# Sampling-based Inference for Large Linear Models, with Application to Linearised Laplace

Cambridge NeurIPS Meetup, Dec. 8 2023

*Shreyas Padhy*

UNIVERSITY OF CAMBRIDGE

# Summary

# Summary

Bayesian Linear Models are very useful in many fields!

1. Uncertainty Estimation in NNs (through linearisation)
2. Climate Prediction, Economics, Geology, Computational Biology
3. Bandits / RL

# Summary

Bayesian Linear Models are very useful in many fields!

1. Uncertainty Estimation in NNs (through linearisation)
2. Climate Prediction, Economics, Geology, Computational Biology
3. Bandits / RL

**Problem:** **Posterior inference** and **hyperparameter selection** is intractable with millions of observations and millions of parameters due to **cubic scaling**.

# Summary

Bayesian Linear Models are very useful in many fields!

1. Uncertainty Estimation in NNs (through linearisation)
2. Climate Prediction, Economics, Geology, Computational Biology
3. Bandits / RL

**Problem:** **Posterior inference** and **hyperparameter selection** is intractable with millions of observations and millions of parameters due to **cubic scaling**.

**Solution:** We cast inference and hyperparameter selection as a sequence of **quadratic optimisation problems**. We can solve these relatively easily for high dimensional problems with *roughly* **linear scaling**.

# Uncertainty estimation through NN Linearisation

# Uncertainty estimation through NN Linearisation

- Given a neural network $f : \mathbb{R}^{d'} \to \mathbb{R}^m$ parameterised by $\theta \in \mathbb{R}^d$

# Uncertainty estimation through NN Linearisation

- Given a neural network $f : \mathbb{R}^{d'} \to \mathbb{R}^m$ parameterised by $\theta \in \mathbb{R}^d$
- We estimate uncertainty in $f(x)$ as uncertainty in the tangent **linear** model around MAP $\bar{w}$

$$h(\theta, x) = f(\bar{w}, x) + \nabla_w f(\bar{w}, x)(\theta - \bar{w}), \qquad \theta \sim \mathcal{N}(0, A^{-1})$$

# Uncertainty estimation through NN Linearisation

- Given a neural network $f : \mathbb{R}^{d'} \to \mathbb{R}^m$ parameterised by $\theta \in \mathbb{R}^d$
- We estimate uncertainty in $f(x)$ as uncertainty in the tangent **linear** model around MAP $\bar{w}$

$$h(\theta, x) = f(\bar{w}, x) + \nabla_w f(\bar{w}, x)(\theta - \bar{w}), \qquad \theta \sim \mathcal{N}(0, A^{-1})$$

- For a Gaussian likelihood (i.e regression), the linear model's posterior is Gaussian.
- Approximate the predictive distribution of the NN as

$$\mathcal{N}(f(\bar{w}, x),\ \phi(x)\Sigma\phi(x)^T)$$

where $\phi(x) = \nabla_w f(\bar{w}, x)$ and $\Sigma$ a posterior covariance over $\theta$

# Uncertainty estimation through NN Linearisation

- Given a neural network $f : \mathbb{R}^{d'} \to \mathbb{R}^m$ parameterised by $\theta \in \mathbb{R}^d$
- We estimate uncertainty in $f(x)$ as uncertainty in the tangent **linear** model around MAP $\bar{w}$

$$h(\theta, x) = f(\bar{w}, x) + \nabla_w f(\bar{w}, x)(\theta - \bar{w}), \qquad \theta \sim \mathcal{N}(0, A^{-1})$$

- For a Gaussian likelihood (i.e regression), the linear model's posterior is Gaussian.
- Approximate the predictive distribution of the NN as

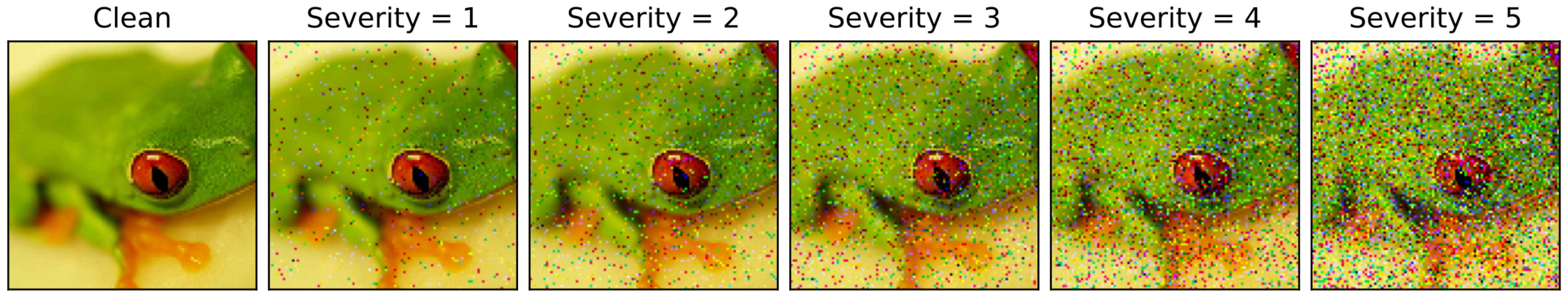
$$\mathcal{N}(f(\bar{w}, x),\ \phi(x)\Sigma\phi(x)^T)$$

