Good Initializations of Variational Bayes for Deep Models
Simone Rossi, Pietro Michiardi, Maurizio Filippone

**Objective and Contributions**

**Initialization** of variational parameters has a huge role in the convergence of stochastic variational inference but received little to no attention in current literature.

Contributions:
- New initialization for svi based on Bayesian linear models;
- Applied to regression, classification and CNNs;
- Experimental comparison against other initializations;
- SoTA performance with Gaussian svi on large-scale CNNs.

**Stochastic Variational Inference - svi**

A DNN is a composition of nonlinear vector-valued functions $f^{(l)}$

$$f(x) = (f^{(L-1)}(W^{(L-1)}x + b^{(L-1)}), \ldots, f^{(1)}(W^{(1)}x + b^{(1)}) \} \{ x \}$$

Objective of Bayesian inference

$$p(W|X,Y) = \frac{p(Y|X,W)p(W)}{p(Y|X)}$$

svi reformulates this problem as minimization of the negative evidence lower bound (or ELBO) under an approximate distribution $q_{\theta}(W)$ [2]:

$$q_{\theta}(W) \text{ s.t. } \theta = \arg \min_{\theta} \text{ELBO}$$

$$\text{ELBO} = E_{q_{\theta}}[\log p(Y|X,W)] + KL(q_{\theta}(W)||p(W))$$

During training, we approximate the posterior $p(W|X,Y)$ with a family of variational distributions $q_{\theta}(W)$, where $\theta = \{ \mu^{(l)}_i, \sigma^{(l)}_i \} : l = 0, \ldots, L - 1$.

How do we initialize $\theta$?

**Iterative Bayesian Linear Modeling Initiator - I-BLM**

- Inspired by residual networks and greedy initialization of DNNs.
- Grounded on Bayesian Linear regression but extended to classification and to convolutional layers.
- Regression on transformed labels obtained through the interpretation of classification labels as the coefficients of a degenerate Dirichlet distribution.
- Scalability achieved thanks to mini-batching.

But how does it work?

Transform the labels if it's a classification task [3]. For each layer [1]:
- Propagate a mini-batch of $X$ up to the previous layer $(1 - l)$;
- Extract the patches if it's a convolutional layer;
- Learn a Bayesian linear model and use its solution to initialize $q_{\theta}(W)$. 

Effect of batch-size: the full training set leads to a better estimate of the posteriors.

**References**


Checkout the Full Paper!


Some more insights!

**Timing profiling** (LENET-5): before training, 4 out of 5 optimal initializers are I-BLM.

**Regression and Classification on Bayesian DNNs**

- Another initialization for Gaussian svi based on a MAP optimization (MAP INIT).
- Loss optimized for the same amount of time required by I-BLM. Solution used to initialize the means, while the log-variances are $-$5.5.
- Models are trained for 100 minutes for the entire end-to-end training (curves are shifted by the initialization time).

I-BLM for Bayesian CNNs - VGG16

- 1.5M+ params
- MNIST: average test error $= 0.035$ 
- VGG16: 3.5M+ params
- MNIST: average test error $= 0.526$ 

Figure & Table: Comparison between Gaussian factorized svi, MCD and NOISY-KFAC on VGG16 with CIFAR10.

**Figure**: Progression of test error and test MNLL with different initializations on a $5x10$ architecture.