Good Initializations of Variational Bayes for Deep Models

Simone Rossi, Pietro Michiardi, Maurizio Filippone

EURECOM Sophia Antipolis

Objectives 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 $\mathbf{f}^{(l)}$

$$\mathbf{f}(\mathbf{x}) = \left(\mathbf{f}^{(L-1)}(\mathbf{W}^{(L-1)}) \circ \ldots \circ \mathbf{f}^{(0)}(\mathbf{W}^{(0)})\right)(\mathbf{x})$$

Objective of Bayesian inference

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

Prior on model parameters

Posterior over the weights
Intractable for DNNs

Marginal Likelihood

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

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

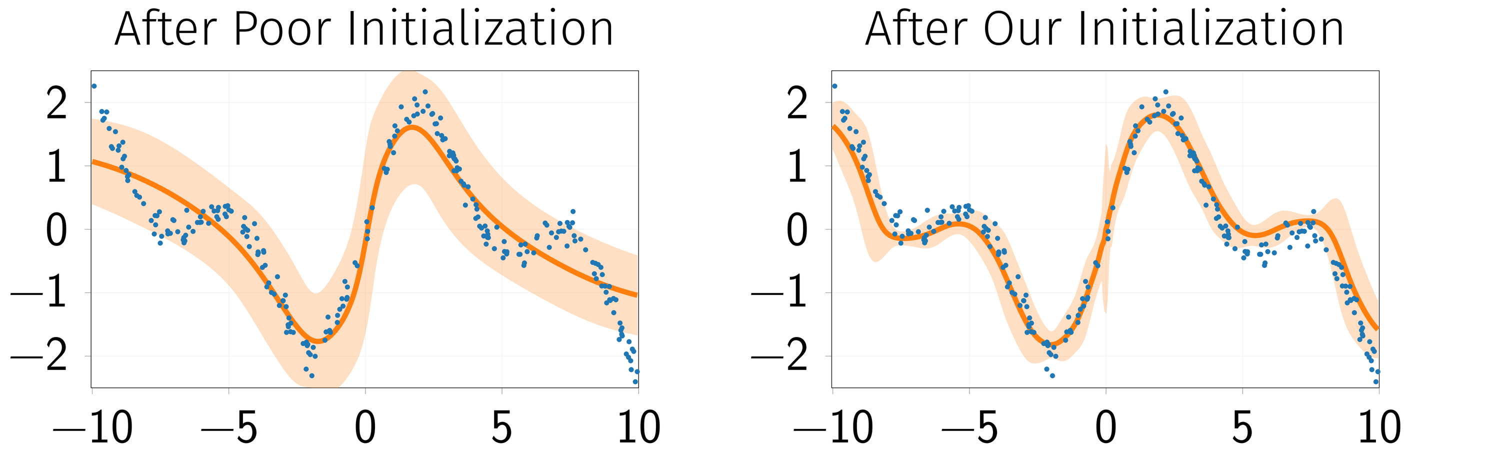
$$\text{NELBO} = \mathbb{E}_{q_\theta}\left[-\log p(Y|X, \mathbf{W})\right] + \text{KL}\left(q_\theta(\mathbf{W}) \| p(\mathbf{W})\right)$$

Commonly used family of variational distribution: **mean field Gaussian** (or fully factorized Gaussian)

$$q(\mathbf{W}^{(l)}) = \prod_{ij} \mathcal{N}(w_{ij}^{(l)} | \mu_{ij}^{(l)}, \sigma_{ij}^{(l)}) \qquad \theta = \{(\mu_{ij}^{(l)}, \sigma_{ij}^{(l)}) : l = 0, \ldots, L-1\}$$

How do we initialize θ?

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.



References

[1] Y. Gal and Z. Ghahramani. "Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference". *Workshop track - ICLR*. June 2015.

[2] A. Graves. "Practical Variational Inference for Neural Networks". *Advances in Neural Information Processing Systems 24*. 2011.

