

# Sampling and prior selection

2023

# Model selection: coherent inference

*First level:* select optimal parameters:

$$w = \arg \max \frac{p(\mathcal{D}|w)p(w|h)}{p(\mathcal{D}|h)},$$

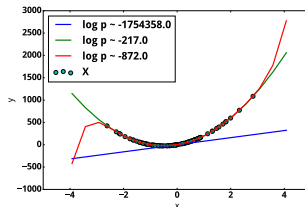
*Second level:* select optimal model (hyperparameters).

Evidence:

$$p(\mathcal{D}|h) = \int_w p(\mathcal{D}|w)p(w|h)dw.$$



Model selection scheme



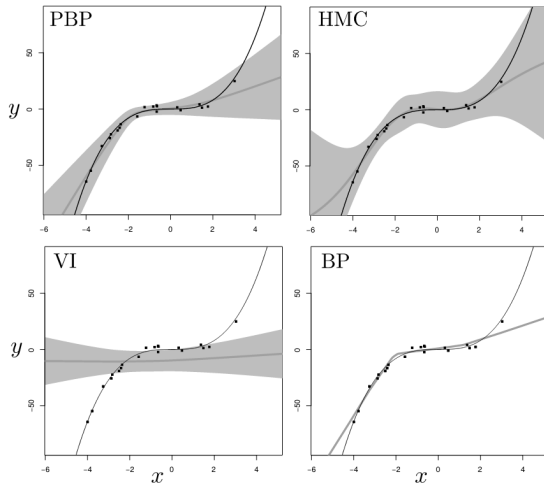
Example: polynomials

# Evidence estimation

$$Ef = \int_{\mathbf{w}} f(\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

- Laplace approximation
  - ▶ Fixed form of approximation distribution
  - ▶ Poorly scales
- Variational inference
  - ▶ Well scales
  - ▶ Can use different forms of approximation distributions
  - ▶ Lower bound of evidence  $\Rightarrow$  biased
- MC
  - ▶ Can use different forms of approximation distributions
  - ▶ Approximates well
  - ▶ Slow

# VI vs MC



# Naive method

$$I = \mathbb{E}f = \int_{\mathbf{w}} f(\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

Approximate:

$$\hat{I} = \frac{1}{N} \sum_{\mathbf{w} \sim p(\mathbf{w})} f(\mathbf{w}).$$

# Properties

Integral estimation:

- strongly consistent :  $\hat{I} \xrightarrow{\text{a.s.}} I$
- Unbiased:  $E\hat{I} = I$
- Asymptotically normal;
- $D\hat{I} = O(\frac{1}{N})$ .
- **Challenge:** we need to sample from  $p$ .

Why this does not work?

# Inverse transform sampling

Let  $T$  be an invertible function from  $u \sim \mathcal{U}(0, 1)$  to some random variable distribution  $p(w)$ .  
Then

$$F_w(t) = p(w \leq t) = p(T(u) \leq t) = p(u \leq T^{-1}(t)) = T^{-1}(u).$$

Therefore  $F_u^{-1} = T$ .

**Example**

$$w = \lambda \exp(-\lambda t).$$

$$F_w(t) = 1 - \exp(-\lambda t).$$

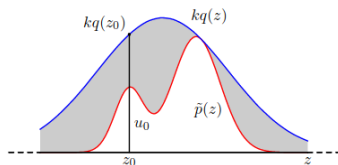
$$F_w^{-1}(t') = -1 \frac{1}{\lambda} \log(1 - t').$$



# Rejection sampling

- Given  $p(w)$  (up to normalizing constant)
- Set distribution  $q$
- Set value  $k$  so that  $kq(w) \geq p(z)$  for all  $z$
- In a loop:
  - ▶ Sample  $w_0 \sim q$
  - ▶ Sample  $u \sim \mathcal{U}(0, kq(w_0))$
  - ▶ If  $u \leq p(w_0)$ , use it as a sample from  $p(w)$

**Core idea:** sample  $u$  are uniform in a region limited by  $p(w)$ .



Bishop, 2006

# Importance sampling

Consider the case when we cannot sample from  $p(w)$ , but we can estimate likelihood and want to estimate the integral

$$Ef = \int f(w)p(w)dw.$$

Let  $q$  be an auxiliary distribution:

$$Ef = \int f(w)p(w)dw = \int f(w)\frac{p(w)}{q(w)}dz \approx \frac{1}{L} \sum_{l=1}^L \frac{p(w^l)}{q(w^l)} f(w^l).$$

# MCMC

**Basic idea:** Sample similar to rejection sampling, but  $q$  is a Markov distribution with conditioning on the previous step.

