{-1}\left(\hat{L}_S^{0-1}(Q), \frac{\mathrm{KL}(Q||P) + \log(\frac{2\sqrt{n}}{\delta})}{n}\right). \tag{45}$$

### 6.3.3 Architectures

We use simple Multi-Layer Perceptron (MLP) architectures in our sampling experiments. Due to the scalability limits of marginal likelihood estimation, the number of parameters are kept small. The architectures along with the corresponding number of parameters are given in Table 2.

### 6.3.4 Hyperparameters

**Prior hyperparameters** We initialize the prior means randomly from a truncated Gaussian distribution of mean 0 and separate variance for each layer, given by $\frac{1}{\sqrt{n_{\mathrm{in}}}}$, where $n_{\mathrm{in}}$ is the input dimensionality of the layer. The constants of the truncation are $\pm 2$ standard deviations. The prior covariance is set to $0.03 \cdot \mathrm{I}$, matching [21]. We use the cross-entropy loss as our generalized likelihood function, in the transformed version (Equation 43), with $p_{\min} = 10^{-4}$. This ensures that our loss is bounded.

Table 2: Our neural network architectures.

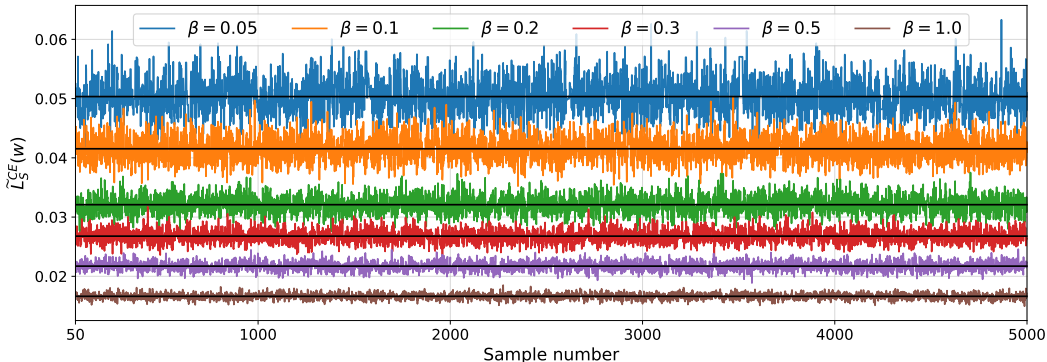| Dataset | Architecture | | | | # params |
|---|---|---|---|---|---|
| Binary MNIST | 784 | 20 | - | 2 | 15742 |
| $14 \times 14$ MNIST | 196 | 128 | 128 | 10 | 43018 |
| MNIST | 784 | 128 | 128 | 10 | 118282 |



Figure 3: HMC sample traces for a single chain without burn-in (50), for various $\beta$ values. The black lines show the mean of each chain. The dataset is Binary MNIST - Half.

**MFVI hyperparameters**   As optimizer, we used SGD with learning rate $0.005$ and momentum $0.95$ and train for 150 epochs. In the case of MFVI, the $\lambda$ bound objective can be easily optimized in $\lambda$ too. To match our proposed method, we fix $\lambda = 1$ for MFVI too.

We ran experiments with optimizing for $\lambda$ too, but the best RCs were achieved when $\lambda$ changed very little during the training (at most $0.02$ across our experiments). For these experiments, we followed [21] in optimizing $\lambda$ separately using SGD with the same learning rate and momentum used for MFVI. We have run initial experiments with higher learning rates for $\lambda$, but these resulted in looser risk certificates using the kl bound.

We choose our risk certificates to hold with probability $1 - \delta$ with $\delta = 0.05$. We use $\delta_1 = 0.025$ to upper bound $\frac{1}{m} \sum_{i=1}^{m} L_S(\mathbf{w}_i)$, $\mathbf{w}_i \sim Q$ and then $\delta_2 = 0.025$ to compute bounds. This ensures a risk certificate that holds with probability $0.95$, via a union bound argument.

