SAMPLING-BASED INFERENCE FOR LARGE LINEAR MODELS WITH APPLICATION TO LINEARISED LAPLACE

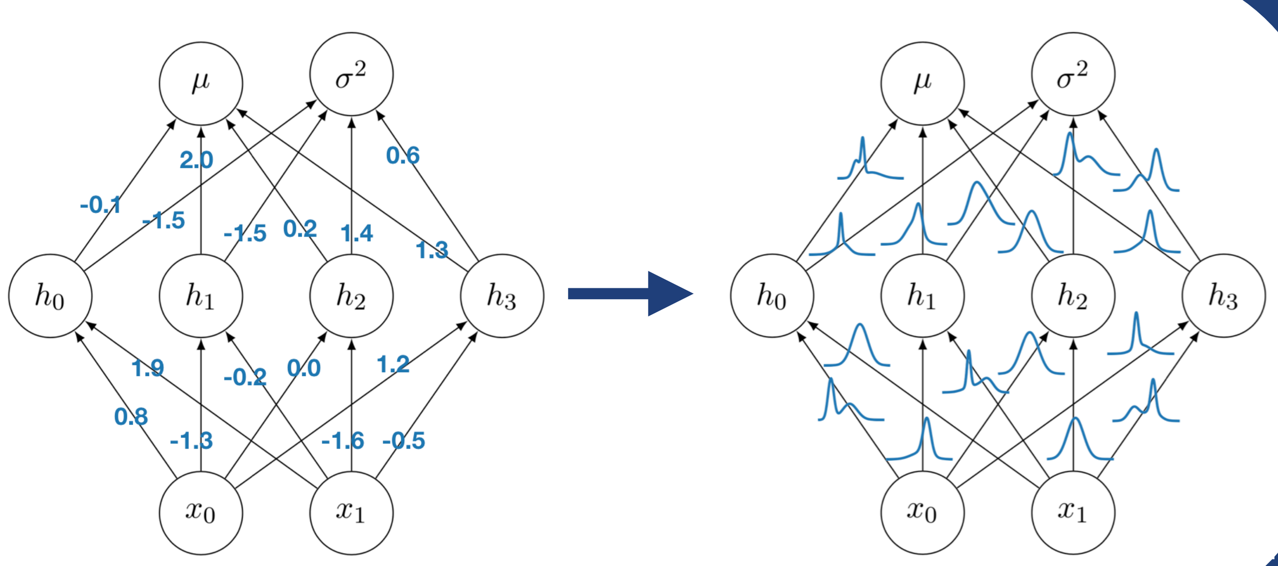
Javier Antorán*, Shreyas Padhy*, Riccardo Barbano, Eric Nalisnick, Dave Janz, Miguel Hernández-Lobato

We draw posterior samples from linear models with millions of parameters and millions of observations and apply it to obtain uncertainty estimates for NNs.

1. Motivation

- **Linearised Laplace turns Neural Nets into Linear Gaussian Models**
- **Linear model inference is $\mathcal{O}(m^3n^3)$ or $\mathcal{O}(d^3)$**

$$y = \Phi\theta + \epsilon$$
$$\theta \sim \mathcal{N}(0, A^{-1})$$
$$\epsilon \sim \mathcal{N}(0, B^{-1})$$



