Posterior evaluation for Bayesian Deep Neural Networks

Аминов Тимур

Московский физико-технический институт

26 февраля 2020 г.

Plan

$$\mathbf{X} \to \mathbf{W}x + b \to \mathbf{P}(y)$$

Prediction

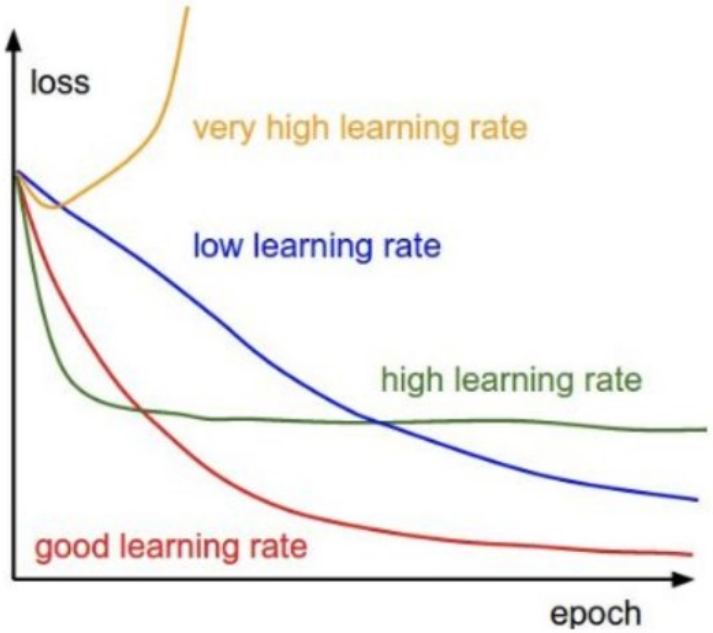
$\mathbf{P}(y|x) = \sigma(w \cdot x + b)$

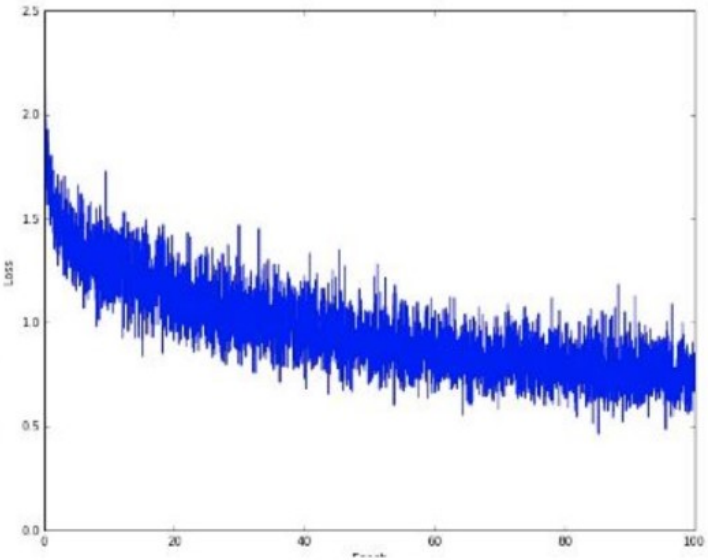
Loglikelihood

$\mathbf{L} = -\sum_{i=1}^{n} y_i \log \mathbf{P}(y|x_i) + (1 - y_i) \log (1 - \mathbf{P}(y|x_i))$

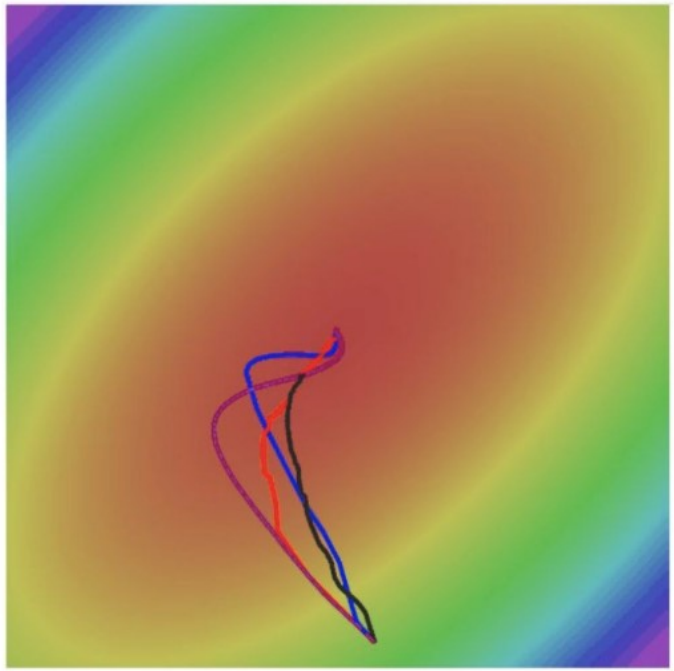
Stochastic gradient descent is used to optimize NN parameters.