We want the stationary (limiting) distribution to be equal to our  $p(w)$ .

Sufficient condition

$$p(w) T(w|w') = p(w') T(w'|w).$$

# Metropolis-Hastings algorithm

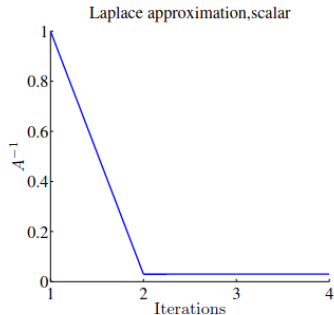
- Sample new  $w' \sim q(w|w^t)$ .
- Accept with probability  $A(w'|w^t) = \min\left(1, \frac{p(w')q(w^t|w')}{p(w^t)q(w'|w^t)}\right)$ .
- If accepted:  $w^{t+1} = w'$ ,
- Otherwise:  $w^{t+1} = w^t$ .

Sufficient condition is satisfied::

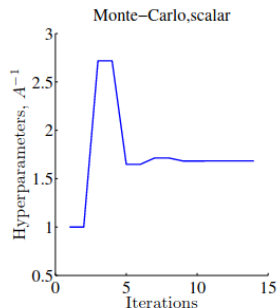
$$\begin{aligned} p(w)T(w|w') &= p(w)T(w'|w) = p(w')T(w'|w^t) = p(w')q(w'|w^t)A(w'|w^t) = \\ &= p(w^t)q(w^t|w')A(w^t|w'). \end{aligned}$$

- Samples are correlated. We can decorrelate sample using each  $k$  sample.
- Works better in high-dimensional settings than rejection sampling.
- Good choice of  $q$  is the main challenge for the algorithm.

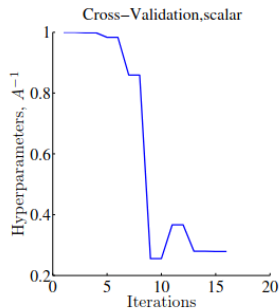
# Hyperparameter selection for linear model



(a) Laplace approximation

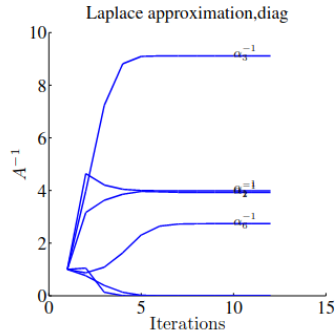


(b) Monte-Carlo

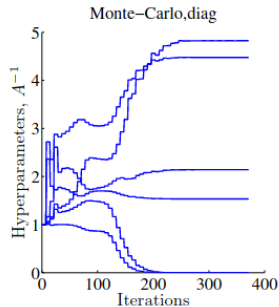


(c) Cross validation

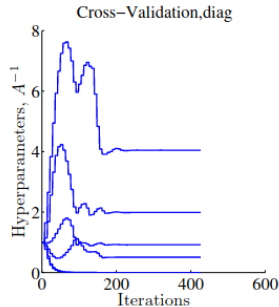
# Hyperparameter selection for linear model



(a) Laplace approximation

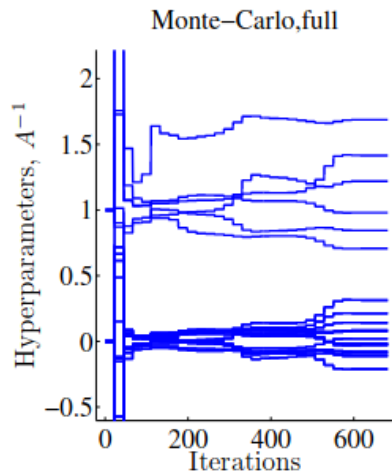


(b) Monte-Carlo

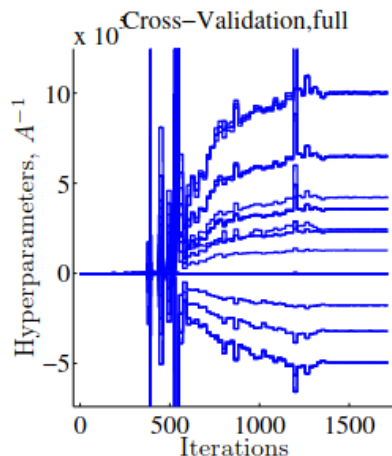


(c) Cross validation

# Hyperparameter selection for linear model



(a) Monte-Carlo



(b) Cross validation

# Stochastic gradient Langevin dynamics

A modification of SGD:

$$T = w - \beta \nabla L + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\beta}{2})$$

where  $\beta$  changes with a number of iterations:

$$\sum_{\tau=1}^{\infty} \beta_{\tau} = \infty, \quad \sum_{\tau=1}^{\infty} \beta_{\tau}^2 < \infty.$$

