Variational Inference based on Robust Divergences

Futoshi Futami (1, 2) Issei Sato (1, 2) Masashi Sugiyama (2, 1)

(1) The University of Tokyo (2) RIKEN

Abstract

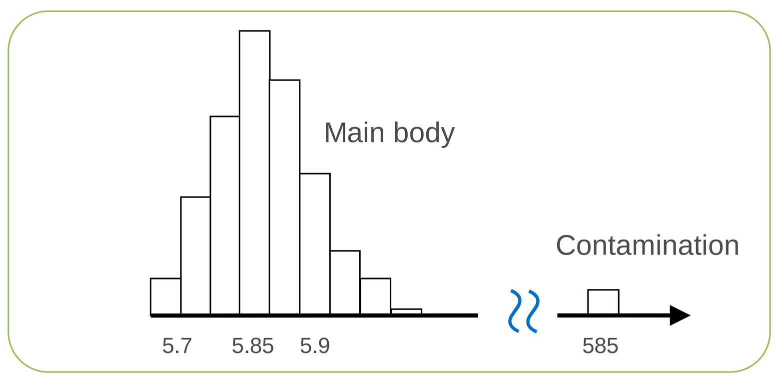
Motivation

- Robustness to outliers is becoming more important these days.
- A standard approach to robust machine learning is a model-based method, e.g. use the Student-t instead of the Gaussian as a likelihood. However, such a model-based method is applicable only to simple setup.

Main Contributions

- To handle more complex models, we employ the optimization and variational formulation of Bayesian inference. In this formulation, the posterior model is optimized to the data under the Kullback-Leibler (KL) divergence, while it is regularized to be close to the prior.
- We propose replacing the KL divergence for data fitting to a robust divergence, such as β-divergence and γ-divergence

Introduction



- Samples are generated from some unknown distribution:

$$\{x_i\}_{i=1}^N \sim p^*(x) \qquad p^*(x) = (1-\varepsilon)p_0^*(x) + \varepsilon\delta(x)$$

Main body Contamination

- In outlier-robust inference, we aim at placing an estimated probability distribution close to the main body of the unknown distribution.

Maximum likelihood (ML) estimation

- We estimate an unknown probability distribution $p^*(x)$ from its independent samples $x_{1:N} = \{x_i\}_{i=1}^N$.
- In ML estimation, we minimize the generalization error measured by the KL divergence from $p^*(x)$ to a parametric model $p(x;\theta)$.
- We approximate $p^*(x)$ by the empirical distribution $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^N \delta(x, x_i)$
- ML estimator is reduced to:

$$\underset{\theta}{\arg\min}\, D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x;\theta)\right) \Rightarrow 0 = \frac{1}{N}\sum_{i=1}^N \frac{\partial}{\partial\theta}\ln p(x_i;\theta)$$

Bayesian inference

- θ is regarded as a random variable, following the prior $p(\theta)$.
- We update our prior belief by Bayes' theorem and obtain the Bayesian posterior $p(\theta|x_{1:N}) = \frac{p(x_{1:N}|\theta)p(\theta)}{p(x_{1:N})}$.

This posterior can also be obtained by solving the following optimization problem

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, L(q(\theta))$$

$$L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right)$$

- $d_{\mathrm{KL}}(\hat{p}(x)\|p(x|\theta)) = -\frac{1}{N}\sum_{i=1}^N \ln p(x_i|\theta)$: Cross entropy
- $\mathcal{P}$: The set of all probability distributions

- This is often intractable analytically, and need some approximation method.
- A popular approach is to restrict the domain of the optimization problem to analytically tractable probability distributions $q(\theta;\lambda)\in\mathcal{Q}$, where λ is a parameter.
- The optimization problem is expressed as $\underset{q(\theta;\lambda)\in\mathcal{Q}}{\arg\min}\, L(q(\theta;\lambda))$
  - This method is called variational inference(VI).
  - $-L(q(\theta;\lambda))$ is called the evidence lower-bound (ELBO).

Robust divergences

ML estimation is sensitive to outliers because it treats all data points equally. To avoid this, robust divergences were proposed.