$$x_{t+1} = x_t - \text{learning rate} \cdot dx$$





Different optimizers



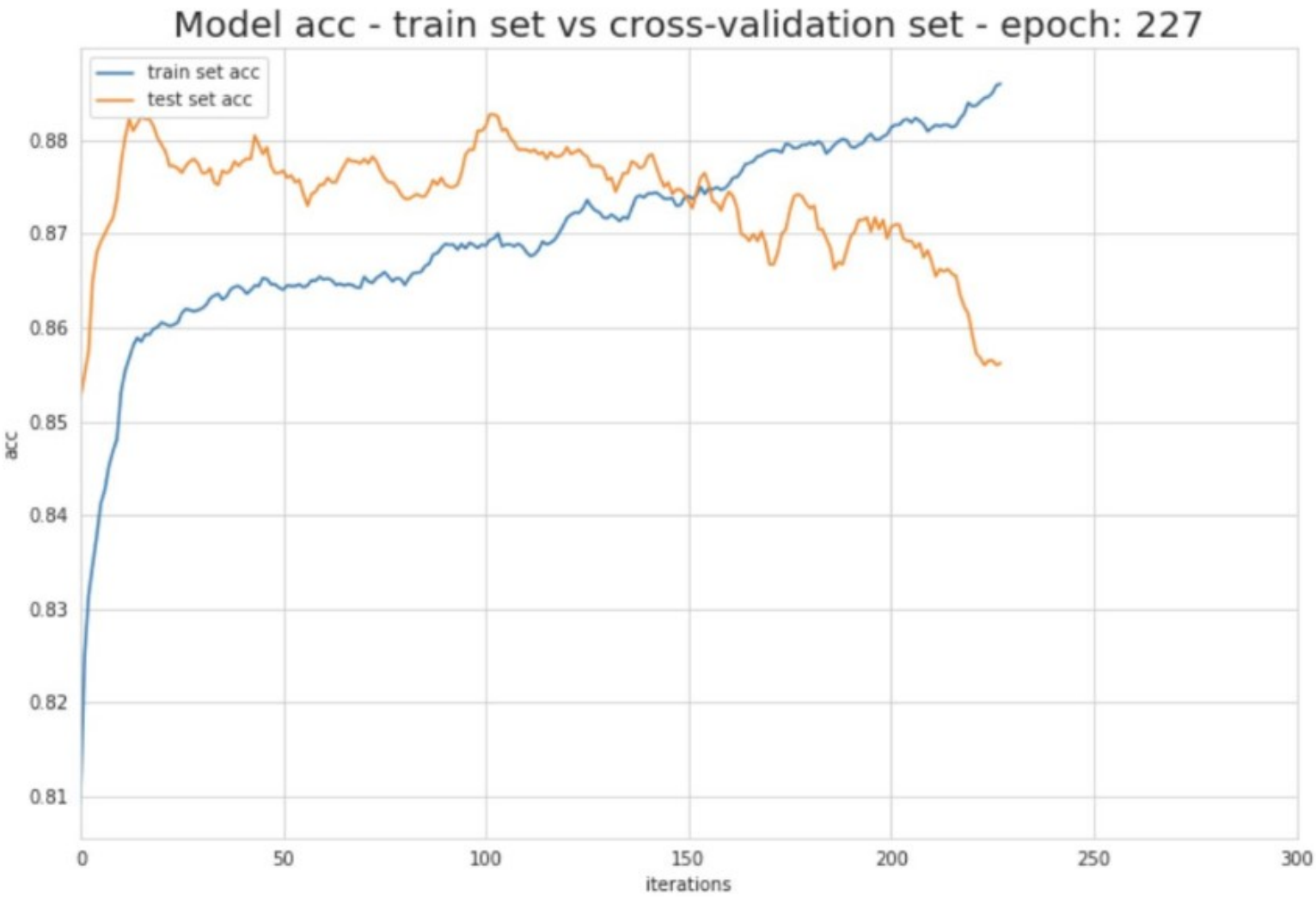
SGD

SGD+Momentum

RMSProp

Adam

Problem: overfitting



Solution: regularization

$$L_{reg}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i|x_i,\boldsymbol{\theta}\right) + \lambda \boldsymbol{R}(\boldsymbol{\theta})$$

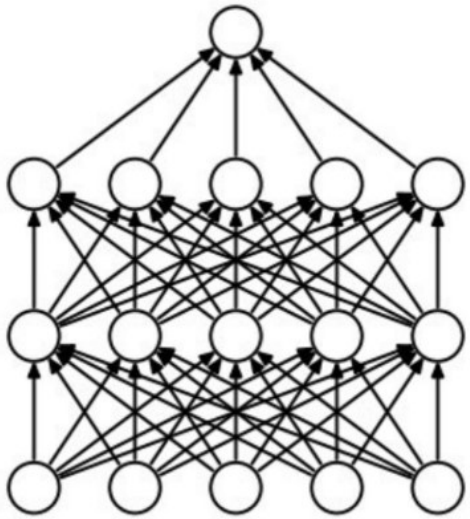
Adding some extra term to the loss function

- L_2 Regularization : $\boldsymbol{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$
- L_1 Regularization : $\boldsymbol{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$
- Elastic Net ($L_1 + L_2$) : $\boldsymbol{R}(\boldsymbol{\theta}) = \beta\|\boldsymbol{\theta}\|_2^2 + \|\boldsymbol{\theta}\|_1$

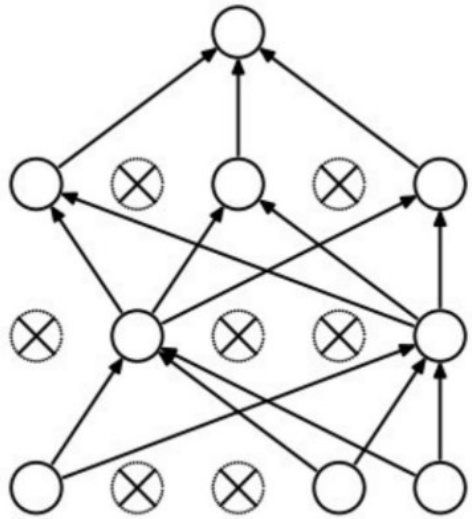
Regularization: Dropout

Some neurons are “dropped” during training.

Prevents overfitting.