[3] D. Milios et al. "Dirichlet-based Gaussian Processes for Large-scale Calibrated Classification". *Advances in Neural Information Processing Systems 31*. 2018.

[4] G. Zhang et al. "Noisy Natural Gradient as Variational Inference". *Proceedings of the 35th International Conference on Machine Learning*. Oct. 2018.

Iterative Bayesian Linear Modeling Initializer - I-BLM

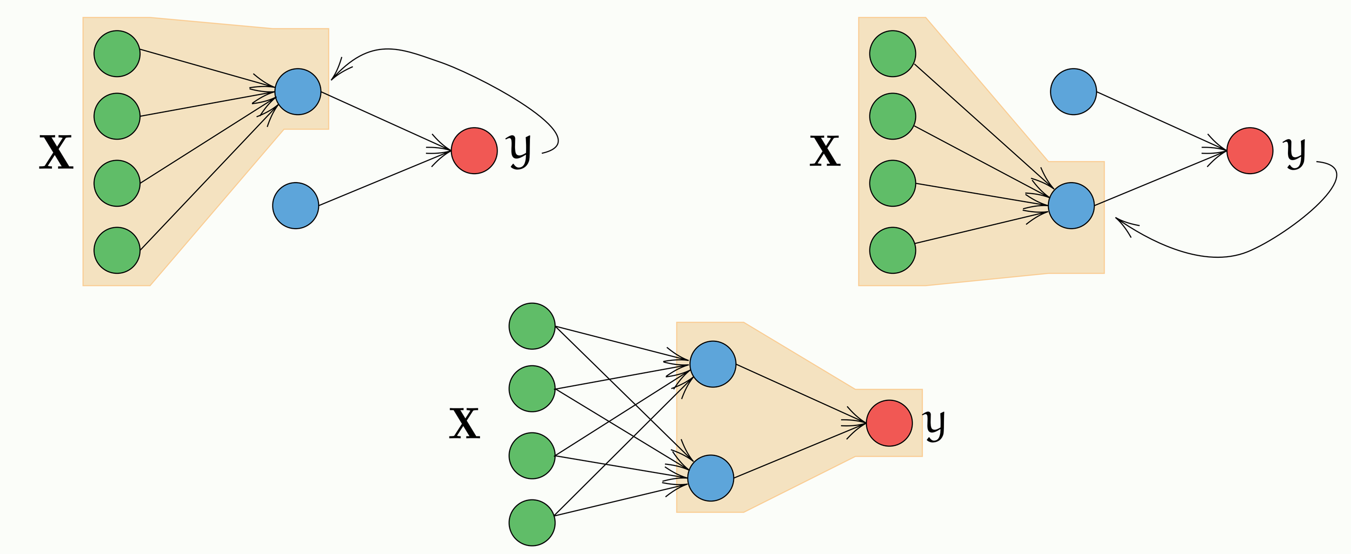


Figure: Representation of I-BLM. On (**top**) we learn two Bayesian linear models, whose outputs are used on the (**bottom**) for the following layer.

In a nutshell:

- 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 (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 $q_\theta(\mathbf{W}^{(l)})$.

Bayesian Linear Regression - BLR

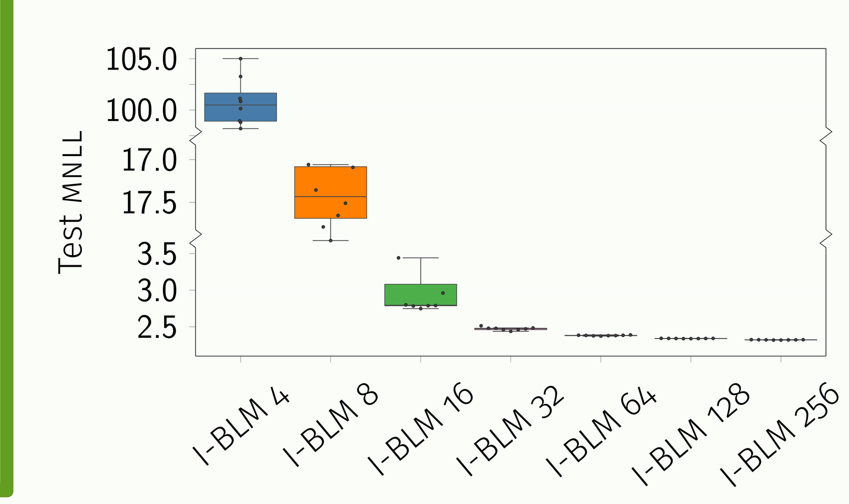
Likelihood:
$$p(Y|W, L) = \prod_i \mathcal{N}(Y_{\cdot i}|XW_{\cdot i}, L)$$

