# Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam

Mohammad Emtiyaz Khan*
RIKEN Center for AI Project (AIP), Tokyo, Japan

Didrik Nielsen* (AIP RIKEN),
Voot Tangkaratt* (AIP RIKEN),
Wu Lin (UBC),
Yarin Gal (University of Oxford),
Akash Srivastava (University of Edinburgh)

*Equal Contribution

# Bayesian Deep Learning

Compute averages over the samples from the posterior distribution

# Approximate Bayesian Inference

Convert Bayesian inference to an optimization problem using Variational Inference (VI), and then use gradient-based methods for optimization

Bayes by Backprop (Blundell et al. 2015), Practical VI (Graves et al. 2011), Black-box VI (Rangnathan et al. 2014) and many more….

# Approximate Bayesian Inference requires more computation, memory, and implementation effort than MLE

Is it possible to reduce these costs?

By replacing gradients with natural-gradients

# Maximum Likelihood Estimation (MLE)

$$\max_\theta \sum_{i=1}^N \log p(\mathcal{D}_i|\theta)$$

Log-likelihood

RMSprop for MLE

$$\theta \leftarrow \mu$$

$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

Backprop on minibatches

$$s \leftarrow (1-\beta)s + \beta g^2$$

Scale vector (gradient-magnitude)

$$\mu \leftarrow \mu + \alpha \, \frac{g}{\sqrt{s} + \delta}$$

Adaptive gradient update

# Gaussian Mean-Field Variational Inference

$$p(\theta) = \mathcal{N}(0, I/\lambda)$$ Known prior precision

$$p(\theta|\mathcal{D}) \approx q(\theta) = \mathcal{N}(\mu, \sigma^2)$$ Covariance matrix = diag($\sigma^2$)

$$\max_{\mu,\sigma^2} \mathcal{L}(\mu, \sigma^2) := \sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathcal{D}_i|\theta)] - KL\left[q(\theta)\|p(\theta)\right]$$

Data-fit term

Regularizer

# MLE vs Gradient-based VI

$$\max_{\theta} \sum_{i=1}^{N} \log p(\mathcal{D}_i|\theta) \qquad \max_{\mu,\sigma^2} \mathcal{L}(\mu,\sigma^2) := \sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathcal{D}_i|\theta)] - KL\Big[q(\theta)\|p(\theta)\Big]$$

RMSprop for Max-likelihood

$$\theta \leftarrow \mu$$

$$g \leftarrow \frac{1}{M}\sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$\mu \leftarrow \mu + \alpha \, \frac{g}{\sqrt{s}+\delta}$$

Gradient-based Variational Inference

$$\mu \leftarrow \mu + \alpha \, \frac{\hat{\nabla}_\mu \mathcal{L}}{\sqrt{s_\mu}+\delta}$$

$$\sigma \leftarrow \sigma + \alpha \, \frac{\hat{\nabla}_\sigma \mathcal{L}}{\sqrt{s_\sigma}+\delta}$$

(Graves et al. 2011, Blundell et al. 2015)

# MLE vs Natural-Gradient VI

RMSprop for Max-likelihood

$$\theta \leftarrow \mu$$

$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$\mu \leftarrow \mu + \alpha \, \frac{g}{\sqrt{s} + \delta}$$

Natural-Gradient VI (Khan, Lin 2017, Khan, Nielsen 2018)

$$\theta \leftarrow \mu + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, Ns + \lambda)$$

$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta \frac{1}{M} \sum_i \nabla^2_{\theta\theta} \log p(\mathcal{D}_i|\theta)$$

$$\mu \leftarrow \mu + \alpha \, \frac{g + \lambda\mu/N}{s + \lambda/N}$$

Variational Online-Newton (VON)

Khan et al. 2017

# MLE vs Natural-Gradient VI

RMSprop for Max-likelihood

$$\theta \leftarrow \mu$$

$$g \leftarrow \frac{1}{M}\sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$\mu \leftarrow \mu + \alpha \frac{g}{\sqrt{s}+\delta}$$

Natural-Gradient VI (Khan, Lin 2017, Khan, Nielsen 2018)

$$\theta \leftarrow \mu + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, Ns+\lambda)$$

$$g \leftarrow \frac{1}{M}\sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta \frac{1}{M}\sum_i \left[\nabla_\theta \log p(\mathcal{D}_i|\theta)\right]^2$$

$$\mu \leftarrow \mu + \alpha \frac{g+\lambda\mu/N}{s+\lambda/N}$$

Variational Online Gauss-Newton (VOGN)

# MLE vs Natural-Gradient VI

## RMSprop for Max-likelihood

$$\theta \leftarrow \mu$$

$$g \leftarrow \tfrac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$\mu \leftarrow \mu + \alpha \, \frac{g}{\sqrt{s}+\delta}$$

