Estimating optimal PAC-Bayes bounds with Hamiltonian Monte Carlo

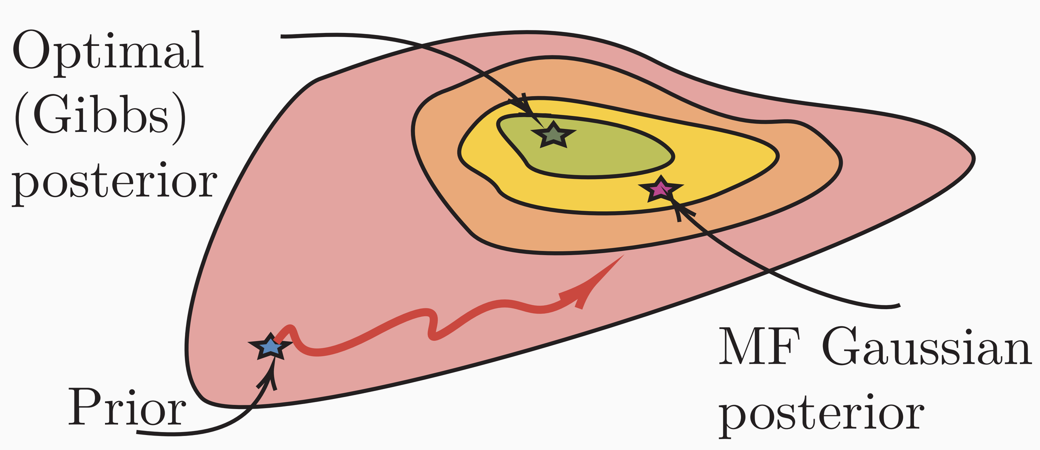
Szilvia Ujváry[1] Gergely Flamich[1] Vincent Fortuin[2] José Miguel Hernández Lobato[1]

[1] *University of Cambridge* [2] *Hemholtz AI, TU Munich*

Mathematics of Modern Machine Learning Workshop @ NeurIPS 2023

Introduction

How far from optimality are data-independent PAC-bounds computed using diagonal covariance posteriors?



- We estimate PAC-Bayes bounds at their **optimal posterior** instead of a MF Gaussian approximation
- Leads to **tighter bounds**
- Shows the **need for better posterior approximations**

Glossary

our task	supervised classification with NNs
MF Gaussian	A Gaussian with diagonal covariance
MFVI	variational inference with MF Gaussians
P	prior distribution on model weights
Q	a (posterior) distribution on the weights
$L(Q)$	expected risk of randomized predictor Q
$\hat{L}_S(Q)$	empirical risk on i.i.d. data sample S
risk certificate	a high-confidence upper bound on $L(Q)$

The bounds

For fixed prior P, for any Q, with probability at least $1-\delta$

$$\text{kl bound [2]: } \mathrm{kl}(\hat{L}_S(Q)||L(Q)) \le \frac{\mathrm{KL}(Q||P) + \log(\frac{2\sqrt{n}}{\delta})}{n}$$

$$\text{linear bound [3]} : L(Q) \le \frac{\hat{L}_S(Q)}{0.5} + \frac{\mathrm{KL}(Q||P) + \log(\frac{2\sqrt{n}}{\delta})}{0.5n}$$

We sample from Q^* (with density q^*) minimizing the linear bound and compute a risk certificate with the kl bound.

