How good is the Bayes Posterior in Deep Neural Networks really?

Wenzel, Florian and Roth, Kevin and Veeling, Bastiaan S and Świątkowski, Jakub and Tran, Linh and Mandt, Stephan and Snoek, Jasper and Salimans, Tim and Jenatton, Rodolphe and Nowozin, Sebastian

Presented by: Pritam Karmokar

Bayesian Learning Series (#11)

Robotic Vision Lab (RVL)
[Weekly Seminars 03/03/2021]

Outline

Outline

Introduction

(in the past five years)

Inference procedures have been developed that allow for **Bayesian inference in Deep Neural Networks**

Introduction
(in the past five years)

Inference procedures have been developed that allow for **Bayesian inference in Deep Neural Networks**

- Increasingly accurate

Introduction
(in the past five years)

Inference procedures have been developed that allow for **Bayesian inference in Deep Neural Networks**

- Increasingly accurate
- Efficiently approximate

Introduction
(in the past five years)

Inference procedures have been developed that allow for **Bayesian inference in Deep Neural Networks**

- Increasingly accurate
- Efficiently approximate

$+$

- Algorithmic progress

Introduction
(in the past five years)

Inference procedures have been developed that allow for **Bayesian inference in Deep Neural Networks**

- Increasingly accurate
- Efficiently approximate

$+$

- Algorithmic progress
- Improved uncertainty quantification

Introduction

(in the past five years)

Inference procedures have been developed that allow for **Bayesian inference in Deep Neural Networks**

- Increasingly accurate
- Efficiently approximate

$+$

- Algorithmic progress
- Improved uncertainty quantification and sample efficiency

Introduction
(as of early 2020)

Despite all that, ..

Introduction
(as of early 2020)

Despite all that, ..

- **No publicized deployments** of Bayesian neural networks **in industrial practice**

Introduction
(as of early 2020)

Despite all that, ..

- **No publicized deployments** of Bayesian neural networks **in industrial practice**

?

Introduction

?..

Cast doubts on,

Introduction

?..

Cast doubts on,

- **current understanding** of Bayes posteriors in popular deep neural networks

Introduction

Demonstrate through careful MCMC sampling

Introduction

Demonstrate through careful MCMC sampling

- Bayesian posterior predictives yield **systematically worse predictions** compared to simpler methods (*e.g., point estimates obtained from SGD*)

Introduction

Demonstrate through careful MCMC sampling

- Bayesian posterior predictives yield **systematically worse predictions** compared to simpler methods (*e.g., point estimates obtained from SGD*)
- predictive performance **improved through use of a cold posterior** that overcounts evidence.

Introduction

- **Demonstrate** through careful MCMC sampling

Introduction

- **Demonstrate** through careful MCMC sampling
- **Put forward** several hypotheses that could explain cold posteriors

Introduction

- **Demonstrate** through careful MCMC sampling
- **Put forward** several hypotheses that could explain cold posteriors
- **Evaluate** hypotheses through experiments

Introduction

- **Demonstrate** through careful MCMC sampling
- **Put forward** several hypotheses that could explain cold posteriors
- **Evaluate** hypotheses through experiments
- **Question** the goal of accurate posterior approximations in Bayesian Deep Learning

Introduction

Question the goal of accurate posterior approximations

- If the true Bayes posterior is poor, ..

Introduction

Question the goal of accurate posterior approximations

- If the true Bayes posterior is poor, **what is the use of more accurate approximations?**

Introduction

Question the goal of accurate posterior approximations

- If the true Bayes posterior is poor, **what is the use of more accurate approximations?**
- **Time to focus** on understanding of **origin of improved performance of cold posteriors**

Introduction

In supervised learning, we use

Introduction

In supervised learning, we use

- training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1,\ldots,n}$

Introduction

In supervised learning, we use

- training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1,\ldots,n}$
- probabilistic model $p(y|x, \boldsymbol{\theta})$

Introduction

In supervised learning, we use

- training dataset $\mathcal{D} = \{x_i, y_i\}_{i=1,\dots,n}$
- probabilistic model $p(y|x, \boldsymbol{\theta})$
- to minimize regularized cross-entropy objective

