Variational Inference based on Robust Divergences

Futoshi Futami[12], Issei Sato[12], Masashi Sugiyama[21]
[1]The University of Tokyo, [2]RIKEN
{futami@ms., sato@, sugi@}k.u-tokyo.ac.jp

Abstract

Robustness to outliers is a central issue in real-world machine learning applications. While replacing a model to a heavy-tailed one (e.g., from Gaussian to Student-t) is a standard approach for robustification, it can only be applied to simple models. In this paper, based on Zellner's optimization and variational formulation of Bayesian inference, we propose an outlier-robust pseudo-Bayesian variational method by replacing the Kullback-Leibler divergence used for data fitting to a robust divergence such as the β- and γ-divergences. An advantage of our approach is that complex models such as deep networks can be handled. We theoretically prove that, for deep networks with ReLU activation functions, the *influence function* in our proposed method is bounded, while it is unbounded in the ordinary variational inference. This implies that our proposed method is robust to both input and output outliers, while the ordinary variational method is not.

1 Introduction

Robustness to outliers is becoming more important these days since recent advances in sensor technology give a vast amount of data with spiky noise and crowd-annotated data is full of human errors. A standard approach to robust machine learning is a *model-based* method, which uses a heavier-tailed distribution such as the Student-t distribution instead of the Gaussian distribution as a likelihood function (Murphy [2012]). However, as pointed out in Wang et al. [2017], the model-based method is applicable only to simple modeling setup.

To handle more complex models, we employ the optimization and variational formulation of Bayesian inference by Zellner [1988]. In this formulation, the posterior model is optimized to fit data under the Kullback-Leibler (KL) divergence, while it is regularized to be close to the prior. In this paper, we propose replacing the KL divergence for data fitting to a robust divergence, such as the β-divergence (Basu et al. [1998]) and the γ-divergence (Fujisawa and Eguchi [2008]).

Another robust Bayesian inference method proposed by Ghosh and Basu [2016], follows a similar line to our method, which adopts the β-divergence for pseudo-Bayesian inference. They rigorously analyzed the statistical efficiency and robustness of the method, and numerically illustrated its behavior for the Gaussian distribution. Our work can be regarded as an extension of their work to variational inference so that more complex models such as deep networks can be handled.

2 Robust divergence minimization and Bayesian inference

Let us consider the problem of estimating an unknown probability distribution $p^*(x)$ from its independent samples $x_{1:N} = \{x_i\}_{i=1}^N$. In maximum likelihood estimation, we minimize the generalization error measured by the KL divergence D_{KL} from $p^*(x)$ to a parametric model $p(x;\theta)$ with parameter θ. Since $p^*(x)$ is unknown in practice, we approximate it by empirical distribution $\hat{p}(x) = \frac{1}{N}\sum_{i=1}^N \delta(x, x_i)$, where δ is the Dirac delta function. It is well known that maximum likelihood estimation is sensitive to outliers because it treats all data points equally. To circumvent

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

this problem, outlier robust divergence estimation has been developed in statistics. The *density power divergence*, which is also known as the β-divergence, is a vital example (Basu et al. [1998]). The β-divergence from function g to f is defined as

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

The γ-divergence (Fujisawa and Eguchi [2008]) is another family of robust divergences:

$$D_\gamma\left(g\|f\right) = \frac{1}{\gamma(1+\gamma)}\ln\int g(x)^{1+\gamma}dx - \frac{1}{\gamma}\ln\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\ln\int f(x)^{1+\gamma}dx. \tag{2}$$

Similarly to maximum likelihood estimation, minimizing the β-divergence (or the γ-divergence) from empirical distribution $\hat{p}(x)$ to $p(x;\theta)$ gives an empirical estimator: $\arg\min_\theta D_\beta\left(\hat{p}(x)\|p(x;\theta)\right)$. This yields $0 = \frac{1}{N}\sum_{i=1}^N p(x_i;\theta)^\beta\partial_\theta \ln p(x_i;\theta) - \mathbb{E}_{p(x;\theta)}\left[p(x;\theta)^\beta\partial_\theta\ln p(x_i;\theta)\right]$, where the second term assures the unbiasedness of the estimator. The first term is the likelihood weighted according to the power of the probability for each data point. Since the probabilities of outliers are usually much smaller than those of inliers, those weights effectively suppress the likelihood of outliers. When $\beta = 0$, all weights become one and thus this estimator is reduced to the maximum likelihood estimator. Therefore, adjusting β corresponds to controlling the trade-off between robustness and efficiency. See Appendices A and B for more details.