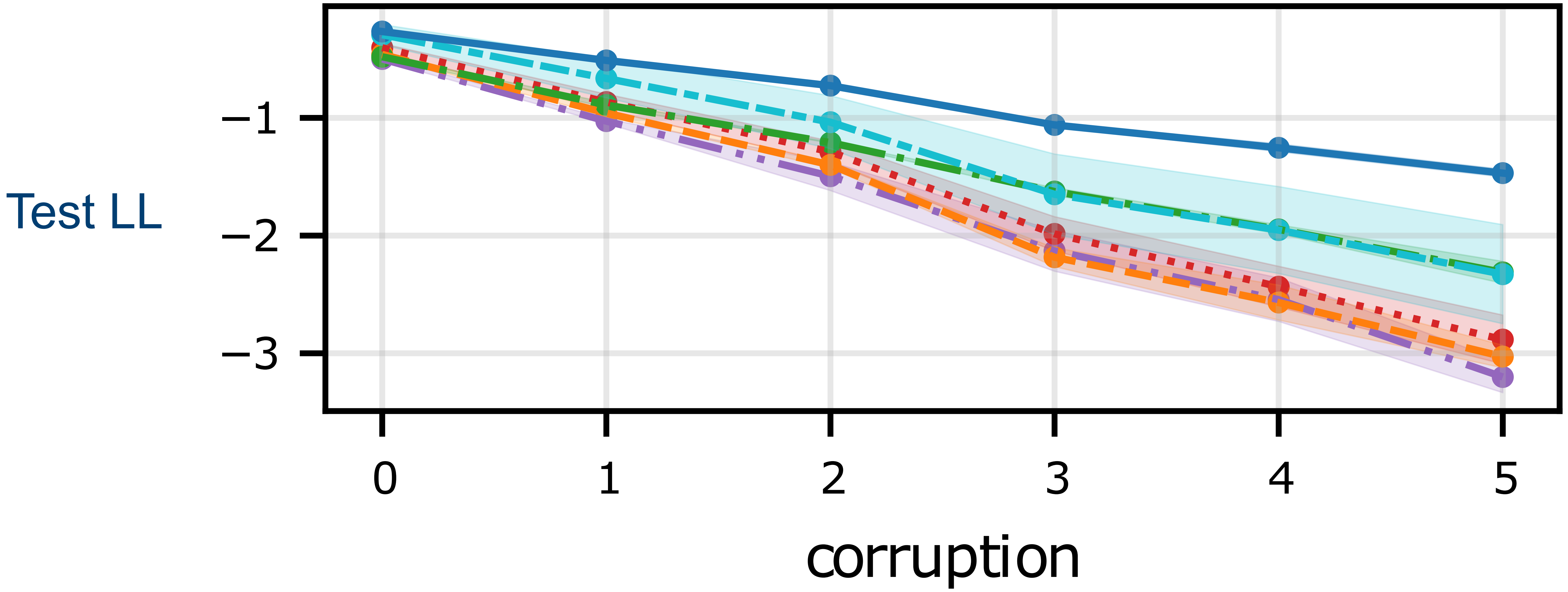
where $\phi(x) = \nabla_w f(\bar{w}, x)$ and $\Sigma$ a posterior covariance over $\theta$

- We can generalise this to non-Gaussian likelihoods (i.e. classification) by 'Gaussianising' with the **Laplace** approximation

# Linearised NNs work well

**Corrupted CIFAR10** (Ovadia 2019)




**Model:**
ResNet-18 with **11M** weights

**Inference:**
**Lin Laplace Subnetwork**
(Daxberger et. al. 2021)

**"Bayesian Deep Learning via Subnetwork Inference"**

**Baselines:**

- MAP
- Diagonal Laplace
- MC Dropout (Gal 2016)
- Deep Ensembles (Lakshminarayanan 2017)
- SWAG (Maddox 2019)

# Consider: Bayesian Linear Models

# Consider: Bayesian Linear Models

$$y_i = \phi(x_i)\theta + \eta_i$$

# Consider: Bayesian Linear Models

$$y_i = \phi(x_i)\theta + \eta_i$$

$$y_i \in \mathbb{R}^m$$

$$\theta \in \mathbb{R}^d$$

$$\phi(x_i) \in \mathbb{R}^{m \times d}$$

$$i \in \{1, \ldots, n\}$$

# Consider: Bayesian Linear Models

$$y_i = \phi(x_i)\theta + \eta_i$$

$$y_i \in \mathbb{R}^m$$

$$\theta \in \mathbb{R}^d$$

$$\phi(x_i) \in \mathbb{R}^{m \times d}$$

$$i \in \{1,\ldots,n\}$$

$$\theta \sim \mathcal{N}(0, A^{-1})$$

$$\eta_i \sim \mathcal{N}(0, B_i^{-1})$$

# Consider: Bayesian Linear Models

$$y_i = \phi(x_i)\theta + \eta_i$$

$$y_i \in \mathbb{R}^m$$

$$\theta \in \mathbb{R}^d$$

$$\phi(x_i) \in \mathbb{R}^{m \times d}$$

$$i \in \{1, \ldots, n\}$$

$$\theta \sim \mathcal{N}(0, A^{-1})$$

$$\eta_i \sim \mathcal{N}(0, B_i^{-1})$$

$$Y = [y_0^T, \ldots, y_n^T]^T \in \mathbb{R}^{nm}$$

# Consider: Bayesian Linear Models

$$y_i = \phi(x_i)\theta + \eta_i$$

$$y_i \in \mathbb{R}^m \quad \rightarrow \quad Y = [y_0^T, \ldots, y_n^T]^T \in \mathbb{R}^{nm}$$

$$\theta \in \mathbb{R}^d$$

$$\phi(x_i) \in \mathbb{R}^{m \times d} \quad \rightarrow \quad \Phi = [\phi(x_0)^T, \ldots, \phi(x_n)^T]^T \in \mathbb{R}^{nm \times d}$$

$$i \in \{1,\ldots,n\}$$

$$\theta \sim \mathcal{N}(0, A^{-1})$$

$$\eta_i \sim \mathcal{N}(0, B_i^{-1})$$

# Consider: Bayesian Linear Models

$$y_i = \phi(x_i)\theta + \eta_i$$

$y_i \in \mathbb{R}^m$ → $Y = [y_0^T, \ldots, y_n^T]^T \in \mathbb{R}^{nm}$

$\theta \in \mathbb{R}^d$

$\phi(x_i) \in \mathbb{R}^{m \times d}$ → $\Phi = [\phi(x_0)^T, \ldots, \phi(x_n)^T]^T \in \mathbb{R}^{nm \times d}$

