(Semi-)Implicit Modeling: New Deep Tool for Approximate Bayesian Inference

Dmitry Vetrov

Bayesian Methods Research Group
National Research University HSE
Samsung AI Center, Moscow

bayesgroup.ru/people/dmitry-vetrov

SAMSUNG
AI Center

Frequentist vs. Bayesian

	Frequentist	Bayesian
Randomness	Objective indefiniteness	Subjective ignorance
Variables	Random and deterministic	Everything is random
Inference	Maximum likelihood $p(D\|\theta) \to \max_\theta$	Bayes theorem $p(\theta\|D) = \frac{p(D\|\theta)p(\theta)}{\int p(D\|\theta)p(\theta)d\theta}$
Estimates	ML-estimates	Posterior or MAP-estimates
Applicability	$n \gg d$	$n \geq 0$

Frequentist vs. Bayesian

- It can be shown that

$$\lim_{n/d \to \infty} p(\theta|D) = \delta(\theta - \theta_{ML}).$$

- There is NO contradiction between the two approaches!
- Frequentist framework is a **limit case** of Bayesian one!

Frequentist vs. Bayesian

- It can be shown that

$$\lim_{n/d \to \infty} p(\theta|D) = \delta(\theta - \theta_{ML}).$$

- There is NO contradiction between the two approaches!
- Frequentist framework is a **limit case** of Bayesian one!
- The number of tunable parameters in modern ML models is comparable with the sizes of training data $d \gtrsim n$
- We have no choice but to be Bayesian :)

Bayesian machine learning

- At the learning stage we need to infer the posterior over the parameters of our model

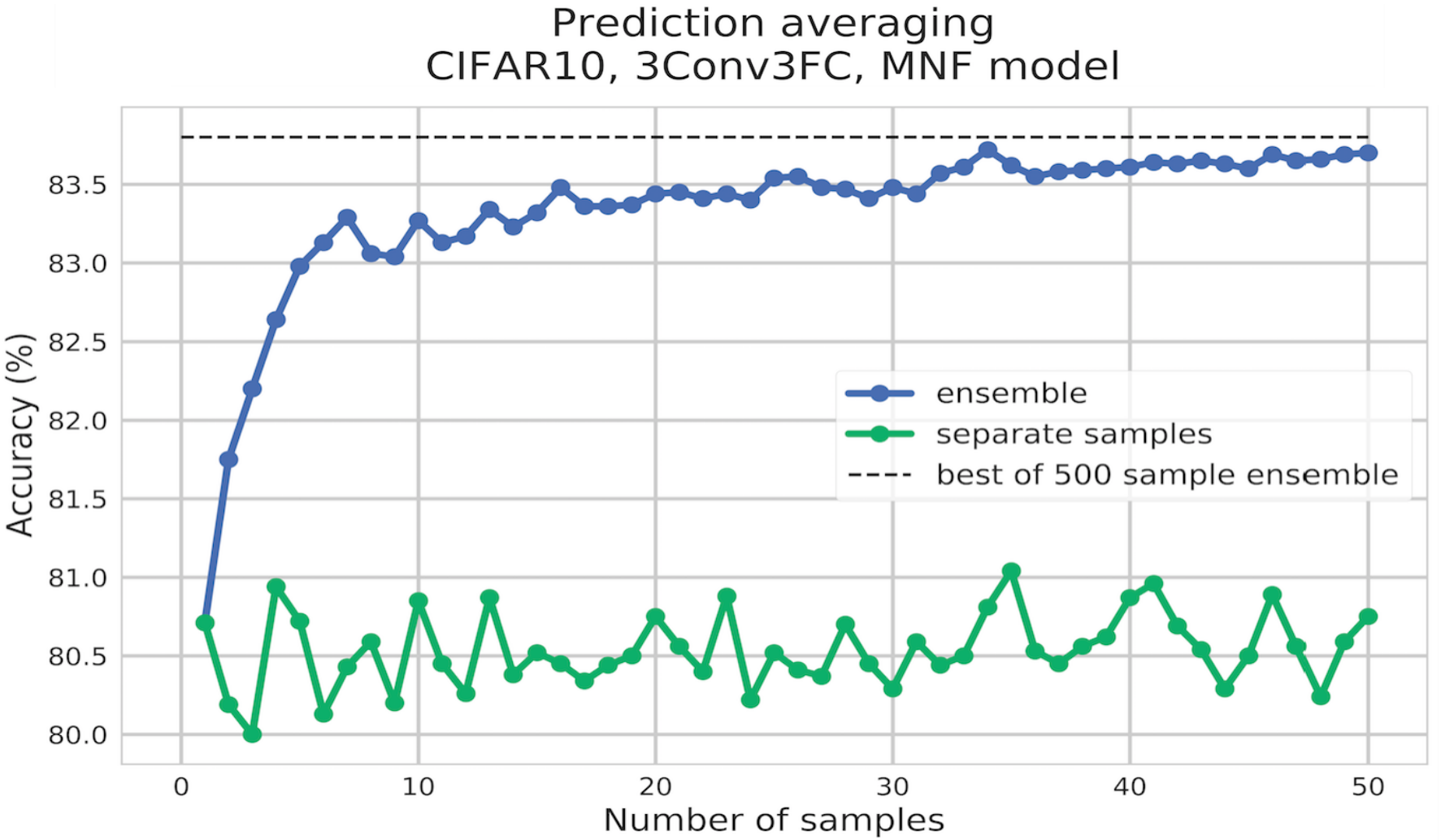
$$p(\theta|D_{tr}) = \frac{p(D_{tr}|\theta)p(\theta)}{\int p(D_{tr}|\theta)p(\theta)d\theta}$$