Prior:
$$p(W|\Lambda) = \prod_i p(W_{\cdot i}) = \mathcal{N}(W_{\cdot i}|\mathbf{0}, \Lambda)$$

Posterior:
$$p(W_{\cdot i}|Y, X, L, \Lambda) = \prod_i \mathcal{N}(W_{\cdot i}|\Sigma_i X^\top L^{-1} Y_{\cdot i}, \Sigma_i)$$

with $\Sigma_i = (\Lambda^{-1} + X^\top L^{-1} X)^{-1}$.

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



Some more insights!

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

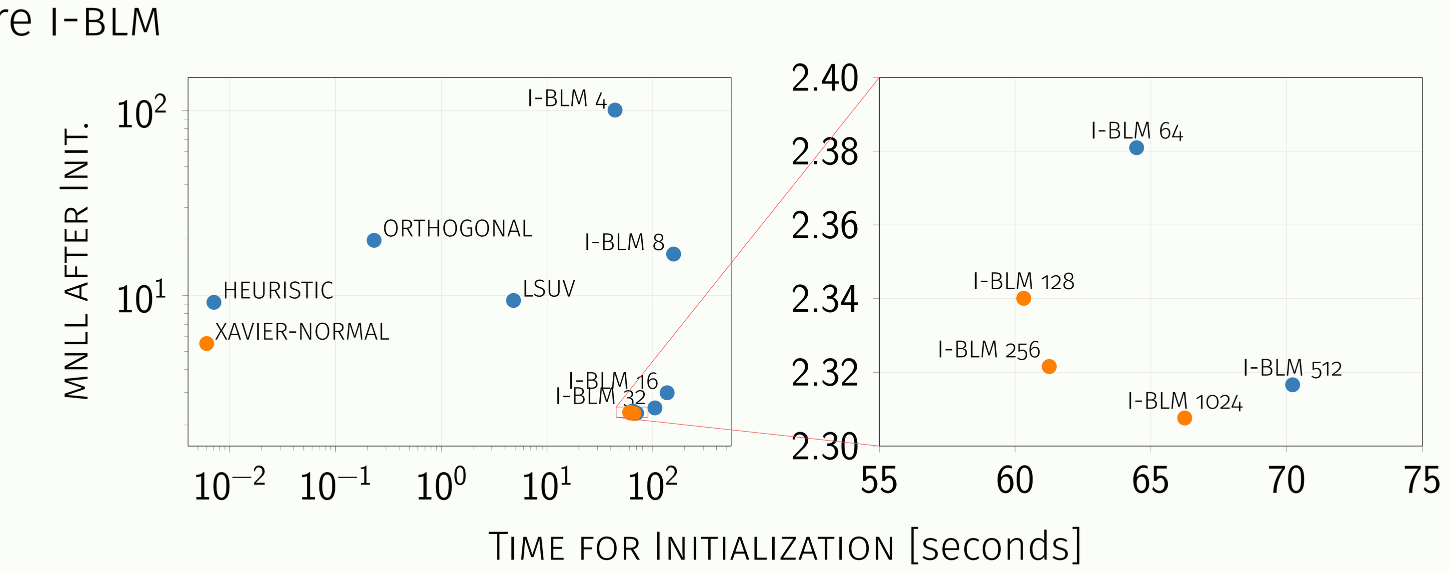


Figure: Comparison of initialization time versus test MNLL.