$i \in \{1,\ldots,n\}$

$\theta \sim \mathcal{N}(0, A^{-1})$

$\eta_i \sim \mathcal{N}(0, B_i^{-1})$ → $B = B_i \otimes I_n \in \mathbb{R}^{nm \times nm}$

# Consider: Bayesian Linear Models

- Number of parameters is large $d > 1e6$
- Observation space is large $n \cdot m > 1e6$

$$y_i = \phi(x_i)\theta + \eta_i$$

$$y_i \in \mathbb{R}^m \rightarrow Y = [y_0^T, \ldots, y_n^T]^T \in \mathbb{R}^{nm}$$

$$\theta \in \mathbb{R}^d$$

$$\phi(x_i) \in \mathbb{R}^{m \times d} \rightarrow \Phi = [\phi(x_0)^T, \ldots, \phi(x_n)^T]^T \in \mathbb{R}^{nm \times d}$$

$$i \in \{1, \ldots, n\}$$

$$\theta \sim \mathcal{N}(0, A^{-1})$$

$$\eta_i \sim \mathcal{N}(0, B_i^{-1}) \rightarrow B = B_i \otimes I_n \in \mathbb{R}^{nm \times nm}$$

# Inference in Bayesian Linear Models

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}||Y - \Phi\theta||_B^2 + ||\theta||_A^2$ is quadratic with curvature $H = \Phi^T B \Phi + A$

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.

2. Tune the L2 regularisation strength $A$.

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}||Y - \Phi\theta||_B^2 + ||\theta||_A^2$ is quadratic with curvature $H = \Phi^T B \Phi + A$

$\rightarrow$ both tasks can be performed in closed form:

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}||Y - \Phi\theta||_B^2 + ||\theta||_A^2$ is quadratic with curvature $H = \Phi^T B \Phi + A$

$\rightarrow$ both tasks can be performed in closed form:

1. Posterior is $\mathcal{N}(\bar{\theta}, H^{-1}), \quad \bar{\theta} = H^{-1}\Phi^T B Y$

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}||Y - \Phi\theta||_B^2 + ||\theta||_A^2$ is quadratic with curvature $H = \Phi^T B \Phi + A$

$\rightarrow$ both tasks can be performed in closed form:

1. Posterior is $\mathscr{N}(\bar{\theta}, H^{-1}), \quad \bar{\theta} = H^{-1}\Phi^T B Y$
2. Model evidence can be used to tune hyperparameters, $\mathscr{M}(A) = \mathscr{L}(\bar{\theta}) - \frac{1}{2}\log\det H$

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}\|Y - \Phi\theta\|_B^2 + \|\theta\|_A^2$ is quadratic with curvature $H = \Phi^T B \Phi + A$

$\rightarrow$ both tasks can be performed in closed form:

1. Posterior is $\mathcal{N}(\bar{\theta}, \underbrace{\boxed{H^{-1}}}_{})$ , $\bar{\theta} = H^{-1}\Phi^T BY$
2. Model evidence can be used to tune hyperparameters, $\mathscr{M}(A) = \mathscr{L}(\bar{\theta}) - \frac{1}{2}\mathrm{logdet} H$

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}||Y - \Phi\theta||_B^2 + ||\theta||_A^2$ is quadratic with curvature $H = \Phi^T B \Phi + A$

$\rightarrow$ both tasks can be performed in closed form:

1. Posterior is $\mathcal{N}(\bar{\theta}, \underbrace{\boxed{H^{-1}}}_{}{}^{\mathcal{O}(d^3)})$, $\quad \bar{\theta} = H^{-1}\Phi^T B Y$
2. Model evidence can be used to tune hyperparameters, $\mathscr{M}(A) = \mathscr{L}(\bar{\theta}) - \frac{1}{2}\overset{\mathcal{O}(d^3)}{\boxed{\log\det H}}$

# Inference in Bayesian Linear Models

**We want to:**

1. Find the posterior distribution over parameters $\theta$.
2. Tune the L2 regularisation strength $A$.

4.5TB (resnet18)

Loss landscape $\mathscr{L}(\theta) = \frac{1}{2}\|Y - \Phi\theta\|_B^2 + \|\theta\|_A^2$ is quadratic with curvature $\boxed{H} = \Phi^T B \Phi + A$

$\rightarrow$ both tasks can be performed in closed form:

$\mathcal{O}(d^3)$ $\mathcal{O}(d^3)$

1. Posterior is $\mathcal{N}(\bar{\theta}, \boxed{H^{-1}})$, $\quad \bar{\theta} = H^{-1}\Phi^T B Y$
2. Model evidence can be used to tune hyperparameters, $\mathscr{M}(A) = \mathscr{L}(\bar{\theta}) - \frac{1}{2}\boxed{\log\det H}$

# Idea 1: Sample from the posterior with stochastic optimisation

$$z^* \sim \mathcal{N}(0, H^{-1}) \;\text{ if }\; z^* = \mathrm{argmin}_z\, L(z)$$

# Idea 1: Sample from the posterior with stochastic optimisation

$$z^* \sim \mathcal{N}(0, H^{-1}) \;\text{ if }\; z^* = \mathrm{argmin}_z\, L(z)$$

$$L(z) = \sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|_{B_i}^2 + \|z - \theta^0\|_A^2$$

$$\epsilon_i \sim \mathcal{N}(0, B_i^{-1})$$

$$\theta^0 \sim \mathcal{N}(0, A^{-1})$$

# Idea 1: Sample from the posterior with stochastic optimisation

$$z^* \sim \mathcal{N}(0, H^{-1}) \;\; \text{if} \;\; z^* = \operatorname{argmin}_z L(z)$$

$$L(z) = \underbrace{\sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|_{B_i}^2}_{1} + \underbrace{\|z - \theta^0\|_A^2}_{2} \qquad \begin{array}{l} \epsilon_i \sim \mathcal{N}(0, B_i^{-1}) \\ \theta^0 \sim \mathcal{N}(0, A^{-1}) \end{array}$$

1. Noise-fit term.
   - Depends on each observation's feature expansion so it needs to be **minibatched**.
     - Very large variance when estimated stochastically.
2. Regularisation term.
   - We can compute its gradient in closed form.