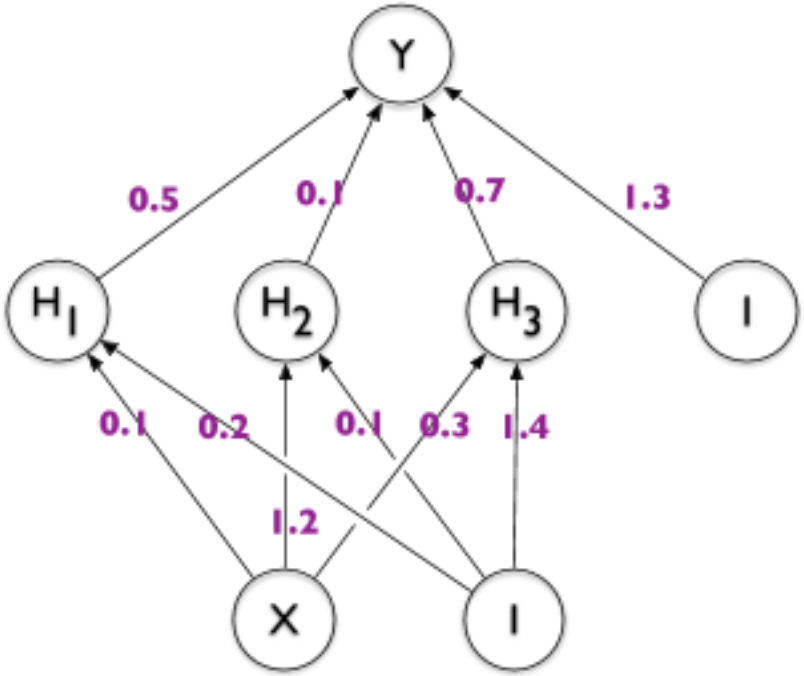


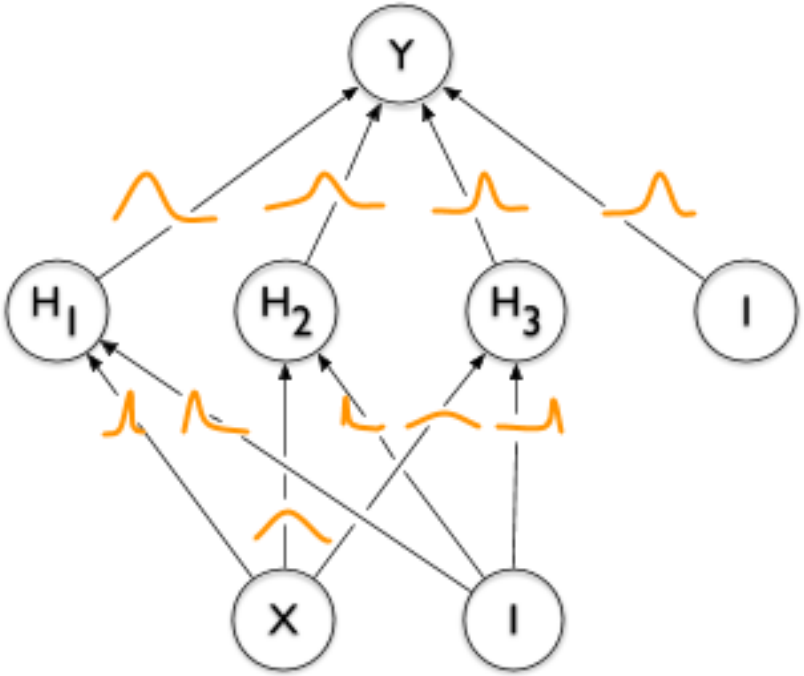
(a) Standard Neural Net



(b) After applying dropout.

Bayesian deep network





Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior

$$p(\theta|\mathcal{D}) = \frac{p(y,\theta|X)p(\theta)}{p(y|X)} \propto p(y,\theta|X)p(\theta)$$

Where $p(y,\theta|X) = p(y|X,\theta)p(\theta)$

Evidence

$$f(\Theta) = p(y|X,\Theta) = \int p(y|X,w)p(w|\Theta)\mathrm{d}\Theta$$

Where Θ - hyperparameters of model, w - parameters of model

Regularized cross-entropy

$$L(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i|x_i,\theta\right) + \Omega(\theta)$$

$\Omega(\theta)$ is a regularizer over model parameters

Temperature

$$p(\theta|\mathcal{D}) \propto exp(-U(\theta)/T)$$

Posterior energy function

$$U(\theta) := -\sum_{i=1}^{n} \log p\left(y_i|x_i,\theta\right) - \log p(\theta)$$

$p(\theta)$ is a proper prior density function

Prediction

$$p(\boldsymbol{y}|\boldsymbol{X},\mathcal{D}) = \int p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\mathrm{d}\boldsymbol{\theta}$$