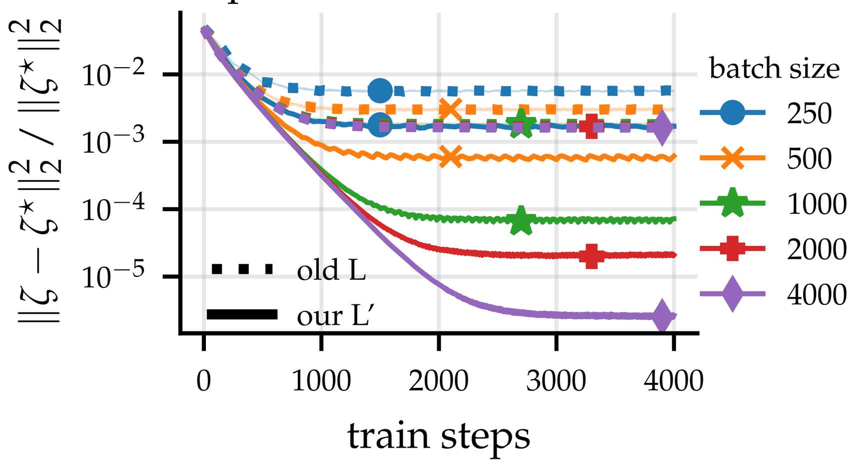
2. Scalable Sampling Inference

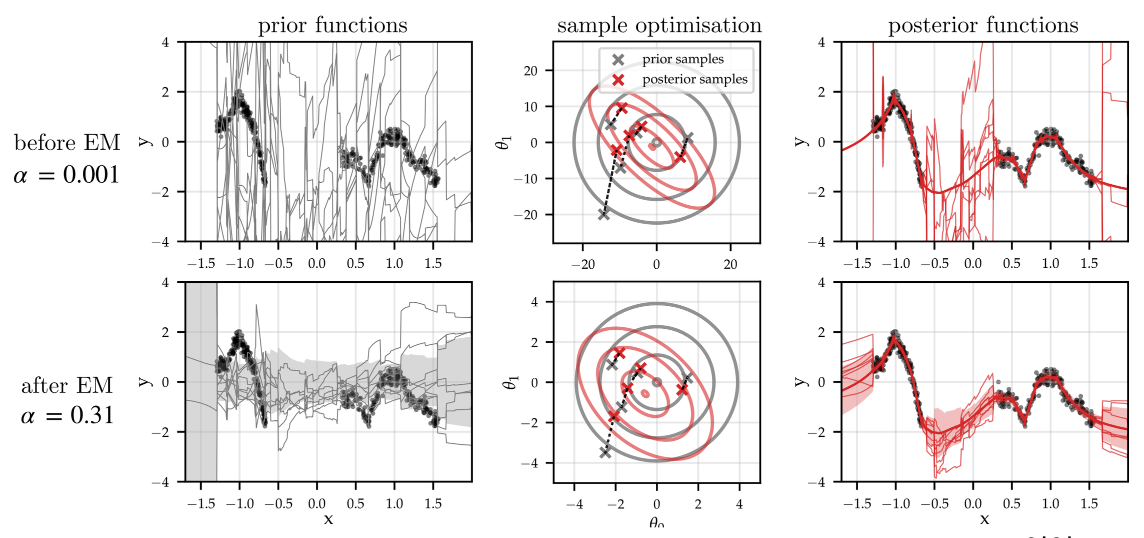
E-step: minibatchable **sample**-then-**optimise**

Naive StO $\|\Phi z - \epsilon_{\text{noise}}\|_B^2 + \|z - \theta_{\text{prior}}\|_A^2$

Low Variance StO
$$\begin{cases} \theta_0 \sim \mathcal{N}(0,\ A^{-1} + A^{-1}\Phi^T B \Phi A^{-1}) \\ \zeta = \operatorname{argmin}_z \|\Phi z\|_B^2 + \|z - \theta_0\|_A^2 \end{cases}$$

Comparison:

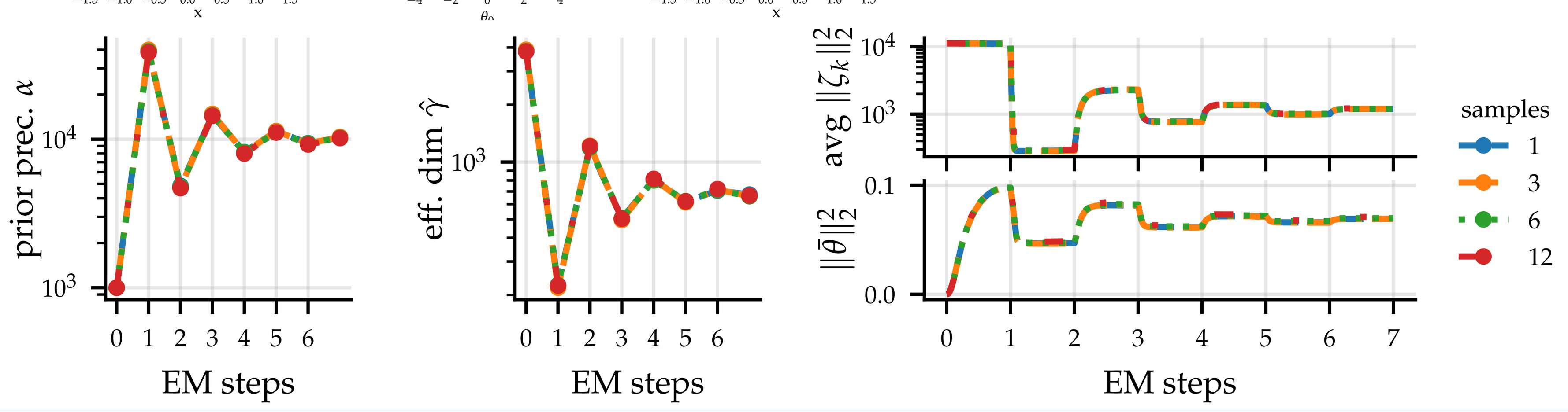




M-step: stochastic **"Mackay update"**

$$A = \alpha I$$

$$\alpha_{\text{new}} = \frac{\mathbb{E}[\zeta^T \Phi^T B \Phi \zeta]}{\|\theta_{\text{MAP}}\|^2}$$



3. Demonstration

Computed Tomography + U-Net

(n=1, m=15k, d=2M)

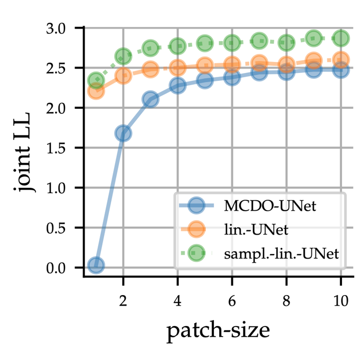
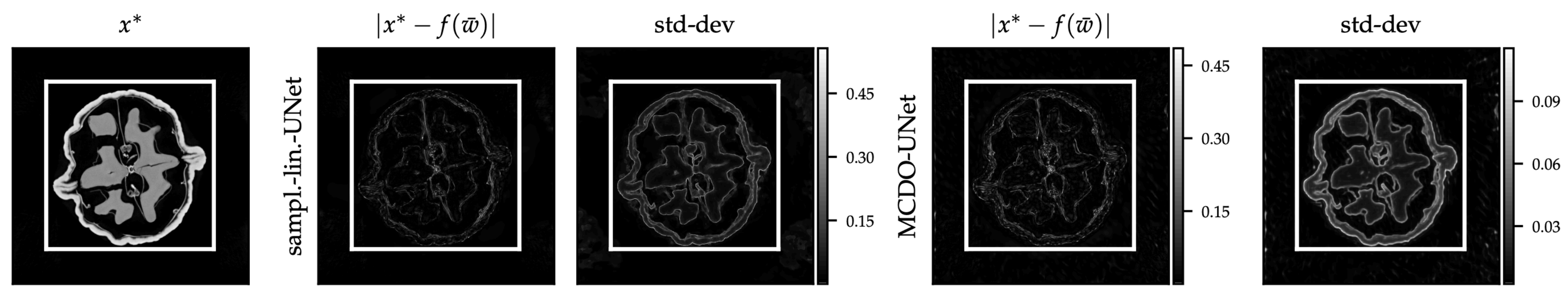


Table 3: Tomographic reconstruction: test LL and wall-clock times (A100 GPU) for both data sizes.