# Solution 1: Variance reduction for stochastic optimisation-based sampler

$$L(z) = \underbrace{\sum_{i=1}^{n} \|\epsilon_i - \phi(x_i) z\|_{B_i}^2}_{1} + \underbrace{\|z - \theta^0\|_A^2}_{2}$$

# Solution 1: Variance reduction for stochastic optimisation-based sampler

$$L(z) = \underbrace{\sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|_{B_i}^2}_{1} + \underbrace{\|z - \theta^0\|_A^2}_{2}$$

$$\downarrow$$

$$L'(z) = \underbrace{\sum_{i=1}^{n} \|\phi(x_i)z\|_{B_i}^2}_{1} + \underbrace{\|z - \theta^n\|_A}_{2}$$

$$\mathscr{E} = [\epsilon_0^T, \ldots, \epsilon_n^T]^T \in \mathbb{R}^{nm}$$

$$\theta^n = \theta^0 + A^{-1}\Phi^T B \mathscr{E}$$

# Solution 1: Variance reduction for stochastic optimisation-based sampler

$$L(z) = \underbrace{\sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|_{B_i}^2}_{1} + \underbrace{\|z - \theta^0\|_A^2}_{2}$$

$$\downarrow$$

$$L'(z) = \underbrace{\sum_{i=1}^{n} \|\phi(x_i)z\|_{B_i}^2}_{1} + \underbrace{\|z - \theta^n\|_A}_{2}$$

$$\mathscr{E} = [\epsilon_0^T, \ldots, \epsilon_n^T]^T \in \mathbb{R}^{nm}$$

$$\theta^n = \theta^0 + A^{-1}\Phi^T B \mathscr{E}$$

Random noise appears as a part of 2, the term for which we can compute exact gradients!

# Solution 1: Variance reduction for stochastic optimisation-based sampler

$$L(z) = \underbrace{\sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|^2_{B_i}}_{1} + \underbrace{\|z - \theta^0\|^2_A}_{2}$$

$$L'(z) = \underbrace{\sum_{i=1}^{n} \overbrace{\|\phi(x_i)z\|^2_{B_i}}^{\mathcal{O}(d)}}_{1} + \underbrace{\|z - \theta^n\|_A}_{2}$$

$$\mathcal{E} = [\epsilon_0^T, \ldots, \epsilon_n^T]^T \in \mathbb{R}^{nm}$$

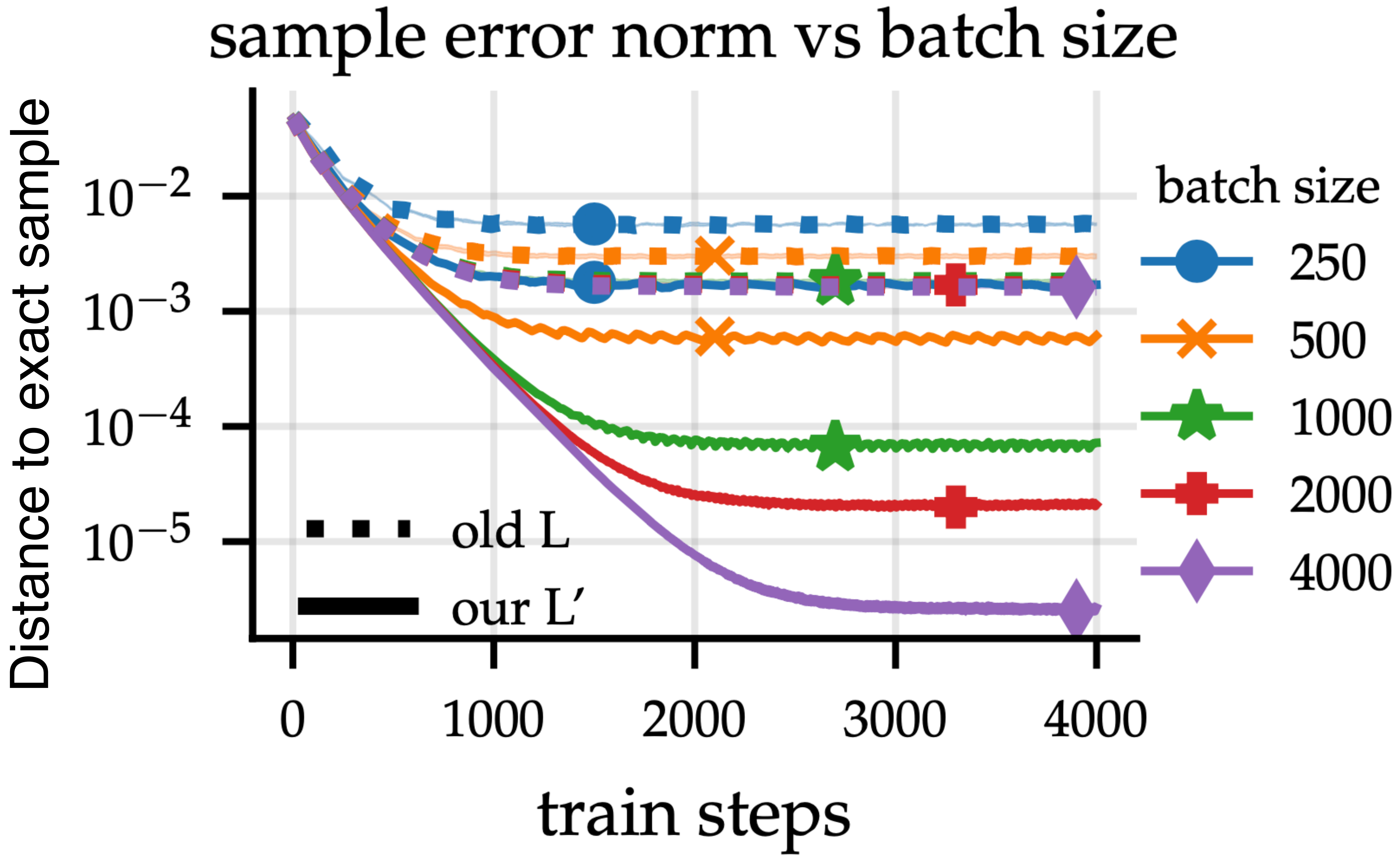
$$\theta^n = \theta^0 + A^{-1}\Phi^T B \mathcal{E}$$