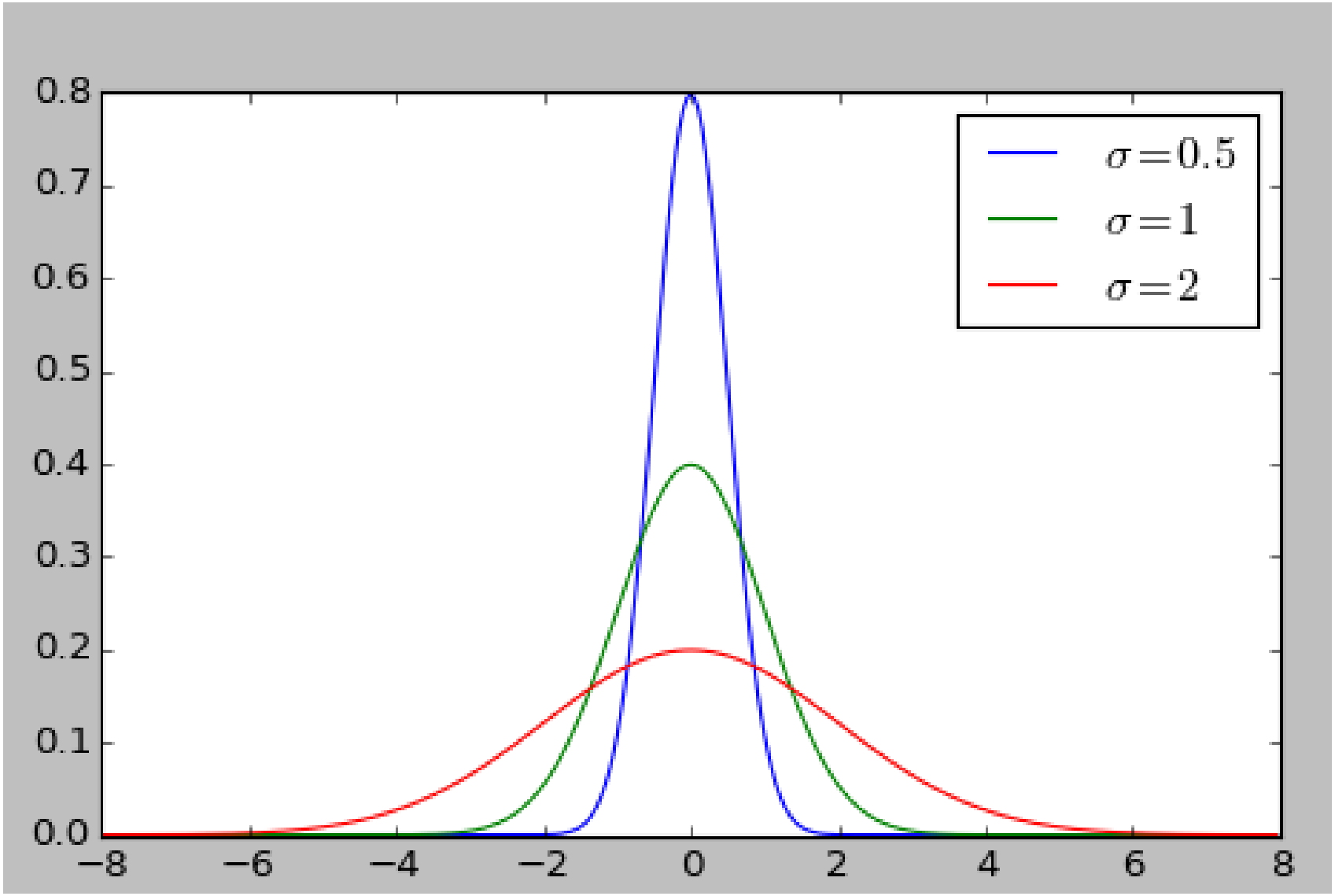
Analytical solution of this integral is not always available

Approximation

$$p(y|x,\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S} p\left(y|x,\boldsymbol{\theta}^{(s)}\right)$$

Where $\boldsymbol{\theta}^{(s)}, s = 1 \ldots S$ sampled from $p(\boldsymbol{\theta}|\mathcal{D})$

Prior: Temperature concentration



Prior: Temperature

T = 0

All posterior probability mass is concentrated on the set of maximum a posteriori (MAP) point estimates

T < 1

Corresponds to artificially sharpening the posterior, which can be interpreted as overcounting the data by a factor of 1/T and a rescaling of the prior as $p(\theta)^{\frac{1}{T}}$

T = 1

Corresponds to the true Bayes posterior and performance gains for T < 1 point to a deeper and potentially resolvable problem with the prior, likelihood, or inference procedure

Why Should Bayes (T = 1) be Better?