On the other hand, in Bayesian inference, parameter θ is regarded as a random variable, having prior distribution $p(\theta)$. With Bayes' theorem, the Bayesian posterior distribution $p(\theta|x_{1:N})$ can be obtained as $p(\theta|x_{1:N}) = \frac{p(x_{1:N}|\theta)p(\theta)}{p(x_{1:N})}$. Zellner [1988] showed that $p(\theta|x_{1:N})$ can also be obtained by solving $\arg\min_{q(\theta)\in\mathcal{P}} L(q(\theta))$, where $\mathcal{P}$ is the set of all probability distributions,

$$L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right), \tag{3}$$

and $d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)$ denotes the *cross-entropy*, $d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right) = -\frac{1}{N}\sum_{i=1}^N \ln p(x_i|\theta)$. In practice, this optimization problem is often intractable analytically, and thus we need to use some approximation method. A popular approach is to restrict the domain of the optimization problem to analytically tractable probability distributions $\mathcal{Q}$. Let us denote such a tractable distribution as $q(\theta;\lambda) \in \mathcal{Q}$, where λ is a parameter. Then the optimization problem is expressed as $\arg\min_{q(\theta;\lambda)\in\mathcal{Q}} L(q(\theta;\lambda))$. This optimization problem is called *variational inference* (VI) and $-L(q(\theta))$ is called *evidence lower-bound* (ELBO).

3 Robust Variational Inference based on Robust Divergences

As detailed in Appendix C, Zellner's optimization problem can be equivalently expressed as

$$\arg\min_{q(\theta)\in\mathcal{P}} \mathbb{E}_{q(\theta)}[D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)] + \frac{1}{N}D_{\mathrm{KL}}\left(q(\theta)\|p(\theta)\right). \tag{4}$$

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

$$\arg\min_{q(\theta)\in\mathcal{P}} \mathbb{E}_{q(\theta)}[D_\beta\left(\hat{p}(x)\|p(x|\theta)\right)] + \frac{1}{N}D_{\mathrm{KL}}\left(q(\theta)\|p(\theta)\right). \tag{5}$$

Note that Eq.(5) can be equivalently expressed as $\arg\min_{q(\theta)\in\mathcal{P}} L_\beta(q(\theta))$, where $-L_\beta(q(\theta))$ is the β-ELBO defined as

$$L_\beta(q(\theta) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) - \int q(\theta)\left(-N d_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right), \tag{6}$$

and $d_\beta(\hat{p}(x)\|p(x|\theta))$ denotes the β-cross-entropy: $d_\beta(\hat{p}(x)\|p(x|\theta)) = -\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^N p(x_i|\theta)^\beta + \int p(x|\theta)^{1+\beta}dx$. The optimal solution is given by $q^*(\theta) = \frac{e^{-N d_\beta(\hat{p}(x)\|p(x|\theta))}p(\theta)}{\int e^{-N d_\beta(\hat{p}(x)\|p(x|\theta))}p(\theta)d\theta}$. Interestingly, the

above expression of $q^*(\theta)$ is the same as the *pseudo posterior* proposed in Ghosh and Basu [2016]. Although the pseudo posterior is not equivalent to the *posterior distribution* derived by Bayes' theorem, the spirit of updating prior information by observed data is inherited (Ghosh and Basu [2016]). We discuss how prior information is updated in the pseudo-Bayes-posterior in Appendix E.

Since our proposed optimization problem is generally intractable, following the same line as the discussion in standard approximate Bayesian inference, let us restrict the set of all probability distributions to a set of analytically tractable parametric distributions, $q(\theta;\lambda) \in \mathcal{Q}$. Then the optimization problem yields $\arg\min_{q(\theta;\lambda)\in\mathcal{Q}} L_\beta(q(\theta;\lambda))$. We call this method β-*variational inference* (β-VI). We optimize objective function L_β by using the re-parameterization trick (Kingma and Welling [2013], Ranganath et al. [2014]). So far, we focused on the unsupervised learning case and the β-divergence. Actually, we can easily generalize the above discussion to the supervised learning case and also to the γ-divergence, by simply replacing the cross-entropy with a corresponding one shown in Appendix F.

4 Influence Function Analysis

We analyze the robustness of our proposed method based on the *influence function* (IF), which have been used in robust statistics to study how much contamination affects estimated statistics. We briefly review the definition of IF. Let G be an empirical distribution of $\{x_i\}_{i=1}^n$: $G(x) = \frac{1}{n}\sum_{i=1}^n \delta(x, x_i)$. Let $G_{\varepsilon,z}$ be a contaminated version of G at z: $G_{\varepsilon,z}(x) = (1-\varepsilon)G(x) + \varepsilon\delta(x,z)$, where ε is a contamination proportion. For a statistic T and cumulative distribution G, IF at point z is defined as follows (Huber and Ronchetti [2011]):

$$\mathrm{IF}(z,T,G) = \frac{\partial}{\partial\varepsilon} T(G_{\varepsilon,z}(x))\Big|_{\varepsilon=0} = \lim_{\varepsilon\to 0}\frac{T(G_{\varepsilon,z}(x)) - T(G(x))}{\varepsilon}. \tag{7}$$

Intuitively, IF is a relative bias of a statistic caused by contamination at z.

