Bayesian Neural Networks

ROY team: Ilya Zharikov,
Roman Isachenko,
Artem Bochkarev

Skolkovo Institute of Science and Technology
Deep Learning course

May 25, 2017

Project goal

Goal

Estimate posterior distributions of the model parameters from data

Probabilistic Programming:

- Uncertainty in predictions;
- Uncertainty in representations;
- Regularizations with priors;
- Transfer learning;
- Hierarchical Neural Networks.

Problem

Monte Carlo sampling is very slow for high-dimensional data

Related work

1. Salvatier J, Wiecki T. V., Fonnesbeck C. Probabilistic programming in Python using PyMC3. // *PeerJ Computer Science*. 2016.

2. Blundell C. et al. Weight Uncertainty in Neural Network // *Proceedings of The 32nd International Conference on Machine Learning*. 2015.

3. Kucukelbir A. et al. Automatic Differentiation Variational Inference // *arXiv preprint arXiv:1603.00788*. – 2017.

Problem Statement

Inference problem

Bayes' theorem states: $\mathbb{P}(\boldsymbol{\theta} \,|\, \mathbf{X}) = \dfrac{\mathbb{P}(\mathbf{X} \,|\, \boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X})}$

Maximum A Posteriori (MAP) estimation

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \left[\ln \mathbb{P}(\boldsymbol{\theta} \,|\, \mathbf{X})\right] = \arg\max_{\boldsymbol{\theta}} \left[\ln \mathbb{P}(\mathbf{X} \,|\, \boldsymbol{\theta}) + \ln \mathbb{P}(\boldsymbol{\theta})\right]$$

Monte Carlo approach:

- Metropolis-Hastings sampling;
- Gibbs sampling;
- No-U-Turn Sampling (NUTS).

Variational Inference

Goal

Approximate posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ by function $q(\boldsymbol{\theta})$ from parametric family.

$$\begin{array}{ccc} \ln p(\mathbf{X}) = & \mathrm{KL}(q||p) & + \ \mathrm{ELBO}(q) \\ & \Updownarrow & \Updownarrow \\ & \displaystyle\int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{X})} \mathrm{d}\boldsymbol{\theta} & \displaystyle\int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \end{array}$$

Minimization of $\mathbf{KL(q||p)}$ ⇔ Maximization of $\mathbf{ELBO(q)}$

Automatic Differentiation Variational Inference (ADVI)

- Automatic transformation of constrained variables $\boldsymbol{\zeta} = T(\boldsymbol{\theta})$;
  Example: $\theta \in \mathbb{R}_+ \Rightarrow \zeta = T(\theta) = \log\theta$, then $\zeta \in \mathbb{R}$.

- $q(\boldsymbol{\zeta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is diagonal;

$$\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* = \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \mathrm{ELBO}(q)$$

- Stochastic optimization;

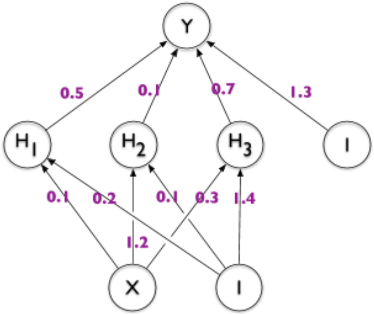
- Reparametrization trick to apply automatic differentiation;

- Adaptive step-size.

Deep Learning

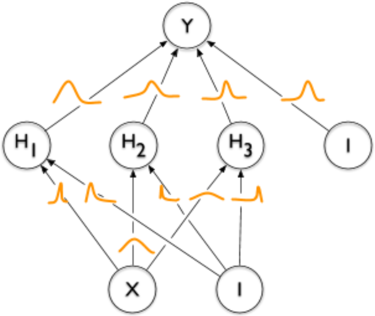
Neural Networks

Predict values of parameters by fitting complex model on the huge dataset



Bayesian Neural Networks

Predict the parameters of the weights distributions from the dataset



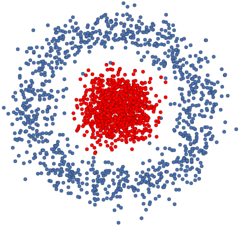
http://bit.ly/2rMQuDq

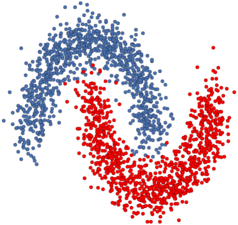
Experiments

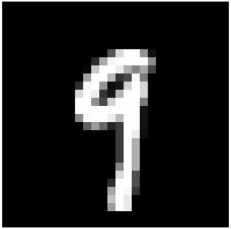
Goals:

- investigate influence of different priors on the predictions
- visualize uncertainties in predictions
- analyze the model behaviour

Datasets:







Course of work

Normal **Laplace** **Cauchy** **Flat**

Fix

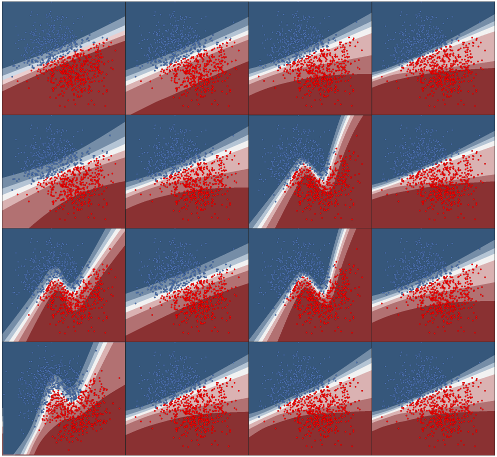
Fixed values

Hierarchical modeling

Half-Normal

Half-Cauchy

Inverse-Gamma

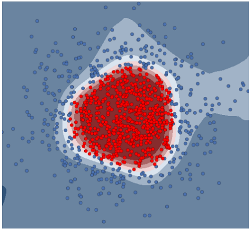


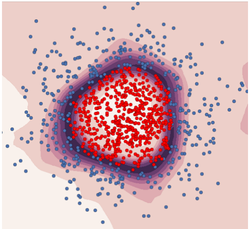
Synthetic data

Posterior Probability

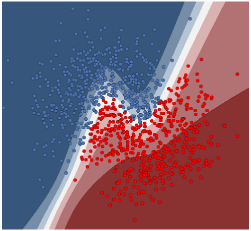
Uncertainty

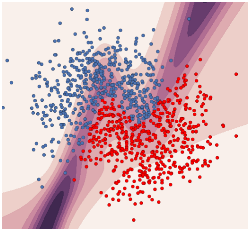
Prior: Cauchy
Hyperprior: Inverse-Gamma
Accuracy: 0.735