- At the test stage we evaluate predictive distribution

$$p(D_{test}|D_{tr}) = \int p(D_{test}|\theta)p(\theta|D_{tr})d\theta$$

by performing weighted averaging or **ensembling** w.r.t. the posterior

Ensembling


Prediction averaging
CIFAR10, 3Conv3FC, MNF model
Accuracy (%)
83.5
83.0
82.5
82.0
81.5
81.0
80.5
80.0
0
10
20
30
40
50
Number of samples
ensemble
separate samples
best of 500 sample ensemble

Bayesian ML: practice

“If everything is so well then why is it so bad?”

— M. Zhvanetsky, Soviet comedian

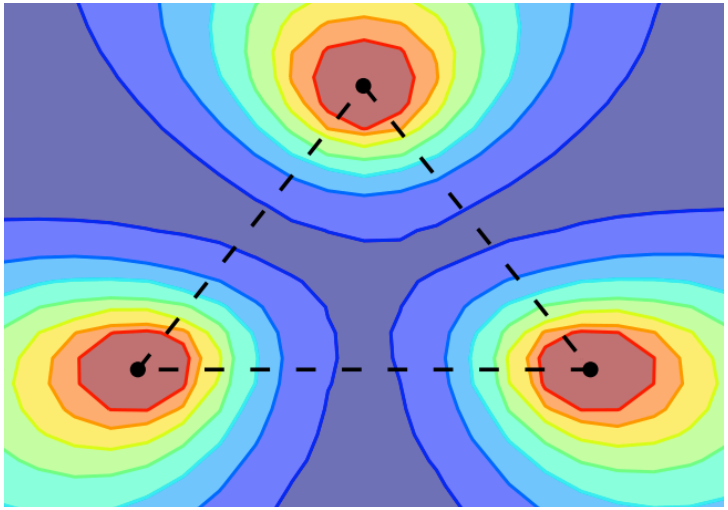
Bayesian ML: practice


Timur Garipov

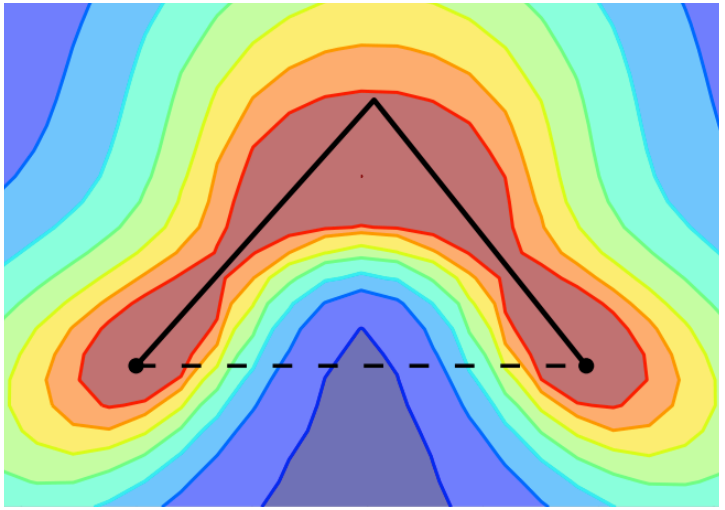

Pavel Izmailov


Dmitry Podoprikhin

- The highly-dimensional integrals are generally intractable
- Mode-connectivity effect (Garipov18, Draxler18) confirms that in DNNs the true posteriors have extremely complicated structure



Connecting three optima for independently trained networks

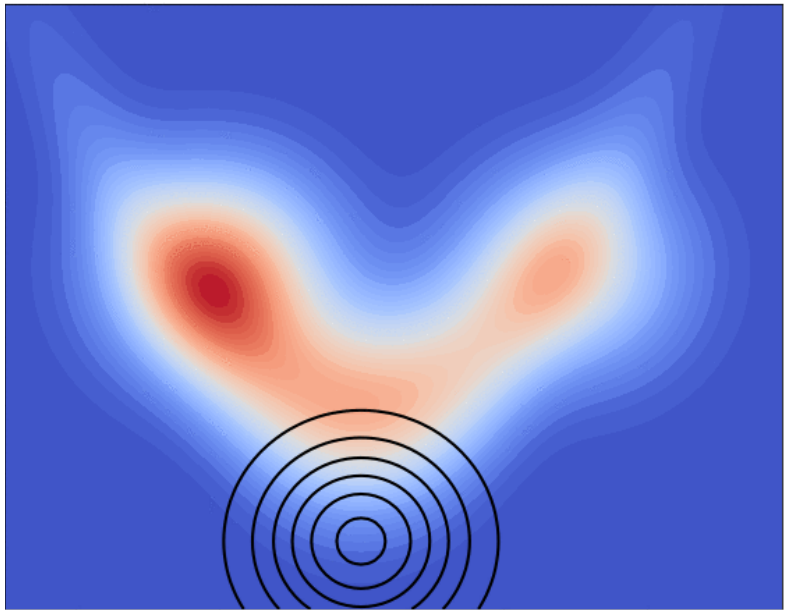


A polygonal chain with one bend, connecting the lower two optima

(Garipov18) T. Garipov et al. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs, NIPS 2018
(Draxler18) F. Draxler et al. Essentially No Barriers in Neural Network Energy Landscape, ICML 2018