## Natural-Gradient VI (Khan, Lin 2017, Khan, Nielsen 2018)

$$\theta \leftarrow \mu + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, Ns+\lambda)$$

$$g \leftarrow \tfrac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$\mu \leftarrow \mu + \alpha \, \frac{g + \lambda\mu/N}{\sqrt{s} + \lambda/N}$$

Variational RMSprop (Vprop)

# Variational Adam (Vadam)

## Adam for Max-likelihood

$$\theta \leftarrow \mu$$

$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$m \leftarrow (1-\gamma)m + \gamma g$$

$$\hat{m} \leftarrow m/(1-(1-\gamma)^t)$$

$$\hat{s} \leftarrow s/(1-(1-\beta)^t)$$

$$\mu \leftarrow \mu + \alpha \, \frac{\hat{m}}{\sqrt{\hat{s}} + \delta}$$

## Vadam for VI

$$\theta \leftarrow \mu + \epsilon, \;\; \text{where } \mathcal{N}(0, Ns + \lambda)$$

$$g \leftarrow \frac{1}{M} \sum_i \nabla_\theta \log p(\mathcal{D}_i|\theta)$$

$$s \leftarrow (1-\beta)s + \beta g^2$$

$$m \leftarrow (1-\gamma)m + \gamma(g + \lambda\mu/N)$$

$$\hat{m} \leftarrow m/(1-(1-\gamma)^t)$$

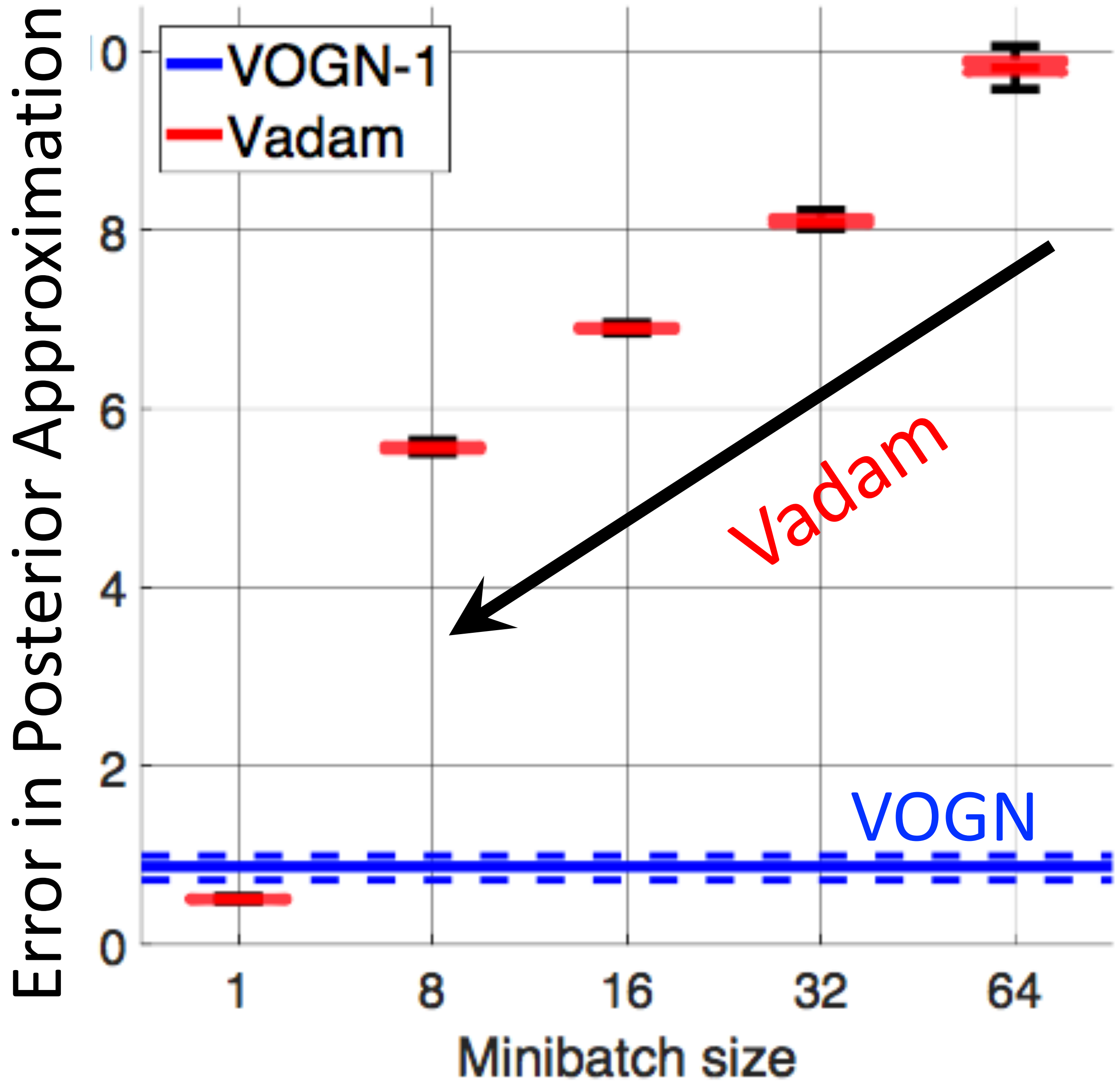
$$\hat{s} \leftarrow s/(1-(1-\beta)^t)$$

