Good Initializations of Variational Bayes for Deep Models

Simone Rossi, Pietro Michiardi and Maurizio Filippone
Department of Data Science, EURECOM

36th International Conference on Machine Learning
9-15 June 2019, Long Beach, USA

Bayesian Inference 101

Posterior over the weights
Intractable for DNNs

Model likelihood

Prior on model parameters

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

Marginal Likelihood

Bayesian Inference 101

Posterior over the weights
Intractable for DNNs

Model likelihood

Prior on model parameters

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

Marginal Likelihood

Variational Inference

SVI reformulates this problem as minimization of the **negative evidence lower bound** (or NELBO) under an approximate distribution $q_{\theta}(\mathbf{W})$:

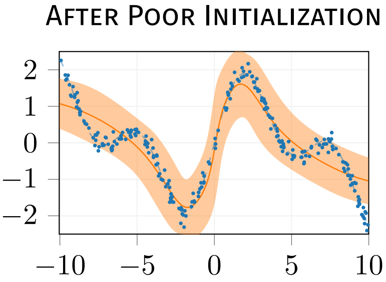
$$q_{\tilde{\theta}}(\mathbf{W}) \quad \text{s.t.} \quad \tilde{\theta} = \arg\min_{\theta}\{\text{NELBO}\}$$

$$\text{NELBO} = \mathbb{E}_{q_{\theta}}\left[-\log p(Y|X,\mathbf{W})\right] + \text{KL}\left(q_{\theta}(\mathbf{W})||p(\mathbf{W})\right)$$

Initialization of Variational Bayes? A Motivating Example

- VI struggles to scale on models with millions of parameters
- Initialization of VI has few mentions in current literature

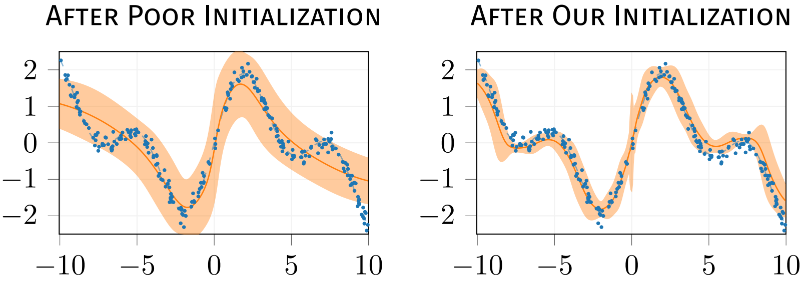
Initialization of θ matters



Initialization of Variational Bayes? A Motivating Example

- VI struggles to scale on models with millions of parameters
- Initialization of VI has few mentions in current literature
- **For the first time**: we propose a solution to this issue

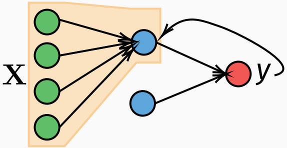
Initialization of θ matters

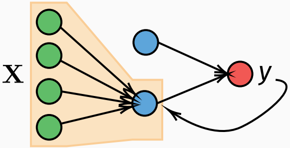


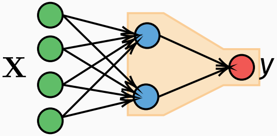
Iterative-Bayesian Linear Modeling: I-BLM

In a nutshell:

- Grounded on **Bayesian Linear regression** but extended to classification and to convolutional layers
- **Regression** with Gaussian likelihood **on transformed labels** for classification tasks
- **Scalability** achieved thanks to mini-batching







Experimental Evaluation - Bayesian Neural Networks

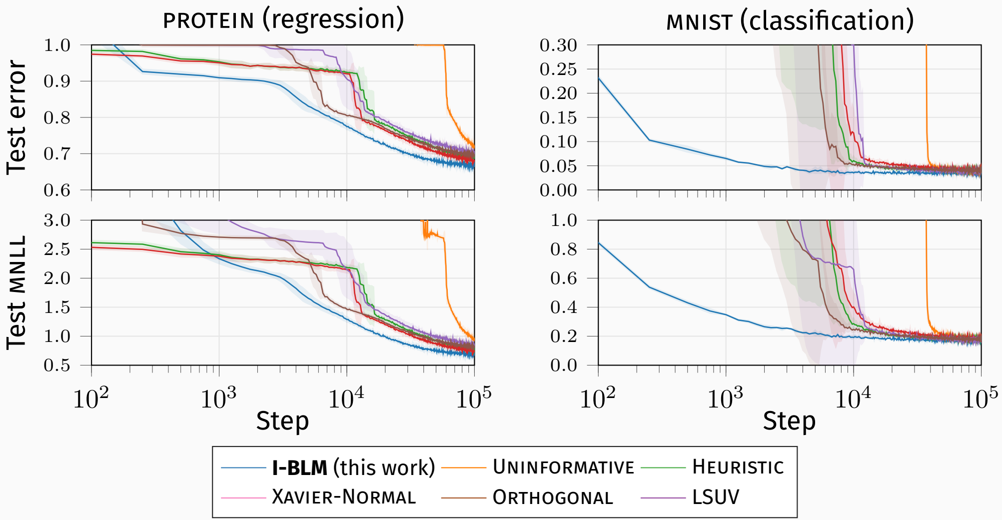
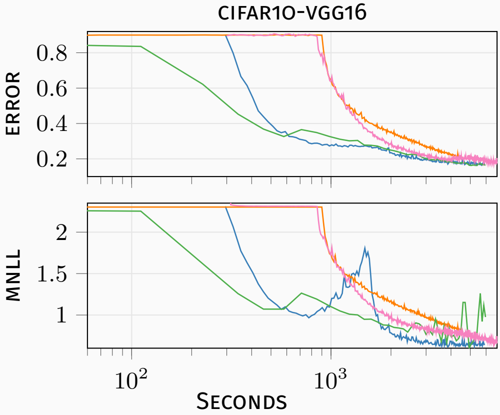