Regression and Classification on Bayesian DNNs

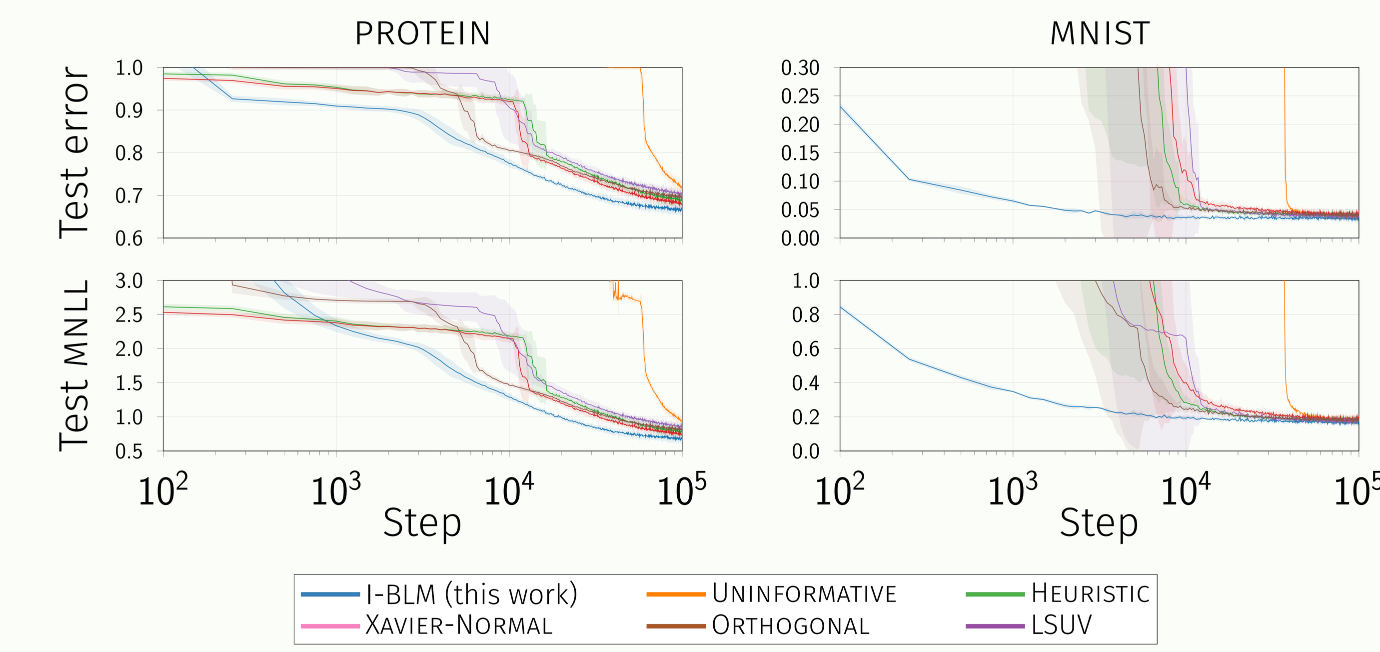
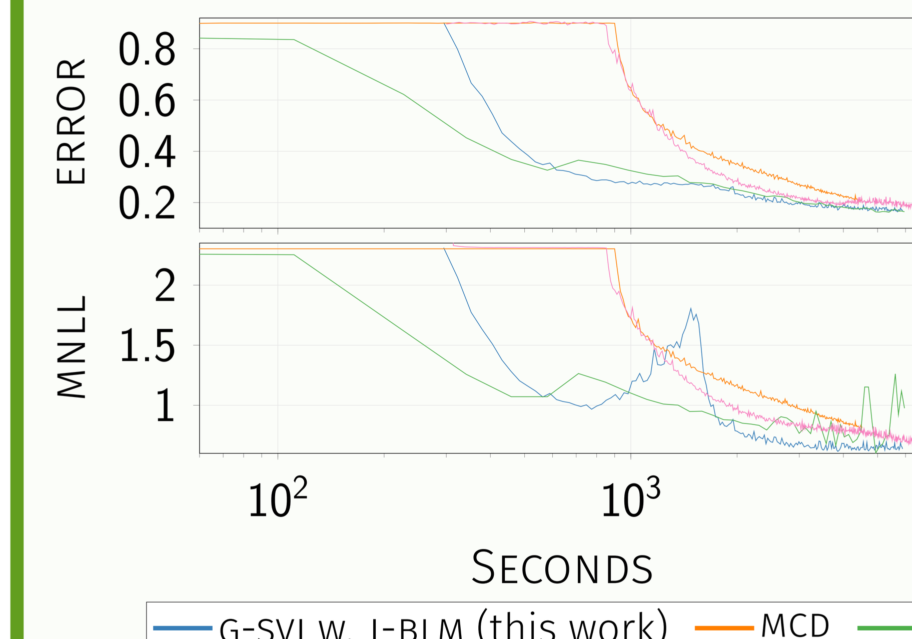