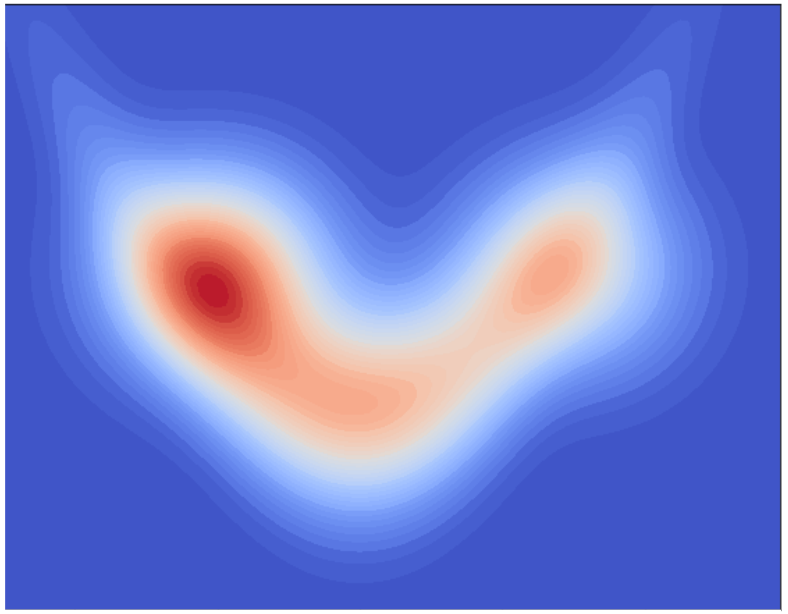
Main inference techniques

Variational inference



- Approximates intractable true posterior with a tractable variational distribution
- Typically KL-divergence is minimized
- Can be scaled up by stochastic optimization

Markov Chain Monte Carlo



- Generates samples from the true posterior
- No bias even if the true distribution is intractable
- Quite slow in practice
- Problematic scaling to large data

Main inference techniques

	MCMC	This Talk	Variational Bayes
Bias	**No**	?	**Strong**
Ensembling	**Inefficient**	?	**Efficient**
Density	**No**	?	**Yes**
Empirical Bayes	**Not available**	?	**Rough**

(Explicit) variational Bayes

- Currently the most efficient scheme due to **(local) reparameterization trick** (Kingma15)
- We mostly have to deal with very simple variational approximations

$$q(\theta) \approx p(\theta|D)$$

such as fully-factorized gaussians

(Kingma15) D.P. Kingma, T. Salimans, M. Welling. Variational Dropout and the Local Reparameterization Trick. NIPS, 2015

(Explicit) variational Bayes

- Currently the most efficient scheme due to **(local) reparameterization trick** (Kingma15)
- We mostly have to deal with very simple variational approximations

$$q(\theta) \approx p(\theta|D)$$

such as fully-factorized gaussians

- Minimization of KL-divergence

$$KL(q(\theta)||p(\theta|D)) \to \min_{q(.)}$$

is equivalent to the maximization of variational lower bound (ELBO)

$$\mathcal{L}(q) = \mathbb{E}_{q(\theta)} \log p(D|\theta) - KL(q(\theta)||p(\theta)) \to \max_{q(.)}$$

(Kingma15) D.P. Kingma, T. Salimans, M. Welling. Variational Dropout and the Local Reparameterization Trick. NIPS, 2015

(Explicit) variational Bayes

- At the learning stage we need to approximate

$$q(\theta) \approx p(\theta|D_{tr}) = \frac{p(D_{tr}|\theta)p(\theta)}{\int p(D_{tr}|\theta)p(\theta)d\theta}$$

- At the test stage we estimate predictive distribution

$$p(D_{test}|D_{tr}) = \int p(D_{test}|\theta)p(\theta|D_{tr})d\theta \approx \int p(D_{test}|\theta)q(\theta)d\theta$$

- Since the integral in intractable in DL models we estimate in via Monte Carlo

$$p(D_{test}|D_{tr}) \approx \frac{1}{m}\sum_{i=1}^{m} p(D_{test}|\theta_i), \quad \theta_i \sim q(\theta)$$