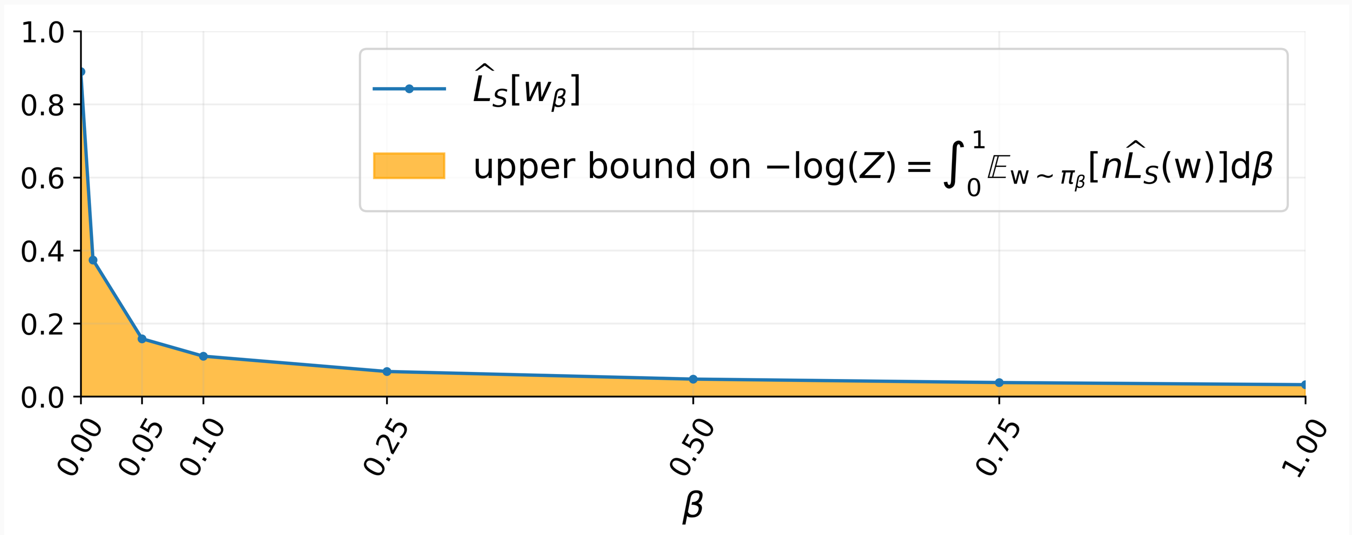
$$q^*(\mathbf{w}) \propto e^{-n\hat{L}_S(\mathbf{w})}p(\mathbf{w})$$

Method part II - KL estimation

We can reduce this problem to estimating the **log marginal likelihood** $\log(Z) = \log \mathbb{E}_{\mathbf{w}\sim P}[e^{-n\hat{L}_S(\mathbf{w})}]$
We compute the **thermodynamic integral** [1]...

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

$$\text{where } \pi_\beta \propto e^{-\beta\hat{L}_S(\mathbf{w})}p(\mathbf{w})$$



...by approximating it with the **trapezium rule**, which we prove to give an upper bound on $-\log(Z)$.

Some results & takeaways

Setup		Train/test stats			0-1 RC with kl bound		
Method	Dataset	Train 0-1	Test 0-1	KL/n	kl inverse	asympt	naive
MFVI	Binary	0.0960	0.0928	0.0105	0.1640	0.1452	0.1640
Gibbs p.	Binary	0.0404	0.0415	0.0195	**0.1080**	**0.0702**	**0.1184**
MFVI	14 × 14	0.1389	0.1313	0.0140	0.2379	0.1991	0.2379
Gibbs p.	14 × 14	0.0695	0.0723	0.0381	**0.1855**	**0.1335**	**0.1920**
MFVI	MNIST	0.1236	0.1200	0.0196	0.2070	0.1987	0.2070
Gibbs p.	MNIST	0.0653	0.0691	0.0334	**0.1759**	**0.1269**	**0.1880**

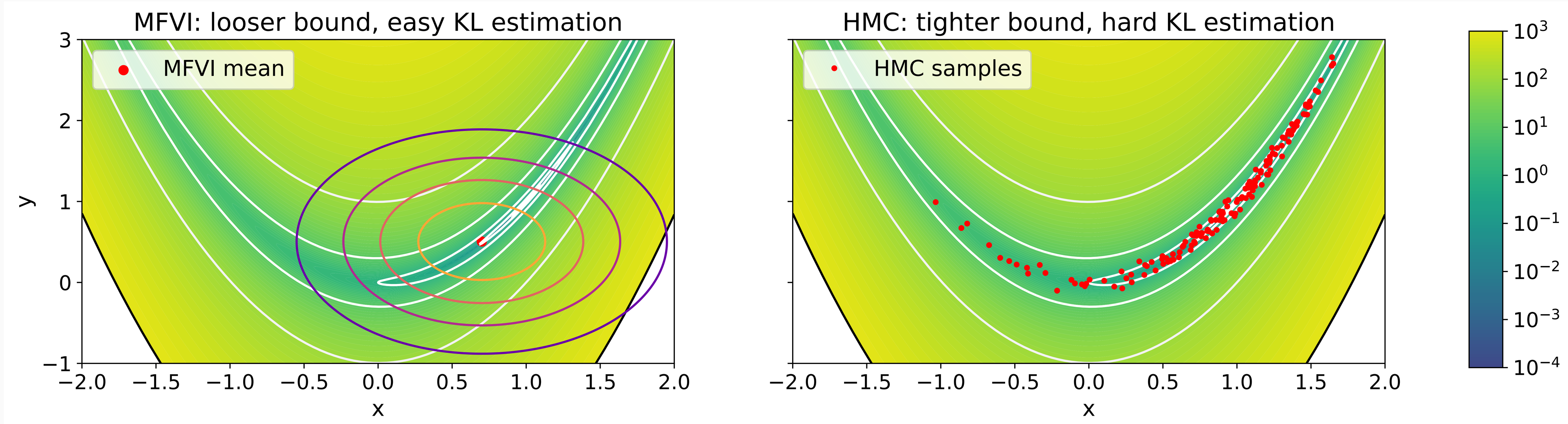
- Reasonable estimates, e.g. no bound violations
- Data-independent bounds can be tightened
- Improvement over MFVI is largest for small models