**Statement [Welling, 2011].** Distribution  $q^{\tau}(w)$  converges to posterior distribution  $p(w|X, f)$ .  
Entropy adjustment:

$$\hat{S}(q^{\tau}(w)) \geq \frac{1}{2} |w| \log \left( \exp \left( \frac{2S(q^{\tau}(w))}{|w|} \right) + \exp \left( \frac{2S(\epsilon)}{|w|} \right) \right).$$

**Special case of Metropolis-Hastings**



# Precondition-matrix for SGLD

SGLD:

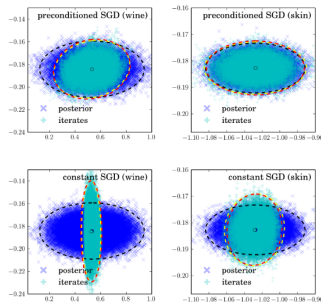
$$T = w - \beta \nabla L + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2})$$

pSGLD:

$$T = w - \beta M \nabla L + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2} M),$$

matrix  $M$  is optimized to make gradient step uniform for each direction (w.r.t. gradient variance).

*Example for SGD, for SGLD the results are similar.*



# Gibbs sampling

Given a graphical model  $w_1, \dots, w_n$ .

Then in a loop over variables:

$$\hat{w}_i \sim p(w_i | w_1, w_{i-1}, w_{i+1}, w_n), w_i := \hat{w}_i$$

Gibbs sampling is also a special case of MH.

# Contrastive Divergence: idea

Energy-based model:

$$p(x|w) = \frac{\exp(-E_w(x))}{Z(w)}, \quad Z = \int_x \exp(-E_w(x)),$$

$$\frac{\partial \log p(x|w)}{\partial w} = \mathbb{E}_{x' \sim p(x|w)} \frac{\partial E(x')}{\partial w} - \frac{\partial E(x)}{\partial w}$$

Algorithm for RBM:

- Take  $x$  from the dataset
- $h_0 \sim p(h_0|x)$
- $x_1 \sim p(x|h_0)$
- ...
- Obtain  $x_k$
- $\frac{\partial \log p(x|w)}{\partial w} = \frac{\partial E(x_k)}{\partial w} - \frac{\partial E(x)}{\partial w}$

# Autoencoder: generative model?

(Alain, Bengio 2012): consider regularized autoencoder:

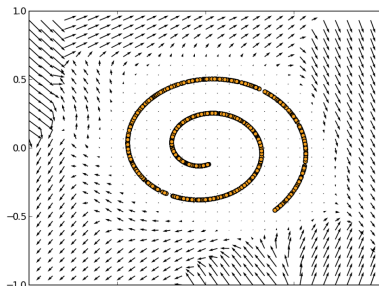
$$\|f(x, \sigma) - x\|^2,$$

where  $\sigma$  is a noise level.

Then

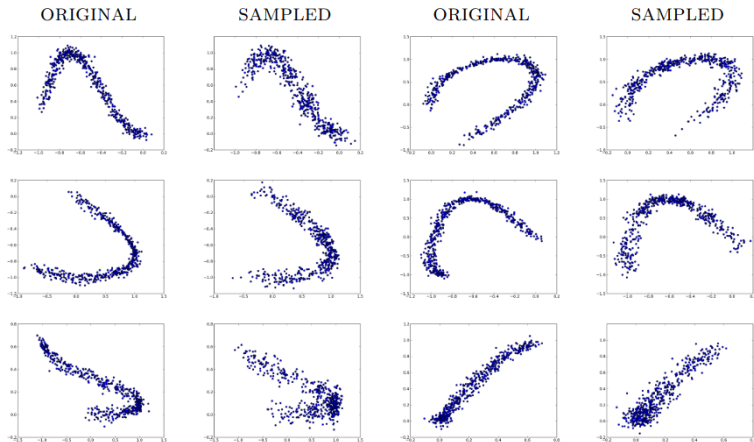
$$\frac{\partial \log p(x)}{\partial x} = \frac{\|f(x, \sigma) - x\|^2}{\sigma^2} + o(1) \text{ при } \sigma \rightarrow 0.$$

Vector field induced by reconstruction error



# Autoencoder for sampling

$$A = \frac{p(x^*)}{p(x)} = \exp(E(x) - E(x^*)) \approx \frac{\partial E(x)^T}{\partial x} (x^* - x) + o(\|x - x^*\|).$$



# Optimization of $q$

Distribution  $q$  can be set using neural networks.