Theory

For several models where the predictive performance can be analyzed it is known that the posterior predictive can dominate common point-wise estimators based on the likelihood

Classical empirical evidence

For classical statistical models, averaged predictions have been observed to be more robust in practice

Model averaging

Recent deep learning models based on deterministic model averages, have shown good predictive performance

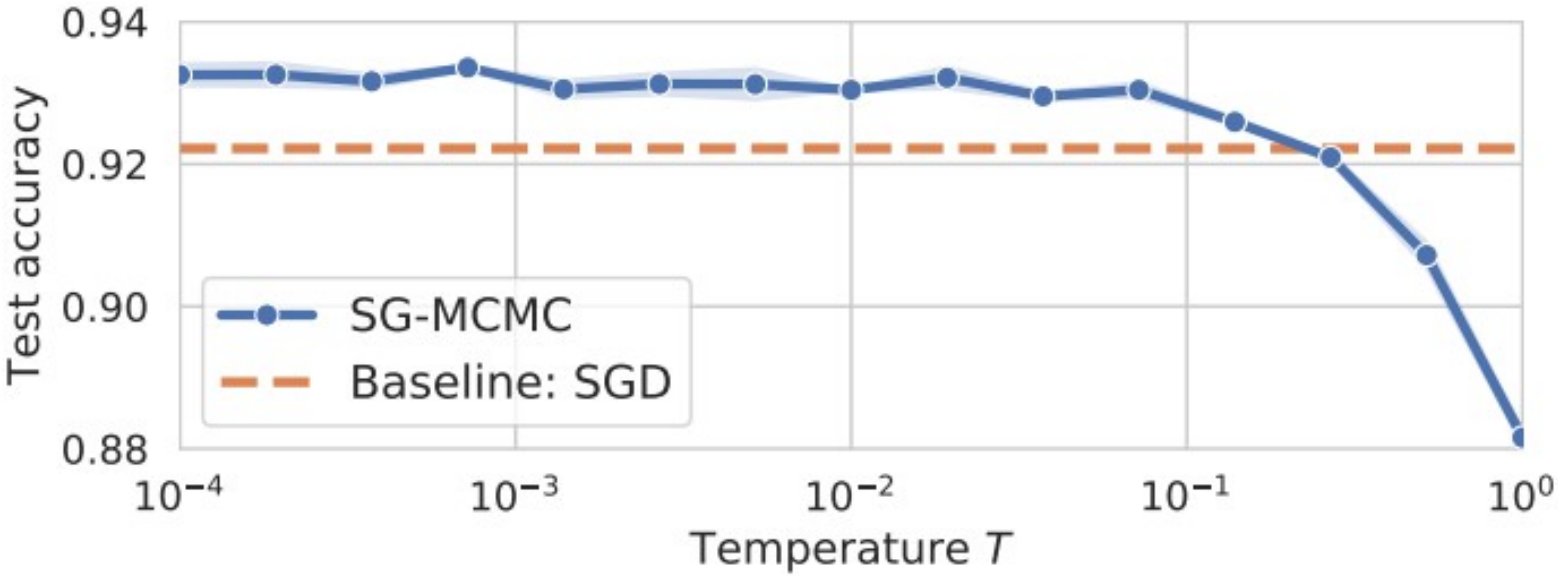


Figure 1. The "**cold posterior**" effect: for a ResNet-20 on CIFAR-10 we can improve the generalization performance significantly by cooling the posterior with a temperature $T \ll 1$, deviating from the Bayes posterior $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta})/T)$ at $T = 1$.