- Then all we need is to be able to sample from $q(\theta)$

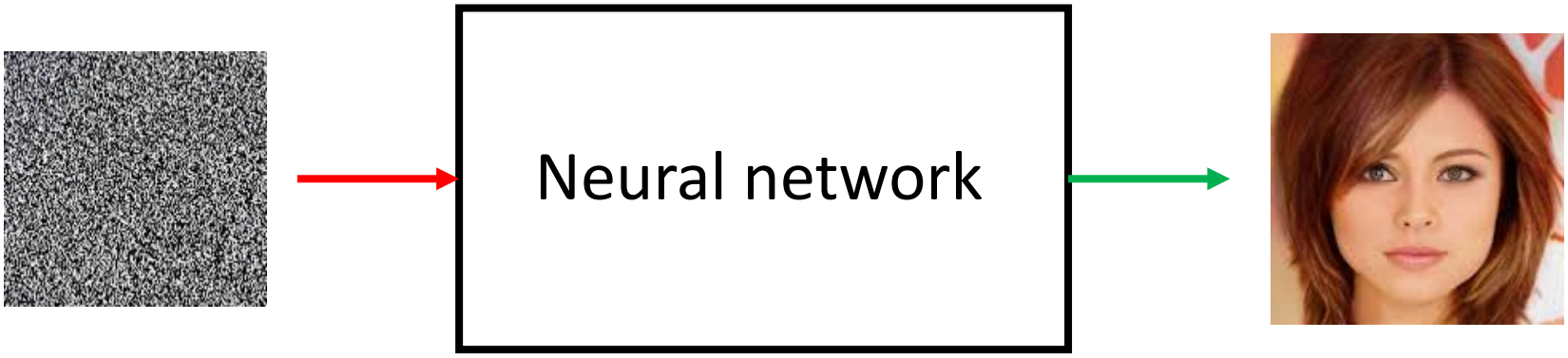
Implicit probabilistic model

- Most known example is generator in GANs
- Converts random noise of simple structure into complicated multi-dimensional distribution

$$\theta = g_\phi(\xi), \quad \xi \sim p(\xi)$$

  using non-linear function parameterized by ϕ

- No access to the density in the space of θ

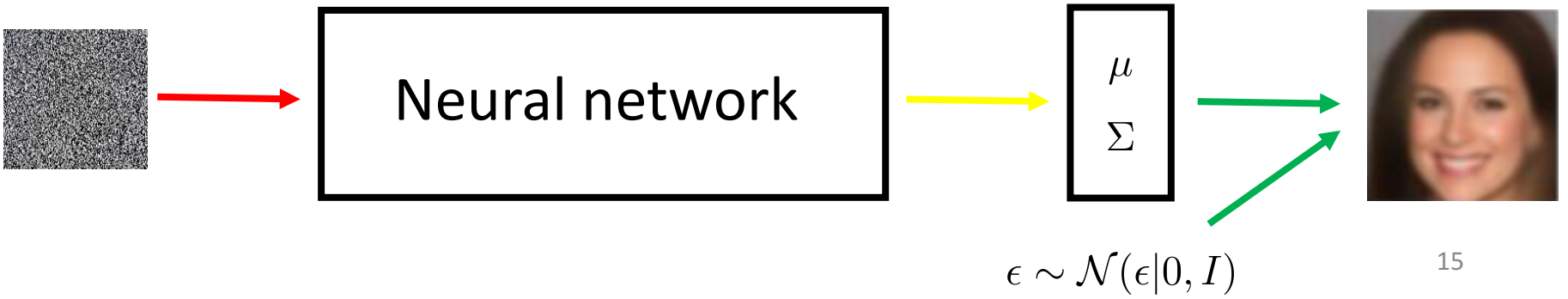


Semi-implicit model

- Most known example: VAE
- We model variational approximation as a **result of marginalization** w.r.t. auxiliary variables

$$q(\theta) = \int q_\phi(\theta|z)q(z)dz$$

- We have an access to both $q(\theta|z)$ and $q(z)$
- Prior $q(z)$ is fixed and has very simple structure
- Parameters of conditional $q_\phi(\theta|z)$ are modelled by DNN with weights ϕ



VI with implicit distributions

Usual ELBO estimation is intractable

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\theta)} \log p(D_{tr}|\theta) - \mathbb{E}_{q_\phi(\theta)} \log \frac{q_\phi(\theta)}{p(\theta)}$$

Implicit distributions:

- GAN-like density ratio estimation
- Spectral gradient estimation
- Different VI objectives (OPVI, ...)
- ...