- **Main requirements:** existence of  $p(x|x')$ ,  $p(x'|x) \rightarrow$  the distribution must be invertible.
- Neural network in a form of  $f(x, w) = x + g(x, w)$  is a flow and invertible.

## Optimization variants:

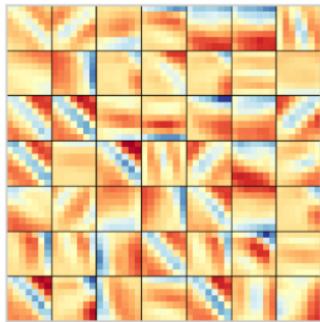
- Entropy \* Acceptance rate (Li et al., 2020)
- GAN between empirical distribution and  $q$  (Song et al., 2017).

# Informative prior vs Uninformative prior

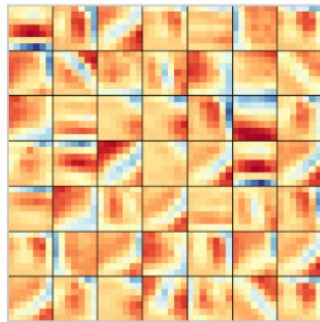
- Informative prior: corresponds to some expert knowledge
  - ▶ Example: air temperature in some region: Gaussian variable with known mean and variance estimated from previous observations.
  - ▶ Mistake in informative prior estimation leads to poor models.
- Uninformative prior: corresponds to some basic knowledge
  - ▶ Example: air temperature in some region: uniform improper prior.
- Weakly-informative prior: somewhere in between
  - ▶ Example: air temperature in some region: uniform distribution in  $[-50, 50]$ .

**What if our prior and posterior are very close?**

## The deep weight prior: Atanov et al., 2019



(b) Learned filters

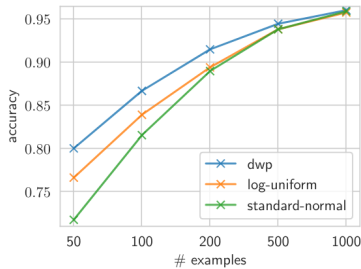


(c) Samples from DWP

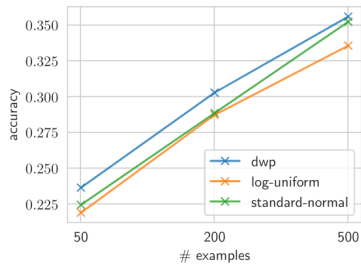
The distribution can be modeled by complex models and can generate rather informative samples!