Prior: Experiment

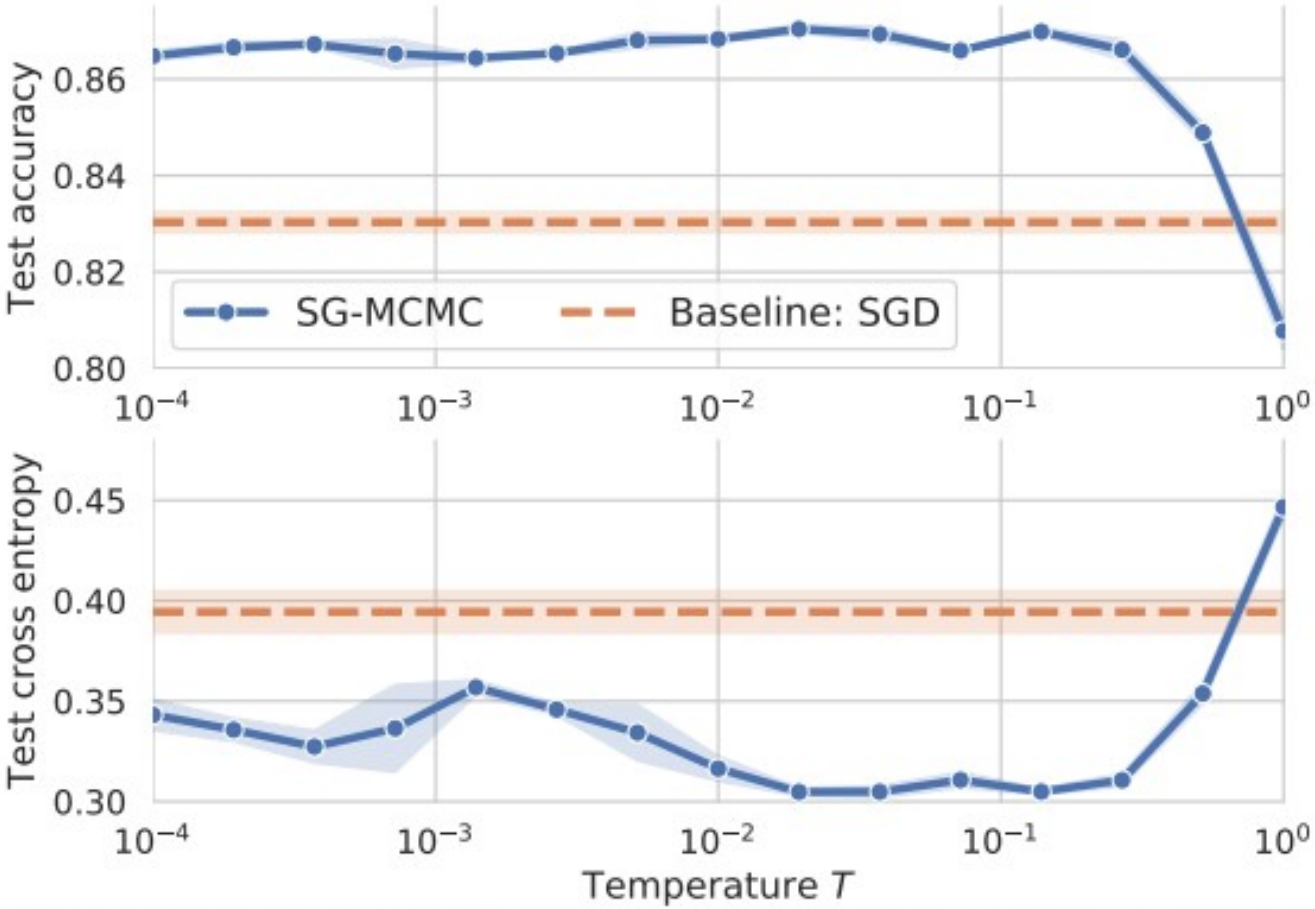


Figure 3. Predictive performance on the IMDB sentiment task test set for a tempered CNN-LSTM Bayes posterior. Error bars are $\pm$ one standard error over three runs. See Appendix A.4.

Biased SG-MCMC Hypothesis

Lack of accept/reject Metropolis-Hastings corrections in SG-MCMC introduces bias

Minibatch Noise Hypothesis

Gradient noise from minibatching causes inaccurate sampling at $T = 1$

Dirty Likelihood Hypothesis

Deep learning practices that violate the likelihood principle cause deviation from the Bayes posterior

Bias-variance Tradeoff Hypothesis

For $T = 1$ the posterior is diverse and there is high variance between model predictions. For $T \ll 1$ we sample nearby modes and reduce prediction variance but increase bias; the variance dominates the error and reducing variance ($T \ll 1$) improves predictive performance

Hypothesis: Problems with the Prior p(θ)

Bad Prior Hypothesis

The current priors used for BNN parameters are inadequate, unintentionally informative, and their effect becomes stronger with increasing model depths and capacity

Implicit Initialization Prior in SGD

The inductive bias from initialization is strong and beneficial for SGD but harmed by SG-MCMC sampling

Literature

1. How Good is the Bayes Posterior in Deep Neural Networks Really?
2. Выбор моделей в машинном обучении
3. Machine learning course at MIPT