Random noise appears as a part of 2, the term for which we can compute exact gradients!

# L vs L'

$$L(z) = \sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|_{B_i}^2 + \|z - \theta^0\|_A \qquad L'(z) = \sum_{i=1}^{n} \|\phi(x_i)z\|_{B_i}^2 + \|z - \theta^n\|_A$$

Both objectives are equal ($L(z) = L'(z)$) but their mini-batch estimators have different variances!

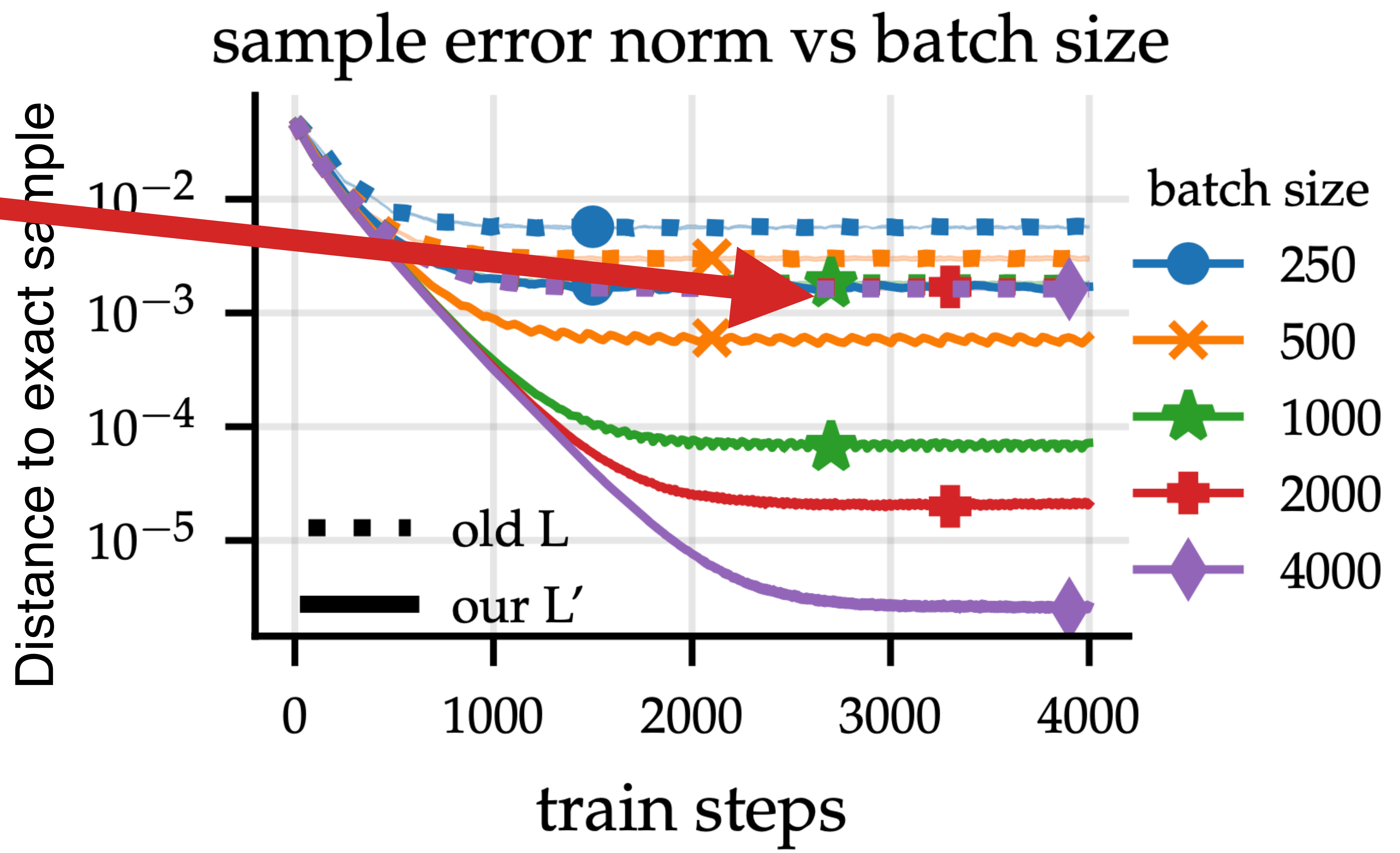


# L vs L’

$$L(z) = \sum_{i=1}^{n} \|\epsilon_i - \phi(x_i)z\|_{B_i}^2 + \|z - \theta^0\|_A \qquad L'(z) = \sum_{i=1}^{n} \|\phi(x_i)z\|_{B_i}^2 + \|z - \theta^n\|_A$$

Both objectives are equal ($L(z) = L'(z)$) but their mini-batch estimators have different variances!

16x reduction in batch size!