Semi-implicit only:

- Hierarchical VI
- Semi-implicit VI
- Unbiased gradients with UIVI

Way 1: Hierarchical Variational Bayes

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\theta)} \log p(D_{tr}|\theta) + \mathbb{E}_{q_\phi(\theta)} \log p(\theta) - \mathbb{E}_{q_\phi(\theta)} \log q_\phi(\theta)$$

- Upper bounding of $\mathbb{E}_{q_\phi(\theta)} \log q_\phi(\theta)$

$$\leq \mathbb{E}_{q_\phi(\theta)}[\log q_\phi(\theta) + KL(q_\phi(z\,|\,\theta) \,\|\, s_\xi(z\,|\,\theta))] = \mathbb{E}_{q_\phi(z,\theta)} \log \frac{q_\phi(\theta,z)}{s_\xi(z\,|\,\theta)} = \mathcal{L}^{aux}_{\mathcal{H}}$$

- Lower bounding of $\mathbb{E}_{q_\phi(\theta)} \log p(\theta)$

$$\geq \mathbb{E}_{q_\phi(\theta)}[\log p(\theta) - KL(r_\psi(z\,|\,\theta) \,\|\, q_\phi(z\,|\,\theta))] = \mathbb{E}_{q_\phi(\theta) r_\psi(z\,|\,\theta)} \log \frac{p(\theta,z)}{r_\psi(z\,|\,\theta)} = \mathcal{L}^{aux}_{CE}$$

- A new lower bound

$$\mathcal{L}(\phi) \geq \mathbb{E}_{q_\phi(\theta)} \log p(D_{tr}|\theta) + \mathcal{L}^{aux}_{CE} - \mathcal{L}^{aux}_{\mathcal{H}} = \underline{\mathcal{L}}(\phi, \xi, \psi)$$

(Salimans15) T. Salimans, D. Kingma, M. Welling. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. ICML, 2015
(Ranganath16) R. Ranganath, D. Tran, D. Blei. Hierarchical Variational Models. ICML, 2016