Figure: Progression of test error and test MNLL with different initializations on a 5x100 architecture.

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).



VGG16: 3.5M+ params

	MNLL	ERROR
G-SVI & I-BLM	**0.637**	**0.167**
G-SVI & MAP	0.750	0.201
MCD [1]	0.821	0.215
NOISY-KFAC [4]	0.750*	**0.164**

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

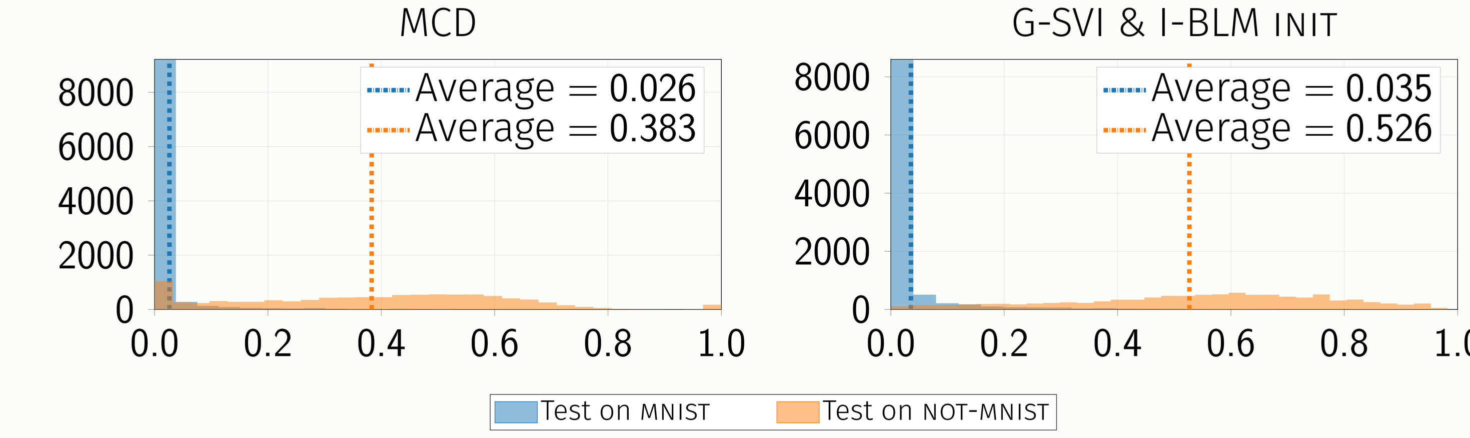


Figure: Entropy distribution while testing on MNIST and NOT-MNIST (higher average entropy on NOT-MNIST means better uncertainty estimation).

Checkout the Full Paper!

S. Rossi, P. Michiardi, and M. Filippone. Good Initializations of Variational Bayes for Deep Models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, USA, 2019*, 2019.