# Solution 2: Optimise the model evidence using only samples

# Solution 2: Optimise the model evidence using only samples

- $\log \det H^{-1}$ cannot be estimated from samples…

# Solution 2: Optimise the model evidence using only samples

- $\log \det H^{-1}$ cannot be estimated from samples…
- MacKay proposed an alternative first order optimal update for $\alpha(\text{assume} A = \alpha I)$

$$\alpha = \frac{\mathrm{Tr}(H^{-1}\ \Phi^T \mathrm{B} \Phi)}{||\bar{\theta}||^2} = \frac{\mathrm{Tr}(H^{-1}\ M)}{||\bar{\theta}||^2}$$

# Solution 2: Optimise the model evidence using only samples

- $\log \det H^{-1}$ cannot be estimated from samples…
- MacKay proposed an alternative first order optimal update for $\alpha (\text{assume} A = \alpha I)$

$$\alpha = \frac{\mathrm{Tr}(H^{-1}\ \Phi^T \mathrm{B} \Phi)}{||\bar{\theta}||^2} = \frac{\mathrm{Tr}(H^{-1}\ M)}{||\bar{\theta}||^2}$$

- This *can be* estimated using only samples from the posterior

$$\mathrm{Tr}\left\{H^{-1}M\right\} = \mathrm{Tr}\left\{H^{\frac{-1}{2}}MH^{\frac{-1}{2}}\right\} = \mathbb{E}\left[z_1^T M z_1\right] \approx \frac{1}{k}\sum_{j=1}^{k} z_j^T \Phi^T\ \mathrm{B}\Phi z_j$$

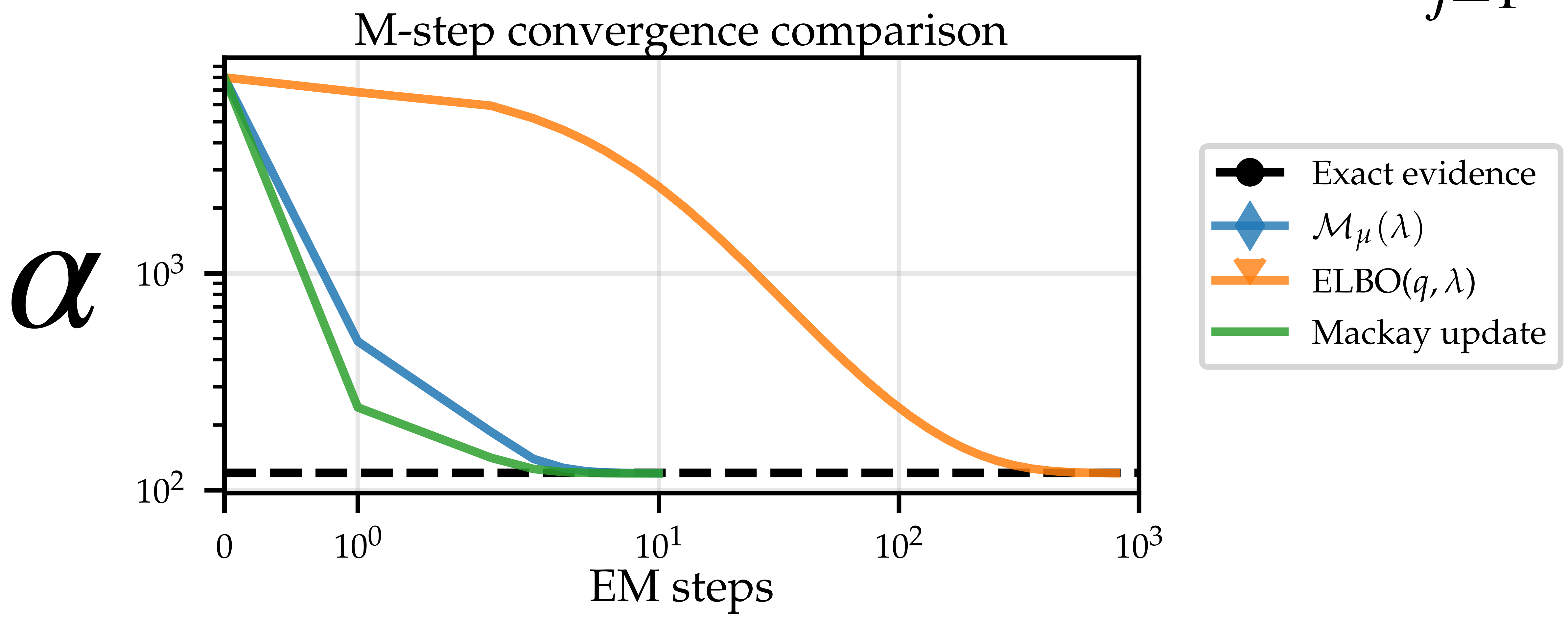
# Solution 2: Optimise the model evidence using only samples

- $\log \det H^{-1}$ cannot be estimated from samples…
- MacKay proposed an alternative first order optimal update for $\alpha (\text{assume} A = \alpha I)$

$$\alpha = \frac{\mathrm{Tr}(H^{-1}\ \Phi^T \mathrm{B} \Phi)}{||\bar{\theta}||^2} = \frac{\mathrm{Tr}(H^{-1}\ M)}{||\bar{\theta}||^2}$$

- This *can be* estimated using only samples from the posterior

$$\mathrm{Tr}\left\{H^{-1}M\right\} = \mathrm{Tr}\left\{H^{\frac{-1}{2}}MH^{\frac{-1}{2}}\right\} = \mathbb{E}\left[z_1^T M z_1\right] \approx \frac{1}{k}\sum_{j=1}^{k} z_j^T \Phi^T\ \mathrm{B}\Phi z_j$$