**HMC hyperparameters**   We use full-batch HMC to sample from Gibbs posteriors of form $\propto e^{-\beta \widetilde{L}_S^{\text{ce}}(\mathbf{w})} p(\mathbf{w})$ for $\beta \in (0, 1)$ using the JAX HMC implementation of [14]. We run HMC with Metropolis-Hastings correction. We use constant step-sizes for the discretization of Hamiltonian dynamics, which we calibrate individually by testing the values $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30\} \times 10^{-3}$ for each dataset and $\beta$. The step-sizes were chosen from this set by running HMC for around 100 sampling iterations for each step size, targeting the ideal acceptance probability for HMC given by $0.65$ [3]. Longer trial runs were not possible due to computational constraints. Lower step sizes were found suitable for higher $\beta$ values. In every experiment, we use a trajectory length of $1.5$. The authors of [14] recommend a trajectory length of at least $\frac{\pi \sigma_{\text{prior}}}{2}$, which is approximately $0.27$ in our case, but we choose higher than this in order to allow for more leapfrog steps. The chosen step-sizes resulted in leapfrog steps between 50 and 750. The difference in leapfrog steps is motivated by the fact that for lower $\beta$ values, the generalized posterior is more similar to the prior and hence is easier to sample from. Each chain was run for 5000 iterations including a burn-in of 50 iterations. A low burn-in was chosen since we observed that the chains converge very quickly in function space ($\widetilde{L}_S^{\text{ce}}(\mathbf{w})$). To support this choice, Figure 3 shows burned-in chains for the Binary MNIST - Half dataset for various values of $\beta$. To be able to test the convergence of the chains, we run four chains started from independent seeds for each $\beta$.

### 6.4 Diagnostics

This section contains the full description and results of our diagnostic analysis on the HMC samples.

13

### 6.4.1 Diagnostic measures

Let $(\mathbf{X}^{(t)})_{t=1}^{\infty}$ be a Markov chain on state space $\mathcal{A}$, whose initial distribution equals its stationary distribution, hence $X_1, X_2, ...$ are identically distributed. Consider a square-integrable function $f : \mathcal{A} \to \mathbb{R}$, whose expected value $\mathbb{E}_{P_{X_1}}[f(X_1)]$ we wish to estimate. Note that square integrability impies $\text{Var}[f(X_1)] \leq \infty$.

**Effective Sample Size**    First, we define the effective sample size (ESS), which quantifies the loss of information caused by correlation.

**Definition 6.1.** *(Effective Sample Size) The Effective sample size of* $(\mathbf{X}^{(t)})_{t=1}^{N}$ *is given by*

$$\text{ESS}[f] = \frac{N}{1 + 2\sum_{\tau=1}^{\infty} \rho(\tau)}, \tag{46}$$

*where* $\rho(\tau) = \text{Cor}(f(X_1), f(X_{1+\tau}))$ *denotes the autocorrelation at lag* $\tau$.

The ESS is typically lower than the number of samples, indicating the presence of positive correlations. In MCMC, where states are typically positively correlated, an ESS equaling the number of samples signals uncorrelated samples.

### MCMC standard error

**Definition 6.2.** *(MCMC standard error) The MCMC standard error of* $(\mathbf{X}^{(t)})_{t=1}^{N}$ *is*

$$\text{MCMC\_SE}[f] = \sqrt{\frac{\text{Var}[f(X_1)]}{\text{ESS}[f]}}. \tag{47}$$

The MCMC SE measures the concentration of a sample mean around the true mean, in the sense of the Markov chain Central Limit theorem (MCCLT). It requires the above conditions on the Markov chain, namely (i) stationary chain and (ii) square-integrable $f$ and states that, for large $N$, approximately

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i) \sim \mathcal{N}(\mathbb{E}_{P_{X_1}}[f(X_1)], \text{MCMC\_SE}[f]). \tag{48}$$

In practice, we don't have access to $\rho(\tau)$ and $\text{Var}[f(X_1)]$, and estimate them from a finite set of samples.

$\hat{R}$ **statistic**    Since ESS requires stationarity to be well-defined, we will also use the $\hat{R}$ (known also as potential scale reduction) to check convergence by comparing multiple independent Markov chains [12]. $\hat{R}$ measures how much the variance of the means between multiple chains exceeds that of identically distributed chains. We have that $\hat{R} \geq 1$, where $\hat{R} = 1$ means perfect convergence. For more discussion on $\hat{R}$, please refer to [12], and for general discussion on MCMC and convergence diagnostics, see [15] and [5].