$$\mu \leftarrow \mu + \alpha \, \frac{\hat{m}}{\sqrt{\hat{s}} + \lambda/N}$$

# Summary: Uncertainty using Adam

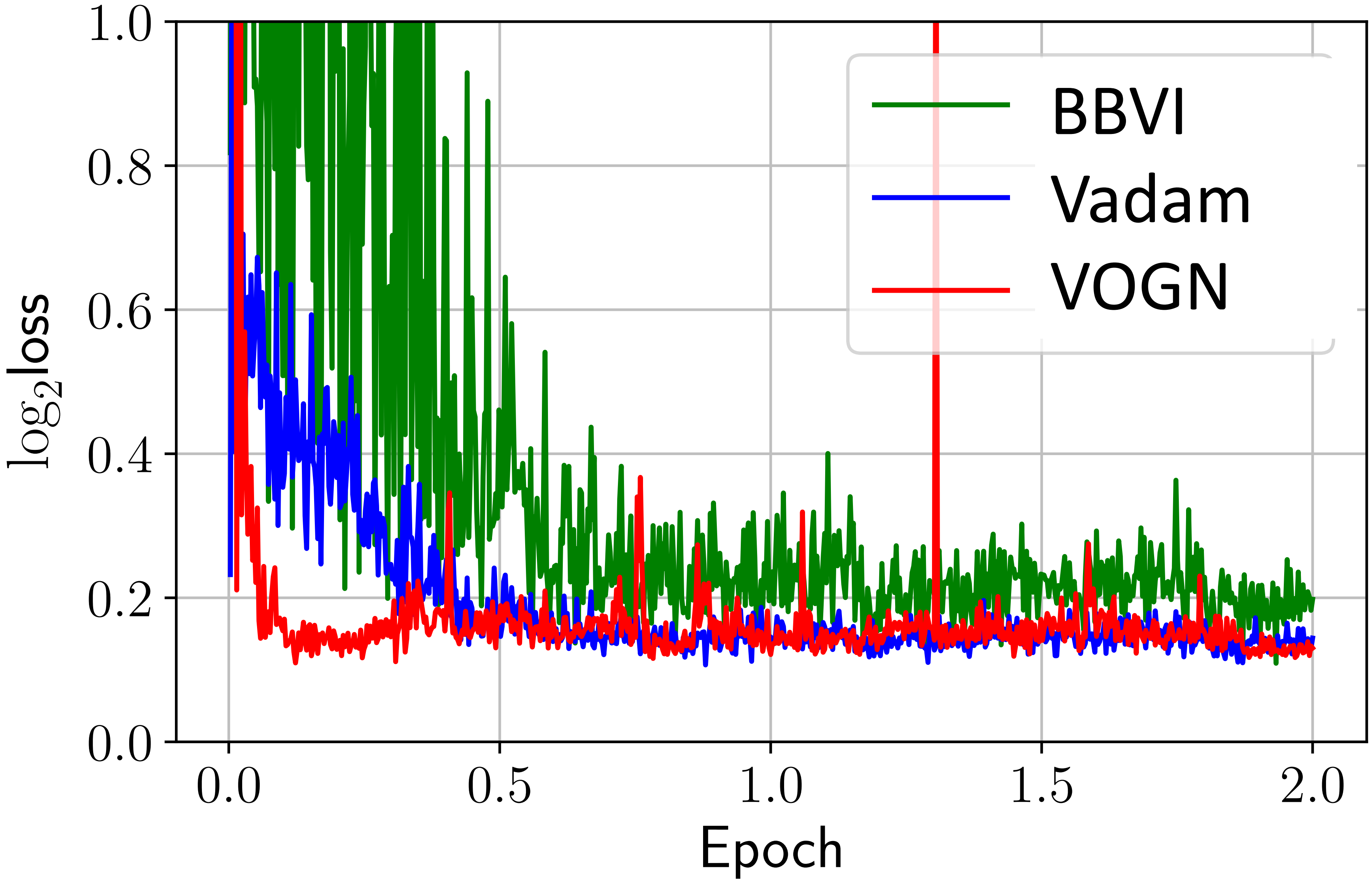
Perturb the weights before backprop.

Choose a small minibatch size.




Bayesian logistic regression on “Breast-Cancer” (N=683, D=8)

As we reduce the minibatch size, Vadam gives similar performance as VOGN.




1 layer 64 hidden Units with ReLu on Breast Cancer [N=683, D=10]

VOGN shows fast convergence

# Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam

Poster tonight (Hall B #190)

Code available at https://github.com/emtiyaz/vadam/

Also check out "Noisy Natural-gradient as VI" by Zhang et al. at this conferene