Figure: Test error and MNLL with different init. on a 5x100 BNN.

Experimental Evaluation - Bayesian CNN

- Another initialization for Gaussian SVI based on a MAP optimization.
- Models are trained for 100 minutes for the entire end-to-end training (curves are shifted by the initialization time).



VGG16: 3.5M+ parameters

	MNLL	ERROR
G-SVI & I-BLM	**0.637**	**0.167**
G-SVI & MAP	0.750	0.201
MCD	0.821	0.215
NOISY-KFAC	0.750*	**0.164**

Good Initializations of Variational Bayes for Deep Models

Simone Rossi [1] **Pietro Michiardi** [1] **Maurizio Filippone** [1]

Abstract

Stochastic variational inference is an established way to carry out approximate Bayesian inference for deep models flexibly and at scale. While there have been effective proposals for good initializations for loss minimization in deep learning, far less attention has been devoted to the issue of initialization of stochastic variational inference. We address this by proposing a novel layer-wise initialization strategy based on Bayesian linear models. The proposed method is extensively validated on regression and classification tasks, including Bayesian Deep Nets and Conv Nets, showing faster and better convergence compared to alternatives inspired by the literature on initializations for loss minimization.

Figure 1: Due to poor initialization (**left**) SVI fails to converge even after 600+ epochs (RMSE $= 0.613$, MNLL $= 29.4$) while with our I-BLM (**right**) SVI easily recovers the function after few epochs (RMSE $= 0.315$, MNLL $= -5.8$). The architecture has three hidden layers with 500 neurons each, and uses the TANH activation function.

1 Introduction

Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) have become the preferred choice to tackle various learning tasks, due to their ability to model complex problems and the mature development of regularization techniques to control overfitting (LeCun et al., 2015; Srivastava et al., 2014). There has been a recent surge of interest in the issues associated with their overconfidence in predictions, and proposals to mitigate these (Guo et al., 2017; Kendall & Gal, 2017; Lakshminarayanan et al., 2017). Bayesian techniques offer a natural framework to deal with such issues, but they are characterized by computational intractability (Bishop, 2006; Ghahramani, 2015).

A popular way to recover tractability is to use variational inference (Jordan et al., 1999). In variational inference, an approximate posterior distribution is introduced and its parameters are adapted by optimizing a variational objective, which is a lower bound to the marginal likelihood. The variational objective can be written as the sum of an expectation of the log-likelihood under the approximate posterior and a regularization term which is the negative Kullback-Leibler (KL) divergence between the approximating distribution and the prior over the parameters. Stochastic Variational Inference (SVI) offers a practical way to carry out stochastic optimization of the variational objective. In SVI, stochasticity is introduced with a doubly stochastic approximation of the expectation term, which is unbiasedly approximated using Monte Carlo and by selecting a subset of the training points (mini-batching) (Graves, 2011; Kingma & Welling, 2014).

While SVI is an attractive and practical way to perform approximate inference for DNNs, there are limitations. For example, the form of the approximating distribution can be too simple to accurately approximate complex posterior distributions (Ha et al., 2016; Ranganath et al., 2015; Rezende & Mohamed, 2015). Furthermore, SVI increases the number of optimization parameters compared to optimizing model parameters through, e.g., loss minimization; for example, a fully factorized Gaussian posterior over model parameters doubles the number of parameters in the optimization compared to loss minimization. This has motivated research on other ways to perform approximate Bayesian inference for DNNs by establishing connections between variational inference and dropout (Gal & Ghahramani, 2016a;b; Gal et al., 2017).

A theoretical understanding of the optimization landscape of DNNs and CNNs is still in its early stages of development (Dziugaite & Roy, 2017; Garipov et al., 2018), and most works have focused on the practical aspects charac-

[1] Department of Data Science, EURECOM, France. Correspondence to: Simone Rossi <simone.rossi@eurecom.fr>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Thanks!

Poster #83:
Pacific Ballroom
Today – 6:30 pm