**Expected Calibration Error**    Following [14], we supply expected calibration error (ECE) estimates averaged over all estimates for a given $\beta$. ECE measures model calibration by quantifying how well a model's output pseudo-probabilities match the true (observed) probabilities [20].

### 6.4.2 Diagnostic results

Figures 4 5, 6 7, 8, 9 show diagnostic results for all datasets, respectively. The target acceptance probability of $0.65$ was not always achieved on average. On Binary MNIST, acceptance probabilities were higher, around $70 - 80\%$. In these cases, increasing the step-size resulted in unstable behaviour, with an acceptance probability close to $0$. We followed the usual recommendation for these cases, which is to keep the step-size slightly lower than the highest stable value [4]. On $14 \times 14$ MNIST, acceptance probabilities were roughly as desired.

However, the MNIST experiments resulted in some acceptance probabilities below $0.4$. In these cases, the ideal step-size is likely smaller. However, for MNIST, decreasing the step-size would have meant the largest increase in running time. Furthermore, running short chains with smaller step-sizes
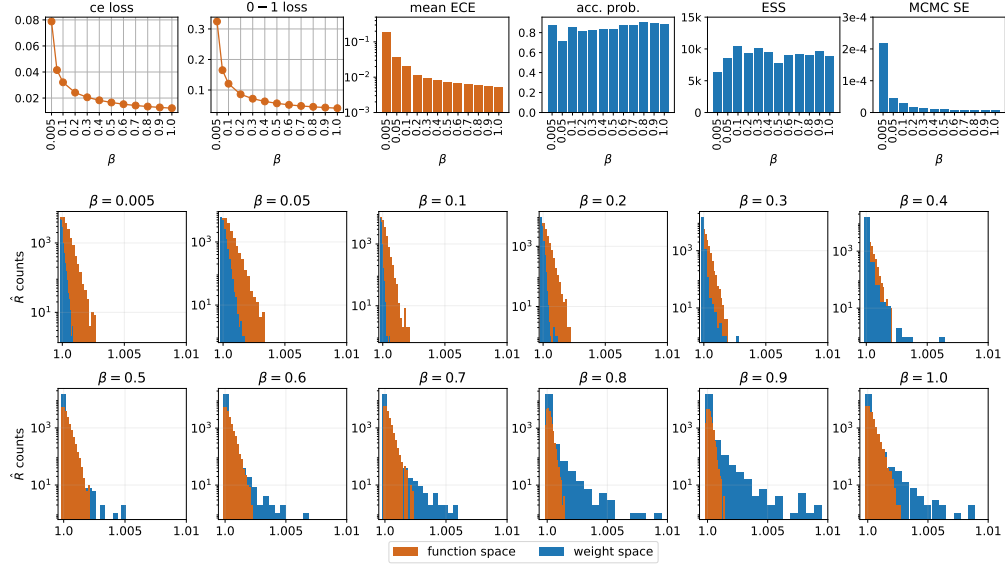
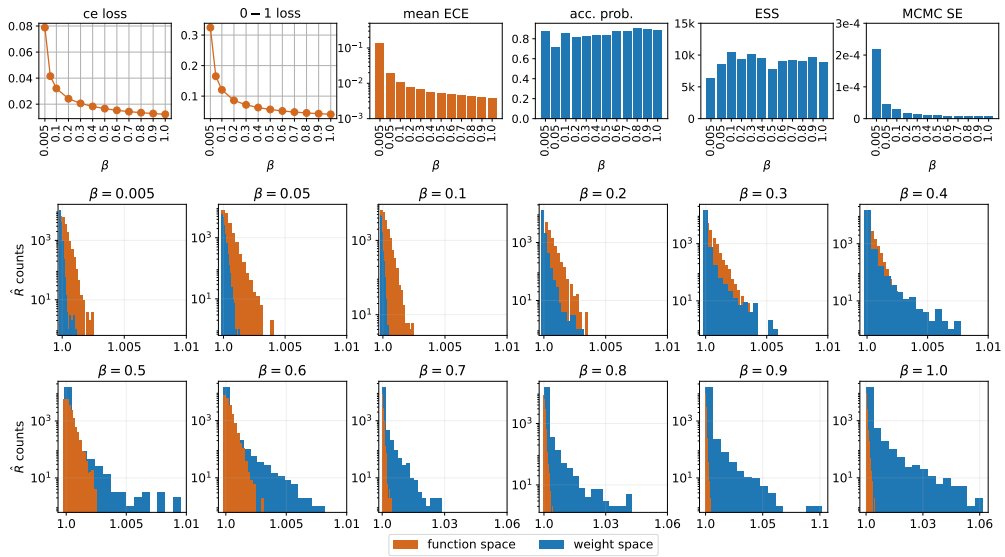Figure 4: HMC diagnostics for Binary MNIST - half, over 4 chains of length 5000, for each $\beta$.



