Good Initializations of Variational Bayes for Deep Models

Simone Rossi, Pietro Michiardi, Maurizio Filippone

Objectives and Contributions

- Initializations 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 \( f^{(l)} \):

\[
f(x) = \left( f^{(l-1)}(W^{(l-1)}) \circ \ldots \circ f^{(0)}(W^{(0)}) \right)(x)
\]

Objective of Bayesian inference

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

How do we initialize \( \theta \)?

A poor initialization can prevent SVI from converging to good solutions even for simple problems. It is even more severe for complex architectures, where SVI systematically converges to trivial solutions.

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

Iterative Bayesian Linear Modeling Initializer - I-BLM

In a nutshell:

- Inspired by residual networks and greedy initialization of DNNs.
- Grounded on Bayesian linear regression but extended to classification and 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 \( l \):

- Propagate a mini-batch of \( X \) up to the previous layer \((l - 1)\);
- Extract the patches if it’s a convolutional layer;
- Learn a Bayesian linear model and use its solution to initialize \( q_k(W^{(l)}) \).

Bayesian Linear Regression - BLR

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

More some insights!

Timing profiling (LENET-5): before training, 3 out of 4 optimal initializers are I-BLM.

Regression and Classification on Bayesian DNNs

I-BLM for Bayesian CNNs - VGG16

- 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).

Some more insights!

Timing profiling (LENET-5): before training, 3 out of 4 optimal initializers are I-BLM.

Figure: Representation of I-BLM. On (top) we learn two Bayesian linear models, whose outputs are used on the (bottom) for the following layer.

Figure: Progression of test error and test MNLL with different initializations on a 5x100 architecture.

Figure: Entropy distribution while testing on MNIST and NOT-MNIST (higher average entropy on NOT-MNIST means better uncertainty estimation).

Check out the Full Paper!


References