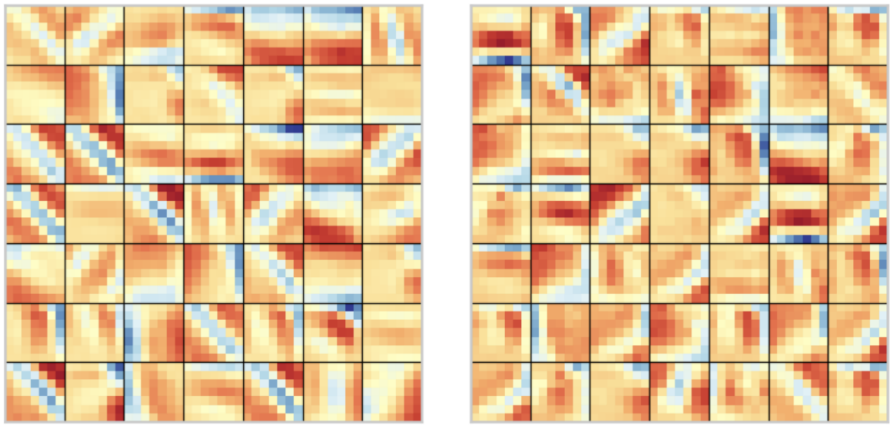
Deep Weight Prior


Andrei Atanov*

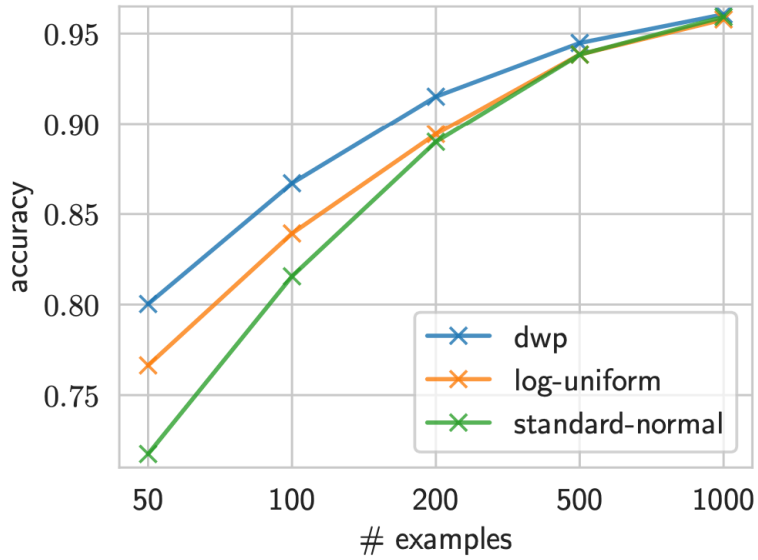

Arsenii Ashukha*


Kirill Struminsky

- Train VAE on convolution kernels from several pre-trained CNNs
- Can exploit knowledge from previous problems for new CNNs trained on different problems with similar but smaller data
- Prior is invariant to the changes in CNN size


Learned filters

Samples from DWP



(Atanov18) A. Atanov, A. Ashukha, K. Struminsky, D. Vetrov, M. Welling. Deep Weight Prior, 2018, https://arxiv.org/abs/1810.06943

Way 2: K-sample estimate

Usage of K-sample estimates (Yin18)

$$q_\phi(\theta) = \int q_\phi(\theta|z)q(z)dz \approx \frac{1}{K}\sum_{k=1}^{K} q_\phi(\theta|z_k), \quad z_k \sim q(z)$$

We want to bound ELBO:

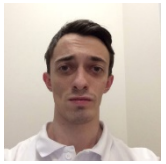
$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\theta)} \log p(D_{tr}|\theta) + \underbrace{\mathbb{E}_{q_\phi(\theta)} \log p(\theta)}_{\geq ?} - \underbrace{\mathbb{E}_{q_\phi(\theta)} \log q_\phi(\theta)}_{\leq ?}$$

(Yin18) M. Yin, M. Zhou. Semi-Implicit Variational Inference. ICML 2018.

Way 2: K-sample estimate


Dmitry Molchanov


Valery Kharitonov


Artem Sobolev

Usage of K-sample estimates (Yin18)

$$q_\phi(\theta) = \int q_\phi(\theta|z)q(z)dz \approx \frac{1}{K}\sum_{k=1}^{K} q_\phi(\theta|z_k), \quad z_k \sim q(z)$$

We want to bound ELBO:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\theta)} \log p(D_{tr}|\theta) + \underbrace{\mathbb{E}_{q_\phi(\theta)} \log p(\theta)}_{\geq ?} - \underbrace{\mathbb{E}_{q_\phi(\theta)} \log q_\phi(\theta)}_{\leq ?}$$

Bounds on the expected log-densities: (Molchanov18)

$$\mathbb{E}_{q_\phi(\theta)} \log q_\phi(\theta) \leq \mathbb{E}_{q(z_0)\ldots q(z_K) q_\phi(\theta|z_0)} \log \left(\frac{1}{K+1} \sum_{k=0}^{K} q_\phi(\theta|z_k) \right)$$

$$\mathbb{E}_{q_\phi(\theta)} \log p(\theta) \geq \mathbb{E}_{p(z_1)\ldots p(z_K) q_\phi(\theta)} \log \left(\frac{1}{K} \sum_{k=1}^{K} p(\theta|z_k) \right)$$

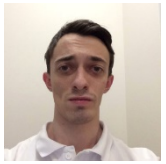
Asymptotically exact for $K \to +\infty$