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

Outline

Bayesian Deep Learning

- do **not** optimize for a ***single*** likely model

Bayesian Deep Learning

- do **not** optimize for a ***single*** likely model

Bayesian Deep Learning

- do **not** optimize for a ***single*** likely model
- instead want to discover ***all*** likely models

Bayesian Deep Learning

- do **not** optimize for a ***single*** likely model
- instead want to discover ***all*** likely models
- **approximate** the ***posterior distribution*** over model parameters, $p(\boldsymbol{\theta}|\mathcal{D}) \propto exp(-U(\boldsymbol{\theta})/T)$

Bayesian Deep Learning

- do **not** optimize for a ***single*** likely model
- instead want to discover ***all*** likely models
- **approximate** the ***posterior distribution*** over model parameters, $p(\boldsymbol{\theta}|\mathcal{D}) \propto exp(-U(\boldsymbol{\theta})/T)$, where $U(\boldsymbol{\theta})$ is the ***posterior energy function***,

$$U(\boldsymbol{\theta}) := -\sum_{i=1}^{n} \log p(y_i|x_i, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \tag{2}$$

Bayesian Deep Learning

- Given $p(\boldsymbol{\theta}|\mathcal{D})$, we *predict* on a new instance x, by averaging over all likely models,

$$p(y|x, \mathcal{D}) = \int p(y|x, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}, \tag{3}$$

Bayesian Deep Learning

- Given $p(\boldsymbol{\theta}|\mathcal{D})$, we *predict* on a new instance x, by averaging over all likely models,

$$p(y|x,\mathcal{D}) = \int p(y|x,\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}, \tag{3}$$

where (**??**) is also known as ***posterior predictive*** or ***Bayes ensemble***.

Bayesian Deep Learning

- solving (**??**) exactly is not possible

Bayesian Deep Learning

- solving (**??**) exactly is not possible
- instead approximate it using sample approximation
  $p(y|x, \mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S} p(y|x, \boldsymbol{\theta}^{(s)})$

Bayesian Deep Learning

- solving (**??**) exactly is not possible
- instead approximate it using sample approximation $p(y|x, \mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S} p(y|x, \boldsymbol{\theta}^{(s)})$, where $\boldsymbol{\theta}^{(s)}$ is approximately sampled from $p(\boldsymbol{\theta}|\mathcal{D})$

Bayesian Deep Learning

the suprising effect called the "Cold Posterior" effect

Cold Posteriors: among all temperized posteriors the best posterior predictive performance on holdout data is achieved at temperature $T < 1$.

Bayesian Deep Learning

the suprising effect called the "Cold Posterior" effect

Cold Posteriors: among all temperized posteriors the best posterior predictive performance on holdout data is achieved at temperature T < 1.

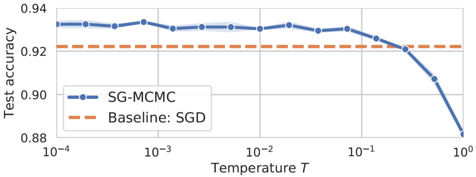

Figure 1. The "**cold posterior**" effect: for a ResNet-20 on CIFAR-10 we can improve the generalization performance significantly by cooling the posterior with a temperature $T \ll 1$, deviating from the Bayes posterior $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta})/T)$ at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

Outline

Why should Bayes ($T = 1$) be better?

Why expect that predictions made by the *ensemble model* (**??**) could improve over predictions made at a single well-chosen parameter?

Why should Bayes ($T = 1$) be better?

Three reasons:

- ***Theory***: known that (**??**) can dominate common point-wise estimators based on likelihood, even in case of misspecification

Why should Bayes ($T = 1$) be better?

Three reasons:

- ***Theory***: known that (**??**) can dominate common point-wise estimators based on likelihood, even in case of misspecification
- ***Classical empirical evidence***: for classical statistical models, averaged predictions (**??**) have been observed to be more robust in practice

Why should Bayes ($T = 1$) be better?

Three reasons:

- ***Theory***: known that (**??**) can dominate common point-wise estimators based on likelihood, even in case of misspecification
- ***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

Why should Bayes ($T = 1$) be better?