Figure 5: HMC diagnostics for Binary MNIST - full, over 4 chains of length 5000, for each $\beta$.

showed very similar values for the train and test losses. Hence a decision was made to use larger step-sizes to be able to produce a longer chain.

The ESS values are relatively high, almost always retaining at least 5000 out of the 19800 samples (burn-in removed) across the four chains. The MCMC SE values are also low. The sudden decrease as $\beta$ is increased is explained by the fact that the true variance of the Gibbs posterior decreases as $\beta$ increases since the posterior becomes more concentrated around the minima of the loss landscape. We observe low ECE values implying that our samples correspond to well-calibrated models. All $\hat{R}$ statistics are very close to 1.0, showing that our chains are approximately stationary. This finding motivates our use of the kl inversion bound (Theorem 2.3) which required an identically distributed sample.
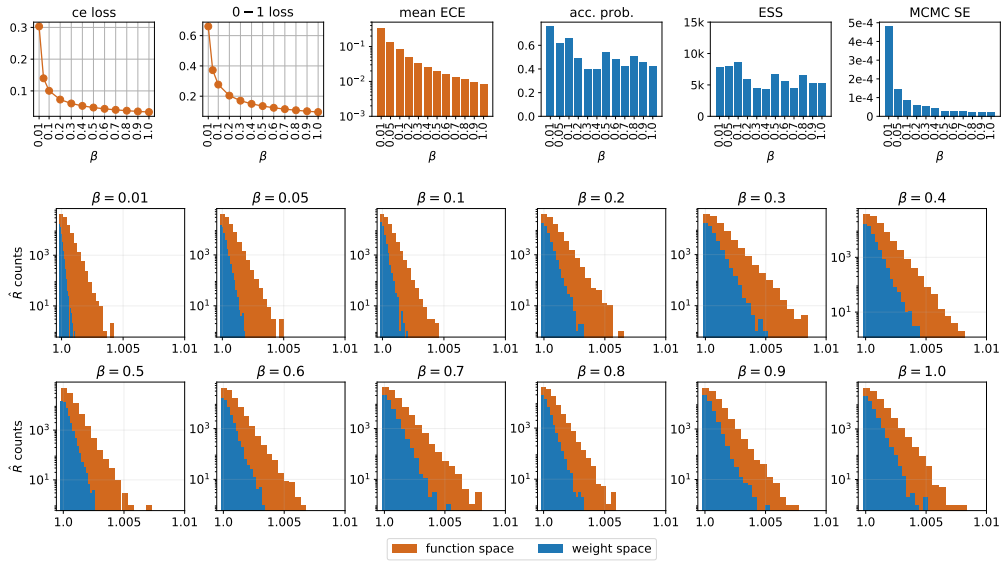
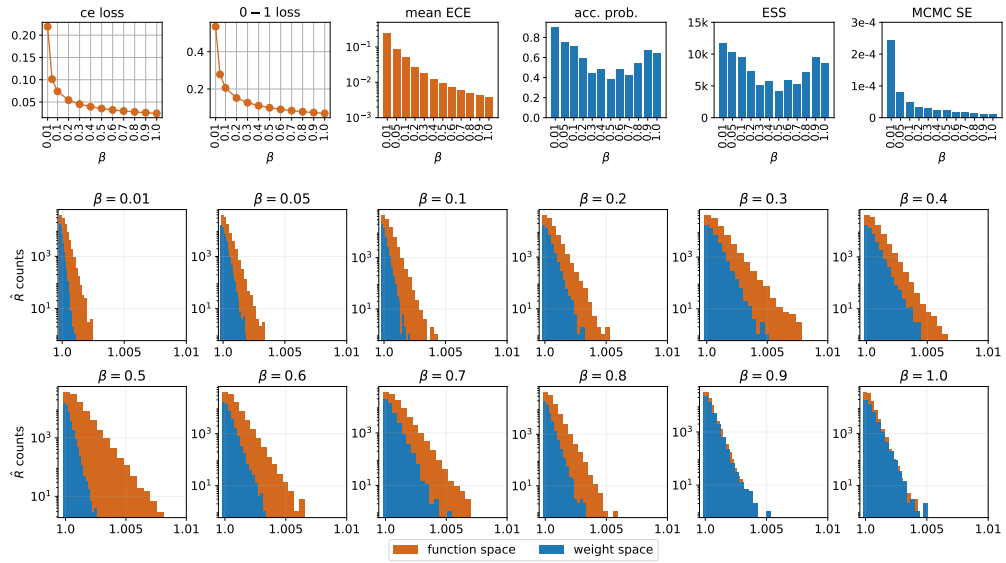Figure 6: HMC diagnostics for $14 \times 14$ MNIST - half, over 4 chains of length 5000, for each $\beta$.