(Yin18) M. Yin, M. Zhou. Semi-Implicit Variational Inference. ICML 2018.
(Molchanov18) D. Molchanov et al. Doubly Semi-Implicit Variational Inference, 2019, https://arxiv.org/abs/1810.02789

Way 2: K-sample estimate


Dmitry Molchanov


Valery Kharitonov


Artem Sobolev

Method	LL
VAE+VampPrior-data	-85.05
VAE+VampPrior	-82.38
VAE+DSIVI-prior (K=2000)	≥ -82.27
VAE+DSIVI-agg (K=500)	≥ -83.02
VAE+DSIVI-agg (K=5000)	$\geq \mathbf{-82.16}$
HVAE+VampPrior-data	-81.71
HVAE+VampPrior	-81.24
HVAE+DSIVI-agg (K=5000)	$\geq \mathbf{-81.09}$

Test log-likelihood for VampPrior vs semi-implicit priors (DSIVI-*)

(Molchanov18) D. Molchanov et al. Doubly Semi-Implicit Variational Inference, 2019, https://arxiv.org/abs/1810.02789

Way 3: Unbiased stochastic gradient

Make use of reparameterization trick and the fact (Titsias18)

$$\frac{\partial}{\partial\theta}\log q_\phi(\theta) = \int q_\phi(z|\theta)\frac{\partial}{\partial\theta}\log q_\phi(\theta|z)dz$$

Then it is possible to rewrite the gradient of ELBO

$$-\frac{\partial}{\partial\phi}\int q_\phi(\theta)\log q_\phi(\theta)d\theta \approx \int q(z)q_\phi(\theta|z)\frac{\partial\theta}{\partial\phi}\left(\int q(z'|\theta)\frac{\partial}{\partial\theta}\log q_\phi(\theta|z')dz'\right)d\theta dz$$

(Titsias18) M. Titsias, F. Ruiz. Unbiased Implicit Variational Inference. https://arxiv.org/abs/1808.02078

Way 3: Unbiased stochastic gradient

Make use of reparameterization trick and the fact (Titsias18)

$$\frac{\partial}{\partial \theta} \log q_\phi(\theta) = \int q_\phi(z|\theta) \frac{\partial}{\partial \theta} \log q_\phi(\theta|z) dz$$

Then it is possible to rewrite the gradient of ELBO

$$-\frac{\partial}{\partial \phi} \int q_\phi(\theta) \log q_\phi(\theta) d\theta \approx$$

$$\int q(z) q_\phi(\theta|z) \frac{\partial \theta}{\partial \phi} \left(\int \boxed{q(z'|\theta)} \frac{\partial}{\partial \theta} \log q_\phi(\theta|z') dz' \right) d\theta dz$$

Need to use MCMC to generate samples

(Titsias18) M. Titsias, F. Ruiz. Unbiased Implicit Variational Inference. https://arxiv.org/abs/1808.02078

Key feature of semi-implicitness

- Any implicit model can be re-written in semi-implicit terms as follows

$$q_\phi(\theta) = \int q_\phi(\theta|z)q(z)dz,$$

  where $q_\phi(\theta|z) = \delta(\theta - g_\phi(z))$

- Key feature is our ability to compute joint densities and derivatives of conditional

$$q_\phi(\theta, z) = q_\phi(\theta|z)q(z), \qquad \frac{\partial}{\partial\phi} q_\phi(\theta|z)$$

- We should also be able to reparameterize θ via differentiable function $g(.)$

$$\theta = g(z, \phi, \epsilon)$$

Main inference techniques

	MCMC	(S)IPM	Variational Bayes
Bias	**No**	**Weak**	**Strong**
Ensembling	**Inefficient**	**Efficient**	**Efficient**
Density	**No**	**No**	**Yes**
Empirical Bayes	**Not available**	**Flexible**	**Rough**

Deep | Bayes

- Summer school on Bayesian Deep Learning, August, 2019, Moscow
- The application deadline is April 30th, more info at http://deepbayes.ru