References

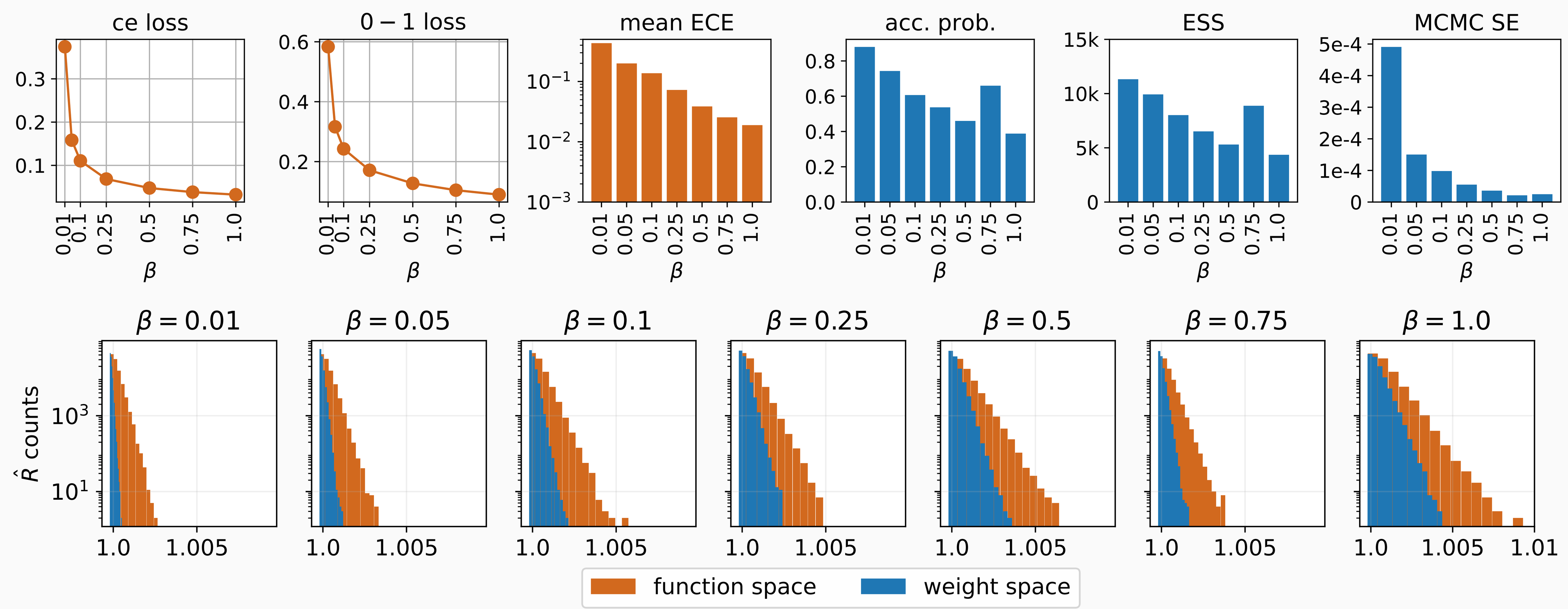
[1] Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*, 2019.
[2] Andreas Maurer. A note on the PAC Bayesian theorem, 2004.
[3] Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, 2017.

Method part I - Sampling from Q^* with HMC

- Why HMC? Can approximate complicated posteriors much better than MF Gaussians
- What's the trade-off? It now becomes harder to estimate the bound



- Does it actually work? → our results are backed up by running extensive diagnostics



Method part III - Ensuring a high-probability bound

We wish to produce a statement such as

$$L(\widehat{Q^*}) \le \{\text{our estimate}\} \text{ with prob. at least } 1-\delta$$

For this we need concentration inequalities on our HMC estimates. It's hard to check convergence assumptions in MCMC, so we give 3 options

1. An i.i.d. concentration inequality on **thinned** samples
2. An asymptotic confidence interval which requires "good estimators"
3. A loose bound that only needs $\mathbf{KL}(\widehat{\mathbf{Q}^*}||\mathbf{Q}^*) < \mathbf{KL}(\mathbf{G}||\mathbf{Q}^*)$ for a baseline MF Gaussian

Experiment details