- Note that a large body of work in Bayesian Deep Learning is **motivated by the assertion** that predicting **using (??) is desirable**

Why should Bayes ($T = 1$) be better?

- Note that a large body of work in Bayesian Deep Learning is **motivated by the assertion** that predicting **using (??) is desirable**
- Authors **confront the assertion** through **simple experiments**

Why should Bayes ($T = 1$) be better?

- Note that a large body of work in Bayesian Deep Learning is **motivated by the assertion** that predicting **using (??) is desirable**
- Authors **confront the assertion** through **simple experiments**
- Show that **our understanding** of the Bayes posterior in deep models **is limited**

Contributions

- **Demonstrate** two models and tasks (ResNet-20 on CIFAR-10 and CNN-LSTM on IMDB) that the Bayes posterior predictive has **poor performance** compared to SGD-trained models

Contributions

- **Demonstrate** two models and tasks (ResNet-20 on CIFAR-10 and CNN-LSTM on IMDB) that the Bayes posterior predictive has **poor performance** compared to SGD-trained models
- **Put forth and systematically examine hypotheses** that could **explain the observed behaviour**

Contributions

- **Demonstrate** two models and tasks (ResNet-20 on CIFAR-10 and CNN-LSTM on IMDB) that the Bayes posterior predictive has **poor performance** compared to SGD-trained models
- **Put forth and systematically examine hypotheses** that could **explain the observed behaviour**
- **Introduce two new diagnostic tools** for assessing the approximation quality of stochastic gradient Markov chain Monte Carlo methods (SG-MCMC) and **demonstrate** that the posterior is accurately simulated by existing SG-MCMC methods

Outline

Deep Learning Models: ResNet-20 and LSTM

ResNet-20 on CIFAR-10

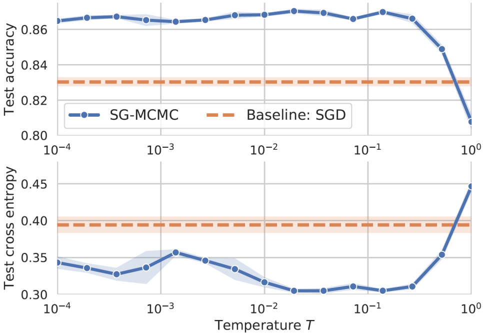

Figure 3. Predictive performance on the IMDB sentiment task test set for a tempered CNN-LSTM Bayes posterior. Error bars are $\pm$ one standard error over three runs. See Appendix A.4.

Deep Learning Models: ResNet-20 and LSTM

CNN-LSTM on IMDB text classification

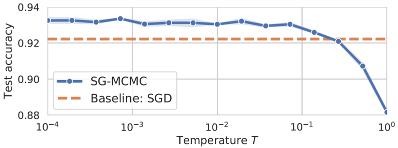

Figure 1. The "**cold posterior**" effect: for a ResNet-20 on CIFAR-10 we can improve the generalization performance significantly by cooling the posterior with a temperature $T \ll 1$, deviating from the Bayes posterior $p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp(-U(\boldsymbol{\theta})/T)$ at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

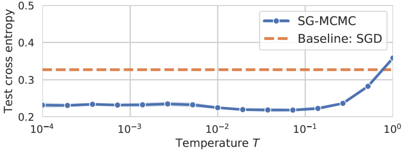

Figure 2. Predictive performance on the CIFAR-10 test set for a cooled ResNet-20 Bayes posterior. The SGD baseline is separately tuned for the same model (Appendix A.2).

Outline

Why is a Temperature of $T < 1$ a Problem?

Two reasons:

1. $T < 1$ corresponds to artificially sharpening the posterior, interpreted as
   - overcounting evidence by a factor of $1/T$

Why is a Temperature of $T < 1$ a Problem?

Two reasons:

1. $T < 1$ corresponds to artificially sharpening the posterior, interpreted as
   - overcounting evidence by a factor of $1/T$
   - rescaling of prior as $p(\boldsymbol{\theta})^{\frac{1}{T}}$

Why is a Temperature of $T < 1$ a Problem?