# Solution 2: Optimise the model evidence using only samples

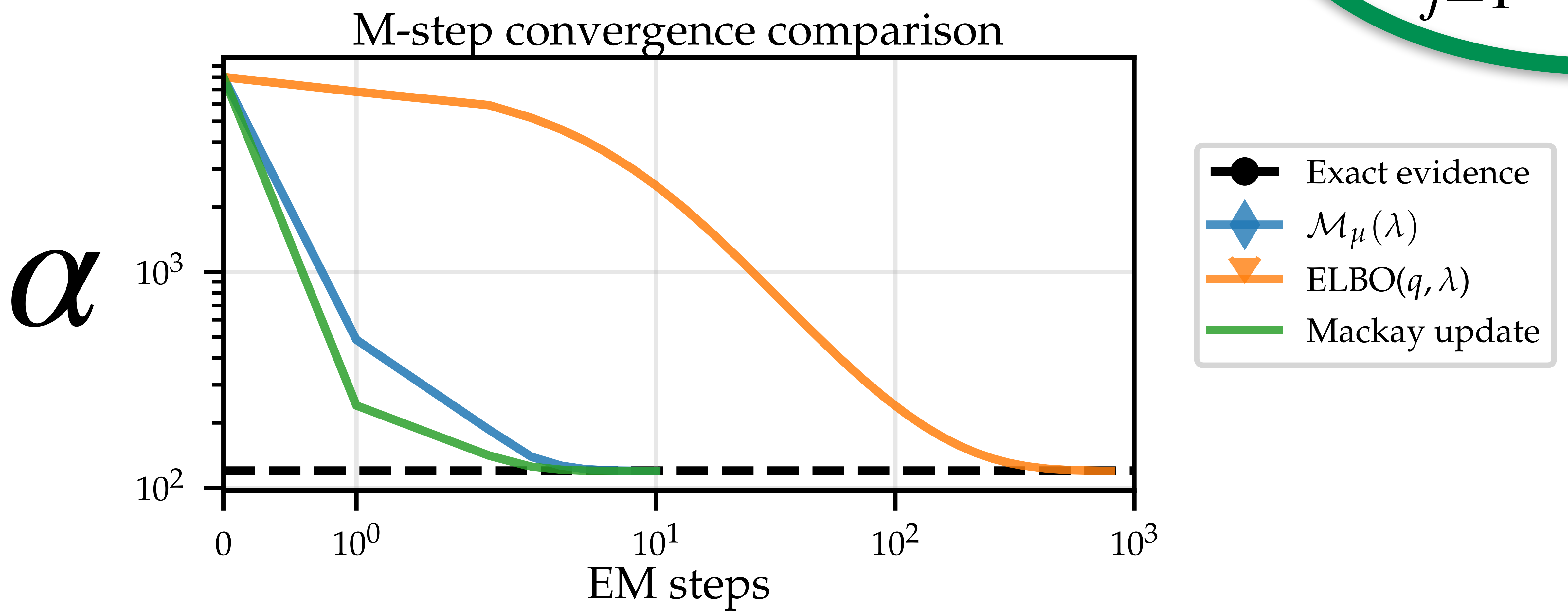
- $\log\det H^{-1}$ cannot be estimated from samples…
- MacKay proposed an alternative first order optimal update for $\alpha (\text{assume} A = \alpha I)$

$$\alpha = \frac{\mathrm{Tr}(H^{-1}\ \Phi^T \mathrm{B}\Phi)}{\|\bar{\theta}\|^2} = \frac{\mathrm{Tr}(H^{-1}\ M)}{\|\bar{\theta}\|^2}$$