# The deep weight prior: Atanov et al., 2019

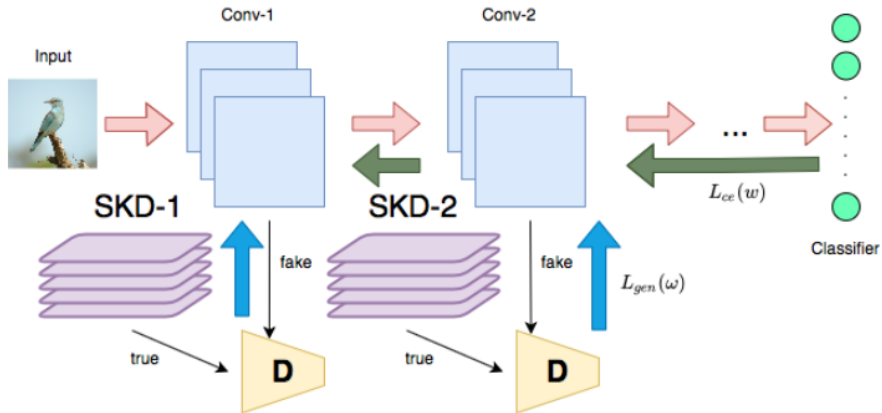


(a) Results for MNIST

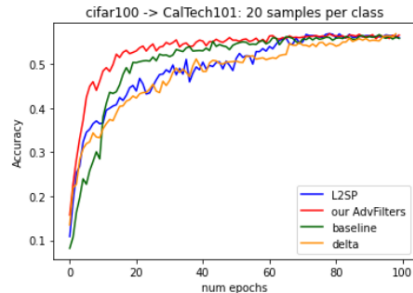
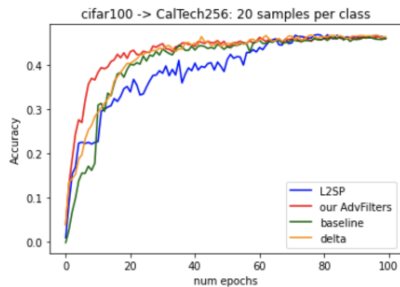


(b) Results for CIFAR-10

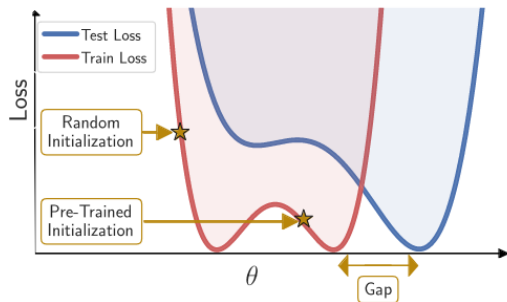
# Deep weight prior for distillation, Kolesov 2022



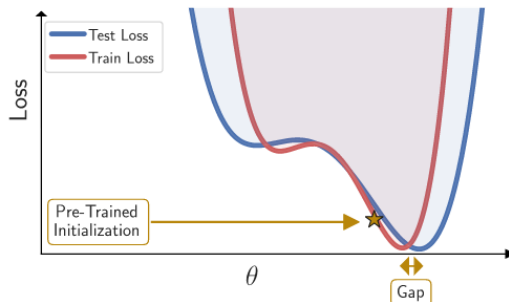
# Deep weight prior for distillation, Kolesov 2022



## Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors, 2022



(a) Standard Transfer Learning



(b) Transfer Learning with Learned Priors

## Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors, 2022

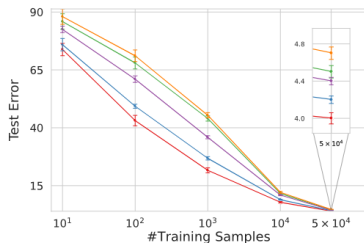
- Decompose model into feature extractor and classifier
- Learn feature extractor's posterior on the source task using SWAG
- Consider target prior to be similar to source posterior:

$$p(w) = \mathcal{N}(\mu, hA^{-1}),$$

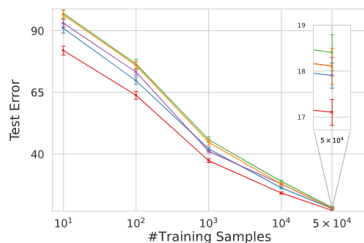
where  $h \in \mathbb{R}$  is calibrated on the target task.

- Learn prior for the classifier on the target task

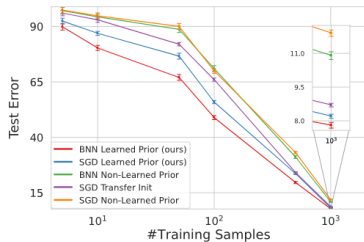
# Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors, 2022



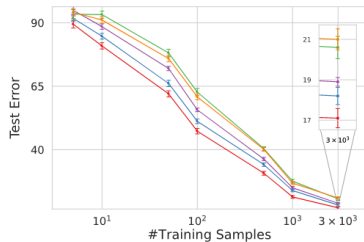
(a) CIFAR-10



(b) CIFAR-100



(c) Oxford Flowers-102



(d) Oxford-IIIT Pets

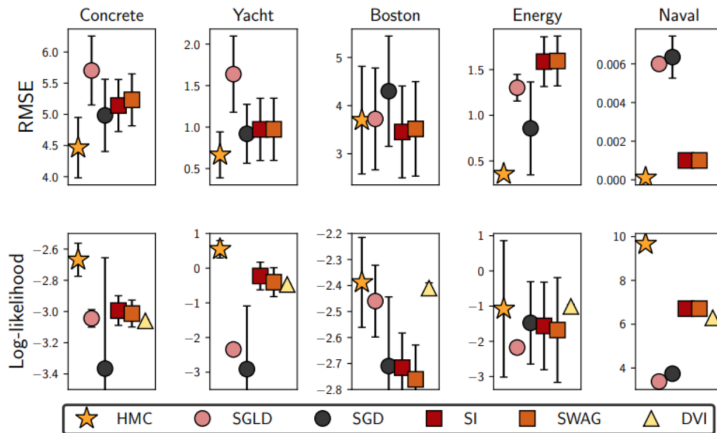
# What Are Bayesian Neural Network Posteriors Really Like?