Method	$m = 7680$ LL marginal	$m = 7680$ LL (10×10)	$m = 7680$ wall-clock time (min.) params optim.	$m = 7680$ wall-clock time (min.) prediction	$m = 15360$ LL marginal	$m = 15360$ LL (10×10)	$m = 15360$ wall-clock time (min.) params. optim.	$m = 15360$ wall-clock time (min.) prediction
MCDO-UNet	0.028	2.474	0	3′	0.002	2.762	0	3′
lin.-UNet	2.214	2.601	1260′	196′	–	–	–	–
sampl.-lin.-UNet	**2.341**	**2.869**	12′	14′	**2.310**	**2.972**	15′	14′



Imagenet + ResNet50

(n=1.1M, m=1000, d=25M)

	κ	MAP	Ensemble 5 NNs	KFAC 5EM	Sampling α=11.4
marginal LL	1	-0.936	**-0.815**	-1.493	-0.917
joint LL	2	-9.347	-6.700	-6.286	**-5.611**
	3	-18.733	-13.268	-12.246	**-10.675**
	4	-28.093	-20.029	-20.493	**-16.154**
	5	-37.416	-26.938	-31.221	**-21.981**

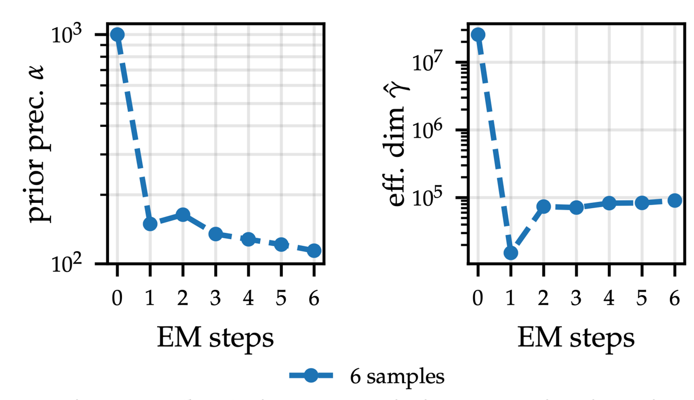


Figure 6: Prior precision optimisation trajectories for ResNet-50 on Imagenet.

CIFAR100 + ResNet18

(n=50k, m=100, d=11M)

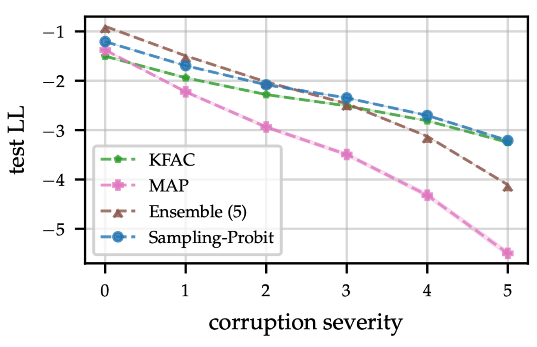


Figure 5: Performance under distribution shift for ResNet-18 on CIFAR100.

	κ	MAP	Ensemble (5)	KFAC	Sampling
marginal LL	1	-1.40 ± 0.00	$\mathbf{-0.90 \pm 0.00}$	-1.12 ± 0.01	-1.07 ± 0.01
joint LL	2	-13.97 ± 0.01	-6.86 ± 0.01	$\mathbf{-4.92 \pm 0.04}$	-5.14 ± 0.04
	3	-27.89 ± 0.03	-14.17 ± 0.03	-10.83 ± 0.12	$\mathbf{-10.77 \pm 0.09}$
	4	-41.83 ± 0.03	-22.29 ± 0.04	-19.02 ± 0.22	$\mathbf{-18.04 \pm 0.18}$
	5	-55.89 ± 0.02	-31.07 ± 0.09	-29.40 ± 0.40	$\mathbf{-26.75 \pm 0.26}$

Table 1: Comparison of methods' marginal and joint prediction performance for ResNet-18 on CIFAR100.