Prior: Normal
Hyperprior: Inverse-Gamma
Accuracy: 0.851





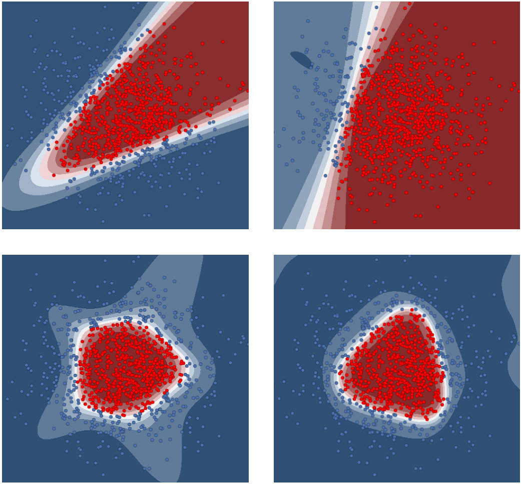
Hierarchical modelling

Prior: Normal

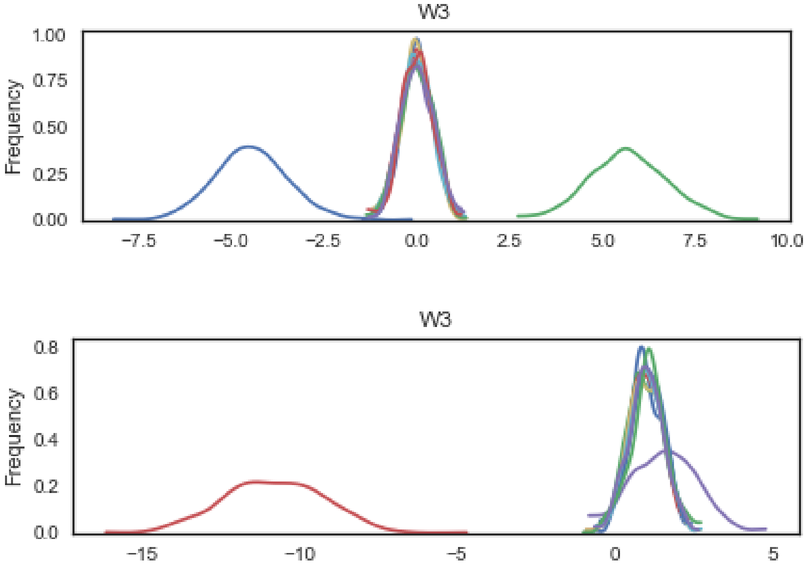
Prior: Laplace

Hyperprior:
Fixed values

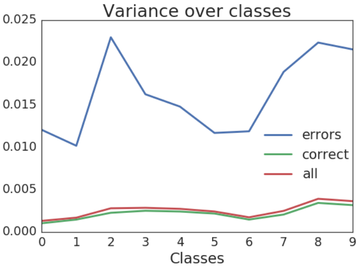
Hyperprior:
Inverse-Gamma

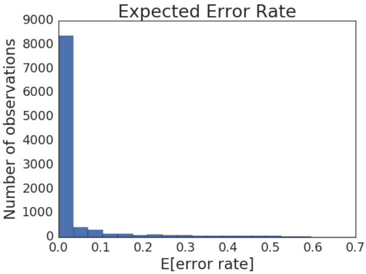


Laplace sparsity



MNIST





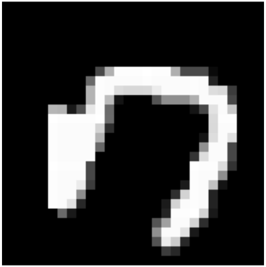
Conclusions:

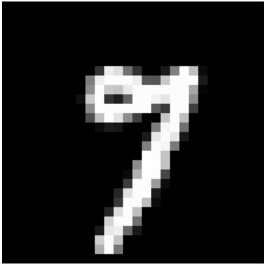
- Accuracy score: 97.7%;
- Variance is much higher for misclassified pictures;
- Model is not always confident.

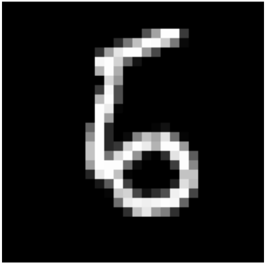
MNIST

Misclassified pictures with **zero expected error rate**:

True	Prediction	True	Prediction	True	Prediction
7	0	9	7	5	6



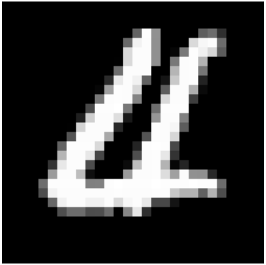


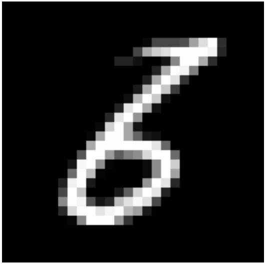


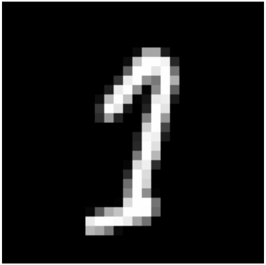
MNIST

Pictures with the **lowest confidence**:

True	Prediction	True	Prediction	True	Prediction
4	0	6	8	1	1







Conclusion

- Posterior distribution has clear advantages over point estimate for deep learning
- Variational inference allows faster approximation of the posterior
- Hierarchical models better adapt to diverse datasets
- Full bayesian approach is still problematic for very complex models