- This *can be* estimated using only samples from the posterior

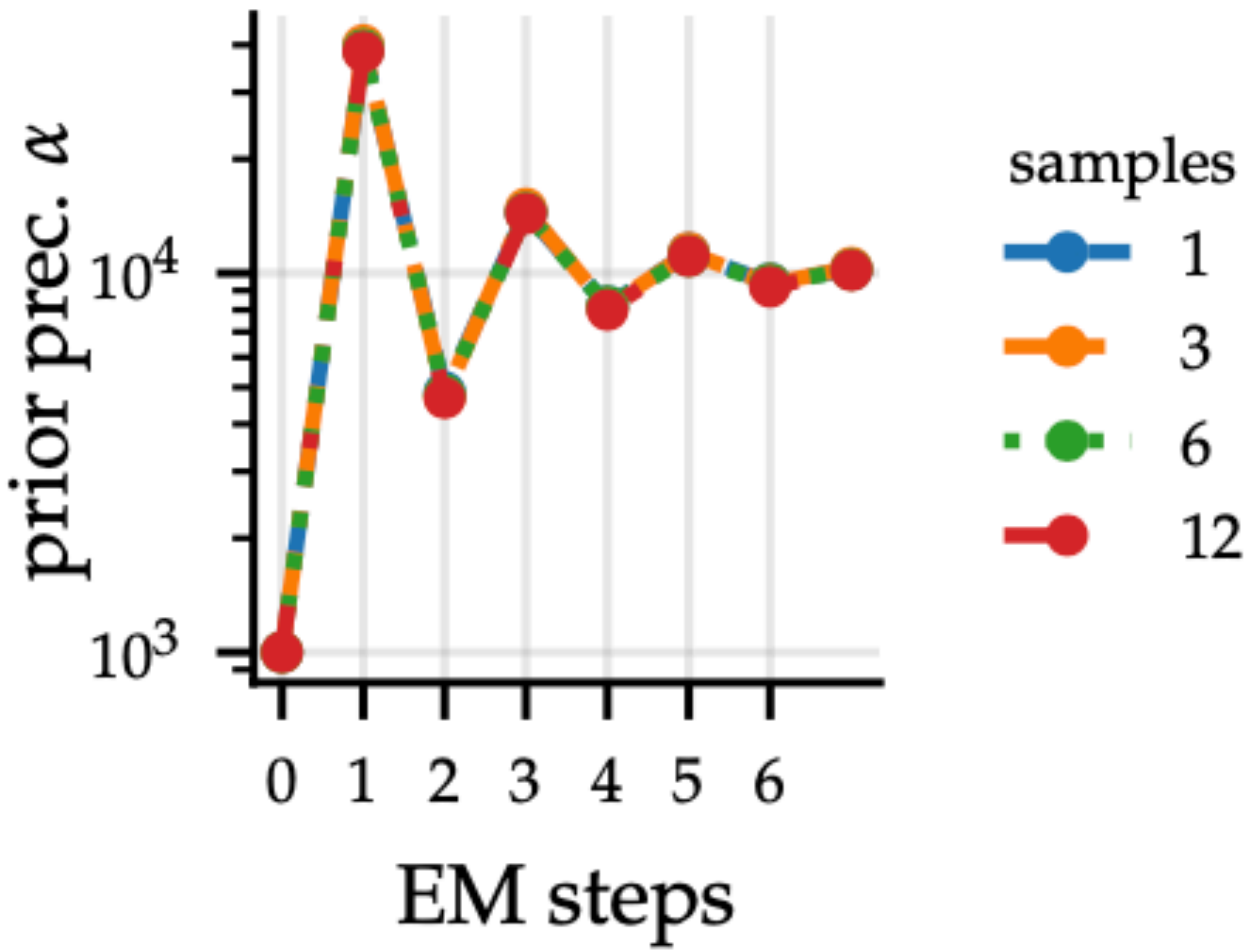
$\mathcal{O}(kdnm)$

$$\mathrm{Tr}\left\{H^{-1}M\right\} = \mathrm{Tr}\left\{H^{\frac{-1}{2}}MH^{\frac{-1}{2}}\right\} = \mathbb{E}\left[z_1^T M z_1\right] \approx \frac{1}{k}\sum_{j=1}^{k} z_j^T \Phi^T\ \mathrm{B}\Phi z_j$$



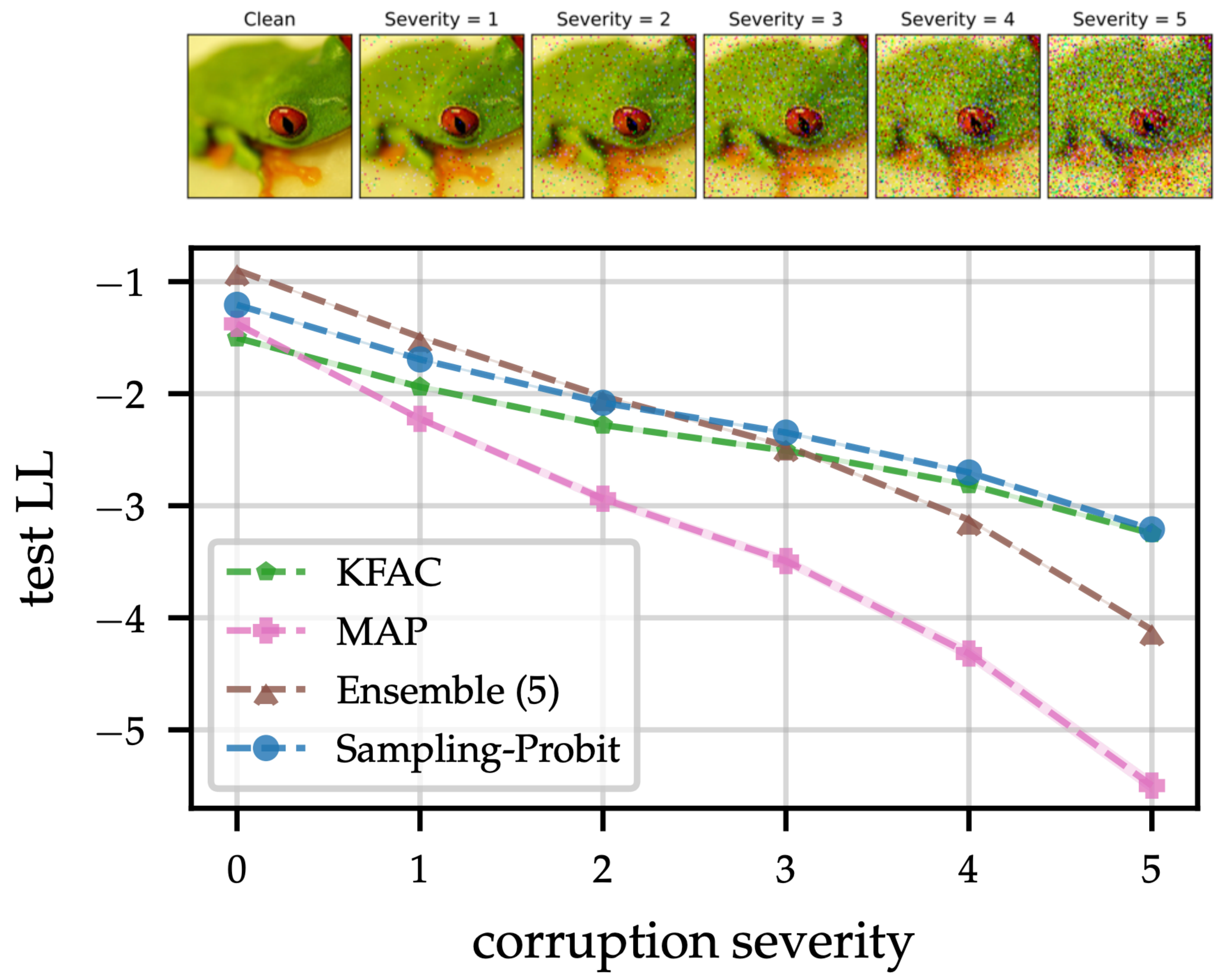
# Stability of estimator: 1 sample is enough

ResNet-18 ($d = 11M$) on CIFAR-100 ($nm = 5M$)



# Demonstration: Scalable Uncertainty Estimation in NNs

ResNet-18 ($d = 11M$) on CIFAR-100 ($nm = 5M$)



# Thank you to my collaborators!

Javier Antorán

Riccardo Barbano

Eric Nalisnick

David Janz

José Miguel
Hernández-Lobato