Two reasons:

1. $T < 1$ corresponds to artificially sharpening the posterior, interpreted as
   - overcounting evidence by a factor of $1/T$
   - rescaling of prior as $p(\boldsymbol{\theta})^{\frac{1}{T}}$
   - equivalent to a Bayes posterior obtained from a dataset consisting of $1/T$ replications of the original data

Why is a Temperature of $T < 1$ a Problem?

Two reasons:

1. $T < 1$ corresponds to artificially sharpening the posterior, interpreted as
   - overcounting evidence by a factor of $1/T$
   - rescaling of prior as $p(\boldsymbol{\theta})^{\frac{1}{T}}$
   - equivalent to a Bayes posterior obtained from a dataset consisting of $1/T$ replications of the original data
   - for $T = 0$ all probability mass is concentrated on thet set of maximum a posteriori (MAP) point estimates

Why is a Temperature of $T < 1$ a Problem?

Two reasons:

1. $T < 1$ corresponds to artificially sharpening the posterior, interpreted as
   - overcounting evidence by a factor of $1/T$
   - rescaling of prior as $p(\boldsymbol{\theta})^{\frac{1}{T}}$
   - equivalent to a Bayes posterior obtained from a dataset consisting of $1/T$ replications of the original data
   - for $T = 0$ all probability mass is concentrated on thet set of maximum a posteriori (MAP) point estimates
2. $T = 1$ corresponds to the true Bayes posterior
   - performance gains for $T < 1$ point to a potentially resolvable problem with the prior, likelihood, or inference procedure

Confirmation from Literature

SG-MCMC on deep neural networks and posterior tempering

Reference	Temperature T
(Li et al., 2016)	$1/\sqrt{n}$
(Leimkuhler et al., 2019)	$T < 10^{-3}$
(Heek & Kalchbrenner, 2020)	$T = 1/5$
(Zhang et al., 2020)	$T = 1/\sqrt{50000}$

Confirmation from Literature

Variational Bayes approach tempering likelihood part of posterior

- Variational bayes approach to Bayesian neural networks optimize τ of a variational distribution $q(\boldsymbol{\theta}|\tau)$ by minimizing the evidence lower bound (ELBO),

$$\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\tau)}\left[\sum_{i=1}^{n}\log p(y_i|x_i,\boldsymbol{\theta})\right] - \lambda D_{KL}(q(\boldsymbol{\theta}|\tau)\|p(\boldsymbol{\theta})) \quad (4)$$

Confirmation from Literature

Variational Bayes approach tempering likelihood part of posterior

- Variational bayes approach to Bayesian neural networks optimize τ of a variational distribution $q(\boldsymbol{\theta}|\tau)$ by minimizing the evidence lower bound (ELBO),

$$\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\tau)}\left[\sum_{i=1}^{n}\log p(y_i|x_i,\boldsymbol{\theta})\right] - \lambda D_{KL}(q(\boldsymbol{\theta}|\tau)\|p(\boldsymbol{\theta})) \qquad (4)$$

- $\lambda = 1$ directly minimizes $D_{KL}(q(\boldsymbol{\theta}|\tau)\|p(\boldsymbol{\theta}))$, for rich variational families approximates the true Bayes $p(\boldsymbol{\theta}|\mathcal{D})$

Confirmation from Literature

Variational Bayes approach tempering likelihood part of posterior

- $\lambda = 1$ directly minimizes $D_{KL}(q(\boldsymbol{\theta}|\tau)\|p(\boldsymbol{\theta}))$, for rich variational families approximates the true Bayes $p(\boldsymbol{\theta}|\mathcal{D})$

Reference	KL term weight λ in (4)
(Zhang et al., 2018)	$\lambda \in \{1/2, 1/10\}$
(Bae et al., 2018)	tuning of λ, unspecified
(Osawa et al., 2019)	$\lambda \in \{1/5, 1/10\}$
(Ashukha et al., 2020)	λ from 10^{-5} to 10^{-3}

- KL-weighted ELBO (**??**) arises from tempering the likelihood part of the posterior

Confirmation from Literature

Cold posterior trail in literature

We are not aware of *any* published work demonstrating well-performing Bayesian deep learning at temperature $T = 1$.