Izmailov et al., 2021:

- HMC for posterior distribution estimation for deep models on some standard datasets.
- Resources: 512 TPU

# What Are Bayesian Neural Network Posteriors Really Like?

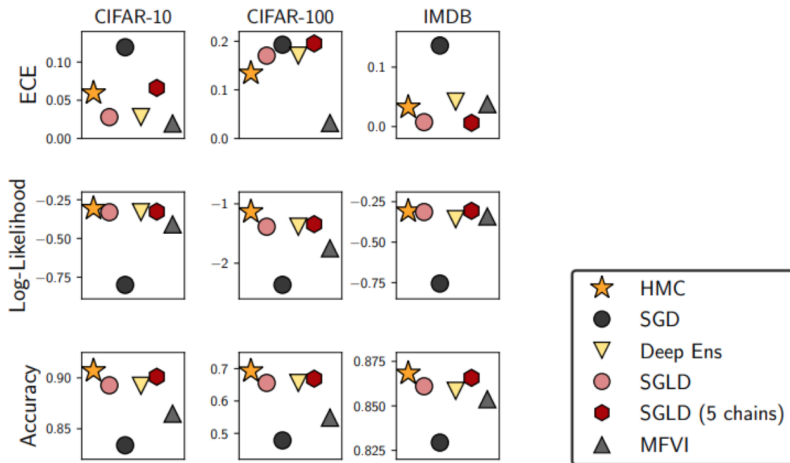
## BNN evaluation: UCI





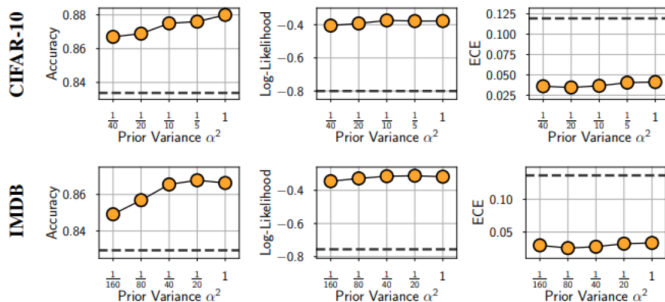
# What Are Bayesian Neural Network Posteriors Really Like?

## BNN evaluation: CIFAR and IMDB



# What Are Bayesian Neural Network Posteriors Really Like?

## Effect of priors



HMC BNNs are fairly robust to Gaussian prior variance.

# References

- Bishop C. M., Nasrabadi N. M. Pattern recognition and machine learning. – New York : springer, 2006. – T. 4. – №. 4. – C. 738.
- Kuznetsov M., Tokmakova A., Strijov V. Analytic and stochastic methods of structure parameter estimation //Informatica. – 2016. – T. 27. – №. 3. – C. 607-624.
- Mandt, Stephan, Matthew Hoffman, and David Blei. "A variational analysis of stochastic gradient algorithms." International conference on machine learning. PMLR, 2016.
- Alain, Guillaume, and Yoshua Bengio. "What regularized auto-encoders learn from the data-generating distribution." The Journal of Machine Learning Research 15.1 (2014): 3563-3593.
- Li Z., Chen Y., Sommer F. T. A neural network mcmc sampler that maximizes proposal entropy //arXiv preprint arXiv:2010.03587. – 2020.
- Song J., Zhao S., Ermon S. A-nice-mc: Adversarial training for mcmc //Advances in Neural Information Processing Systems. – 2017. – T. 30.
- Atanov, Andrei, et al. "The deep weight prior." arXiv preprint arXiv:1810.06943 (2018).
- Kolesov A. An adversarial method for neural network fine-tuning for transfer learning problem, Master thesis, 2022.
- Shwartz-Ziv, Ravid, et al. "Pre-train your loss: Easy bayesian transfer learning with informative priors." Advances in Neural Information Processing Systems 35 (2022): 27706-27715.
- Izmailov, Pavel, et al. "What are Bayesian neural network posteriors really like?." International conference on machine learning. PMLR, 2021.