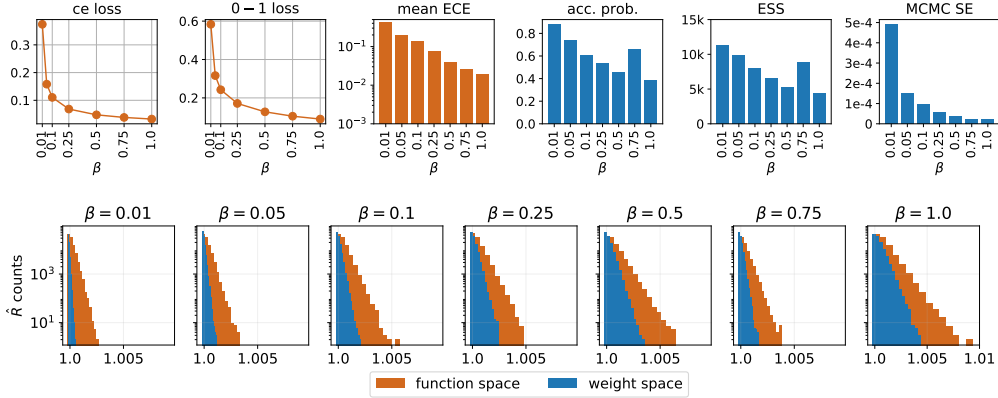Figure 7: HMC diagnostics for $14 \times 14$ MNIST - full, over 4 chains of length 5000, for each $\beta$.

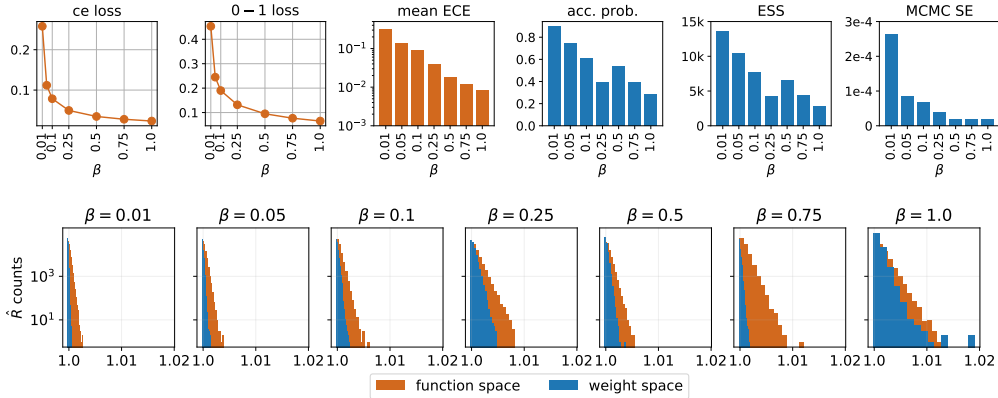Figure 8: HMC diagnostics for MNIST - half, over 4 chains of length 5000, for each value of $\beta$.



Figure 9: HMC diagnostics for MNIST - full, over 4 chains of length 5000, for each value of $\beta$.