β-divergence [1]

$$D_\beta(g\|f) = \frac{1}{\beta}\int g(x)^{1+\beta}dx + \frac{\beta+1}{\beta}\int g(x)f(x)^\beta dx + \int f(x)^{1+\beta}dx$$

γ-divergence [2]

$$D_\gamma(g\|f) = \frac{1}{\gamma(1+\gamma)}\ln\int g(x)^{1+\gamma}dx - \frac{1}{\gamma}\ln\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\ln\int f(x)^{1+\gamma}dx$$

- Similarly to ML estimation, minimizing the β-divergence (or the γ-divergence) from $p^*(x)$ to $p(x;\theta)$ yields:

$$\underset{\theta}{\arg\min}\, D_\beta\left(\hat{p}(x)\|p(x;\theta)\right)$$

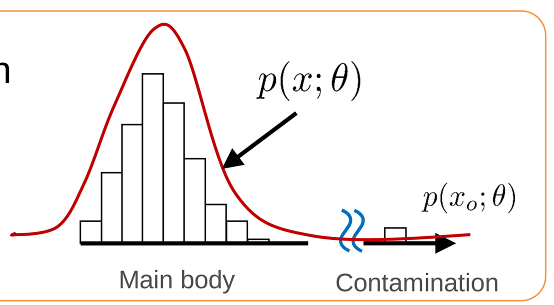
$$0 = \frac{1}{N}\sum_{i=1}^N p(x_i;\theta)^\beta\frac{\partial}{\partial\theta}\ln p(x_i;\theta) - \mathbb{E}_{p(x;\theta)}\left[p(x;\theta)^\beta\frac{\partial}{\partial\theta}\ln p(x;\theta)\right]$$

- The first term is the likelihood weighted according to the power of the probability for each data point.
- The probabilities of outliers are usually much smaller than those of inliers, and thus those weights effectively suppress the likelihood of outliers.

Intuition

We want to model the distribution of the main body of data

Outliers x_o have small $p(x_o;\theta)$



Proposed method

Robust variational inference

Bayesian Inference

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right)$$

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, \mathbb{E}_{q(\theta)}\left[D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right] + \frac{1}{N}D_{\mathrm{KL}}\left(q(\theta)\|p(\theta)\right).$$

- The first term can be regarded as the expected likelihood, while the second term "regularizes" $p(\theta)$ to be close to the prior $q(\theta)$.
- To enhance the robustness to outliers, we replace the KL divergence in the expected likelihood term with the β-divergence:

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, \mathbb{E}_{q(\theta)}\left[D_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right] + \frac{1}{N}D_{\mathrm{KL}}\left(q(\theta)\|p(\theta)\right).$$

$$\underset{q(\theta)\in\mathcal{P}}{\arg\min}\, L_\beta(q(\theta)),$$

$$L_\beta(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right)$$

β cross-entropy: $d_\beta(\hat{p}(x)\|p(x|\theta)) = -\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^N p(x_i|\theta)^\beta + \int p(x|\theta)^{1+\beta}dx$

- This is also generally intractable, let's use VI by restricting the domain of the optimization $q(\theta;\lambda)\in\mathcal{Q}$
- The optimization problem is expressed as $\underset{q(\theta;\lambda)\in\mathcal{Q}}{\arg\min}\, L_\beta(q(\theta;\lambda))$ and we call this β-VI
- We can make the objective function by replacing the cross-entropy with a corresponding cross entropy shown in the following table.

	Unsupervised	Supervised
β	$-\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^N p(x_i\|\theta)^\beta + \int p(x\|\theta)^{1+\beta}dx$	$-\frac{\beta+1}{\beta}\left\{\frac{1}{N}\sum_{i=1}^N p(y_i\|x_i,\theta)^\beta\right\} + \left\{\frac{1}{N}\sum_{i=1}^N \int p(y\|x_i,\theta)^{1+\beta}dy\right\}$
γ	$-\frac{1}{N}\frac{\gamma+1}{\gamma}\sum_{i=1}^N \frac{p(x_i\|\theta)^\gamma}{\{\int p(x\|\theta)^{1+\gamma}dx\}^{\frac{\gamma}{1+\gamma}}}$	$-\frac{1}{N}\frac{\gamma+1}{\gamma}\sum_{i=1}^N \frac{p(y_i\|x_i,\theta)^\gamma}{\{\int p(y\|x_i,\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}$

Theoretical analysis

An influence function (IF) represents relative bias of an estimated static caused by outliers. [3]

empirical distribution: $G(x) = \frac{1}{n}\sum_{i=1}^n \delta(x, x_i)$

contaminated version of G at z: $G_{\varepsilon,z}(x) = (1-\varepsilon)G(x) + \varepsilon\delta(x,z)$

- ε :contamination proportion

For a static T and empirical distribution G , IF at point z is defined as:

$$\mathrm{IF}(z,T,G) = \left.\frac{\partial}{\partial\varepsilon}T\left(G_{\varepsilon,z}(x)\right)\right|_{\varepsilon=0} = \lim_{\varepsilon\to 0}\frac{T\left(G_{\varepsilon,z}(x)\right) - T\left(G(x)\right)}{\varepsilon}$$

How to use IF?