## 6.5 Further results and analysis

Table 3 shows RC estimation results for the (bounded) cross-entropy loss. We first observe that there are no bound violations, i.e. the test loss/error is always smaller than the corresponding RC. Further, the RCs for MFVI and the Gibbs posterior have the same magnitude. This suggests that there are no apparent problems with our estimates. Interestingly, the computed RCs for $\widetilde{L}_S^{\text{ce}}(Q)$ (denoted as $\widetilde{\text{ce}}$ in the tables) are almost always better for MFVI than for the approximate Gibbs posterior samples, despite the Gibbs posterior being the minimizer of the $\lambda$ bound. However, the RC-s on the $0-1$ losses (Table 1 above) show considerable improvement with respect to MFVI, especially for Binary MNIST. This discrepancy in cross-entropy and $0-1$ losses may indicate that we have considerably overestimated the KL divergence term. Since our KL divergences are small compared to the $0-1$ losses, they affect the $0-1$ RCs less. We are mainly interested in $0-1$ RCs, hence this is not a problem for us.

Out of the three high-probability bounds compared, the asymptotic bound gives the lowest risk certificates, which is explained by the fact that (i) more samples were used, and (ii) that this interval only guarantees a probability of $0.95$ as the number of samples tends to infinity.

The gap between the $0-1$ RCs for the Gibbs posterior and MFVI increases as the amount of train data is increased. MFVI gives similar RCs in both cases while Gibbs RCs improve by $3-7\%$ when the amount of data is increased. This is reasonable since Gibbs posteriors concentrate more around minima of the loss as $n$ is increased. However, our results show that MFVI is not able to utilize the extra data very well.

17

Table 3: Training and test metrics and RC estimates using our three bounds, in terms of the cross-entropy loss. **Bold** numbers indicate the tighter certificate out of the Gibbs and MFVI ones for the same dataset.

| Setup | | | Train/test stats | | | ce RC with kl bound | | ce RC with $\lambda$ bound | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Dataset | Model | Train $\tilde{c}e$ | Test $\tilde{c}e$ | KL/n | kl inverse | asympt | kl inverse | asympt |
| MFVI | Binary Half | 1L | 0.0256 | 0.0281 | 0.0106 | **0.0703** | 0.0581 | 0.0934 | 0.0756 |
| Gibbs p. | Binary Half | 1L | 0.0166 | 0.0166 | 0.0205 | 0.0706 | **0.0511** | **0.0915** | **0.0663** |
| MFVI | Binary | 1L | 0.0270 | 0.0272 | 0.0105 | **0.0707** | 0.0580 | 0.0941 | 0.0756 |
| Gibbs p. | Binary | 1L | 0.0122 | 0.0125 | 0.0195 | 0.0598 | **0.0315** | **0.0777** | **0.0402** |
| MFVI | $14 \times 14$ Half | 2L | 0.0481 | 0.0481 | 0.0148 | **0.1238** | 0.0947 | **0.1710** | 0.1263 |
| Gibbs p. | $14 \times 14$ Half | 2L | 0.0335 | 0.0361 | 0.0477 | 0.1410 | **0.0946** | 0.1888 | **0.1237** |
| MFVI | $14 \times 14$ | 2L | 0.0463 | 0.0460 | 0.0140 | 0.1196 | 0.0906 | 0.1631 | 0.1194 |
| Gibbs p. | $14 \times 14$ | 2L | 0.0246 | 0.0259 | 0.0381 | **0.1118** | **0.0702** | **0.1484** | **0.0910** |
| MFVI | MNIST Half | 2L | 0.0430 | 0.0437 | 0.0199 | **0.1176** | 0.0966 | **0.1570** | 0.1263 |
| Gibbs p. | MNIST Half | 2L | 0.0324 | 0.0356 | 0.0428 | 0.1347 | **0.0935** | 0.1792 | **0.1223** |
| MFVI | MNIST | 2L | 0.0423 | 0.0419 | 0.0196 | **0.1010** | 0.0947 | **0.1317** | 0.1226 |
| Gibbs p. | MNIST | 2L | 0.0233 | 0.0253 | 0.0334 | 0.1065 | **0.0673** | 0.1401 | **0.0872** |

When increasing model size, the gap between the $0 - 1$ RCs decreases as we move from Binary MNIST to $14 \times 14$ MNIST and remains roughly the same as we move to MNIST. This supports the hypothesis of variational approximations become more accurate as model depth is increased.