1. Investigate whether $\sup_z|\mathrm{IF}(z,m,G)| < \infty$ or not.
   (If it diverges, the model can be sensitive to small contamination of data.)
2. How much the predictive distribution is affected by outliers

$$\frac{\partial}{\partial\epsilon}\mathbb{E}_{q^*(\theta)}\left[p(x_{\mathrm{test}}|\theta)\right] = \frac{\partial\mathbb{E}_{q^*(\theta)}\left[p(x_{\mathrm{test}}|\theta)\right]}{\partial m}\frac{\partial m^*\left(G_{\varepsilon,z}(x)\right)}{\partial\varepsilon}$$

For VI, if an approximate posterior is describes as $q(\theta;m)$, the IF is given as

- For usual VI,

$$\frac{\partial m^*(G_{\varepsilon,z}(x))}{\partial\varepsilon} = \left(\frac{\partial^2 L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta)) + N\ln p\left(z|\theta\right)\right]$$

- For β-VI,

$$\frac{\partial m^*(G_{\varepsilon,z}(x))}{\partial\varepsilon} = \left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta)) + N\frac{\beta+1}{\beta}p(z|\theta)^\beta - \int p(x|\theta)^{1+\beta}dx\right]$$

- Consider regression and logistic regression for Bayesian neural networks

Input related outlier : $x_o \not\sim p^*(x)$ / Output related outlier : $y_o \not\sim p^*(y|x)$

Behavior of $\sup_z|\mathrm{IF}(z,W,G)|$ for the Bayesian neural network

Activation function	Regression	β- and γ-Regression	Classification	β- and γ-Classification
Linear	$(x_o:U, y_o:U)$	$(x_o:B, y_o:B)$	$(x_o:U)$	$(x_o:B)$
ReLU	$(x_o:U, y_o:U)$	$(x_o:B, y_o:B)$	$(x_o:U)$	$(x_o:B)$
tanh	$(x_o:B, y_o:U)$	$(x_o:B, y_o:B)$	$(x_o:B)$	$(x_o:B)$

- "Regression" and "Classification"indicate the cases of ordinary VI, while " β- and γ-Regression or Classification" is proposed methods.
- $(x_o:U, y_o:U)$ means that IF is unbounded while $(x_o:B, y_o:U)$ means that IF is bounded for input related outliers, but unbounded for output related outliers.

IF of our proposing method is always bounded.

Experiments on UCI datasets

- We compare the performance of our proposed robust variational inference on UCI datasets with an existing robust method or variational inference method.
  - Neural net which has two hidden layers each with 20 units and the ReLU activation function.
  - For solving the optimization problem, we used the re-parameterization trick with 10 Monte Carlo samples.
  - We can determine β or γ by cross-validation. (from 0.1 to 0.9 for the experiment. We found that range from 0.1 to 0.5 is enough.)
  - We added outliers to training data with proportion increased from 0% to 20%.

Dataset	Outliers	KL(G)	KL(St)	WL	Réyni	BB-α	β	γ
concrete	0%	7.46(0.34)	7.36(0.4)	8.04(1.01)	7.16(0.39)	7.18(0.30)	7.27(0.28)	**5.53(0.48)**
N=1030	10%	8.58(0.46)	7.63(0.52)	10.37(1.16)	8.04(0.43)	7.37(0.38)	7.58(0.25)	**6.20(0.74)**
D=8	20%	9.40(1.01)	8.37(0.70)	11.46(0.93)	8.63(0.52)	7.81(0.51)	8.50(0.87)	**6.85(1.15)**
powerplant	0%	4.49(0.15)	4.46(0.16)	4.46(0.18)	4.49(0.14)	4.41(0.13)	4.36(0.11)	**4.28(0.14)**
N=9568	10%	4.71(0.17)	4.59(0.15)	4.81(0.23)	4.66(0.19)	4.56(0.17)	4.41(0.16)	**4.33(0.15)**
D=4	20%	5.12(0.26)	4.65(0.10)	5.04(0.25)	4.82(0.23)	4.70(0.13)	4.52(0.15)	**4.38(0.15)**
protein	0%	5.88(0.50)	**4.78(0.07)**	5.77(0.56)	4.82(0.04)	4.81(0.04)	4.87(0.05)	**4.78(0.05)**
N=45730	10%	6.14(0.03)	**4.84(0.06)**	6.14(0.028)	4.88(0.04)	4.86(0.04)	4.96(0.06)	4.86(0.07)
D=9	20%	6.14(0.03)	4.90(0.08)	6.14(0.031)	4.90(0.05)	**4.86(0.05)**	4.97(0.06)	**4.86(0.07)**

Main References

[1] Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. Biometrika, 85(3):549–559, 1998. ISSN 00063444.
[2] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis, 99(9):2053 – 2081, 2008. ISSN 0047-259X.
[3] P.J. Huber and E.M. Ronchetti. Robust Statistics. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118210338.
[4] Jun Zhu, Ning Chen, and Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. Journal of Machine Learning Research, 15:1799–1847, 2014.

東京大学 THE UNIVERSITY OF TOKYO RIKEN