Now we analyze how posterior distributions derived by VI are affected by contamination. In ordinary VI, we derive a posterior by minimizing Eq.(3). Let us consider an approximate posterior as $q(\theta;m)$ which is parametrized by m. Therefore the objective function given by Eq.(3) can be regarded as a function of m. The first-order optimality condition yields $0 = \frac{\partial}{\partial m} L\Big|_{m=m^*}$. For notational simplicity, we denote $q(\theta;m^*)$ by $q^*(\theta)$.

In Eq.(7), T corresponds to m^*, and G is approximated empirically by the training dataset in VI. Based on this expressions, we can derive the IF of ordinary VI and β-VI in the following (proof is available in Appendix H):

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

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

Using these expressions, we analyze how estimated variational parameters can be perturbed by outliers. In practice, it is important to calculate $\sup_z|\mathrm{IF}(z,\theta,G)|$, because if it diverges, the model can be sensitive to small contamination of data.

In our analysis, we consider two types of outliers—outliers related to input x and outliers related to output y. For true data generating distributions $p^*(x)$ and $p^*(y|x)$, input-related outlier x_o does not obey $p^*(x)$ and output-related outlier y_o does not obey $p^*(y|x)$. Below we investigate whether such outlier-related terms are bounded even when $x_\mathrm{o}\to\infty$ or $y_\mathrm{o}\to\infty$.

As models, we consider neural network models for regression and classification (logistic regression). In neural networks, there are parameters $\theta = \{W, b\}$ where outputs of hidden units are calculated by multiplying W to input and then adding b. Our analysis shows that $\sup_z|\mathrm{IF}(z,b,G)|$ is always bounded (see Appendix K for details), and the result for $\mathrm{IF}(z,W,G)$ is summarized in Table 1.

From Table 1, we can confirm that ordinary VI is always non-robust to output-related outliers. As for input-related outliers, ordinary VI is robust for the "tanh"-activation function, but not for the ReLU

Table 1: Behavior of $\sup_z |\mathrm{IF}(z, W, G)|$ in neural networks, "Regression" and "Classification" indicate the cases of ordinary VI, while "β- and γ-Regression or Classification" mean that we used β-VI or γ-VI. "Activation function" means the type of activation functions used. "Linear" means that there is no nonlinear transformation, inputs are just multiplied W and added b. $(x_\mathrm{o} : U, y_\mathrm{o} : U)$ means that IF is unbounded while $(x_\mathrm{o} : B, y_\mathrm{o} : U)$ means that IF is bounded for input related outliers, but unbounded for output related outliers.

Activation function	Regression	β- and γ-Regression	Classification	β- and γ-Classification
Linear	$(x_\mathrm{o} : U, y_\mathrm{o} : U)$	$(x_\mathrm{o} : B, y_\mathrm{o} : B)$	$(x_\mathrm{o} : U)$	$(x_\mathrm{o} : B)$
ReLU	$(x_\mathrm{o} : U, y_\mathrm{o} : U)$	$(x_\mathrm{o} : B, y_\mathrm{o} : B)$	$(x_\mathrm{o} : U)$	$(x_\mathrm{o} : B)$
tanh	$(x_\mathrm{o} : B, y_\mathrm{o} : U)$	$(x_\mathrm{o} : B, y_\mathrm{o} : B)$	$(x_\mathrm{o} : B)$	$(x_\mathrm{o} : B)$

Table 2: Regression results in RMSE

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)**

and no activation functions. On the other hand, IFs of our proposed method are bounded for all three activation functions including ReLU. We have further conducted IF analysis for Student-t likelihood, which is summarized in Appendix K.

Actually, what we really want to know is a *predictive distribution* at test point x_test. Therefore, it is important to investigate how the predictive distribution is affected by outliers. We can analyze the influence of outliers on the predictive distributions by using IFs of the posterior distribution:

$$\frac{\partial}{\partial \epsilon} \mathbb{E}_{q^*(\theta)} \left[p(x_\mathrm{test}|\theta) \right] = \frac{\partial \mathbb{E}_{q^*(\theta)} \left[p(x_\mathrm{test}|\theta) \right]}{\partial m} \frac{\partial m^* \left(G_{\varepsilon,z}(x) \right)}{\partial \varepsilon}, \tag{8}$$

where $\frac{\partial m^*(G_{\varepsilon,z}(x))}{\partial \varepsilon}$ can be analyzed with the IFs derived above. Since analytical discussion on this expression is difficult, we numerically examined this value in Appendix M.

5 Experiments

We studied our proposed method on the UCI benchmark datasets. We considered both regression and classification problems and used a neural net which has two hidden layers with each 20 units and the ReLU activation function. Detailed experimental setups can be found in Appendix M. We chose β and γ by cross validation. The regression results are shown in Table 2 (the classification results are shown in Appendix M). KL(G) means the Gaussian likelihood (Ordinal VI), KL(St) is Student-t likelihood, WL means the method proposed in Wang et al. [2017] and Réyni is Réyni divergence minimization proposed in Li and Turner [2016] and BB-α is black box alpha divergence minimization proposed in Hernández-Lobato et al. [2016] and Li and Gal [2017]. Our method compares favorably with ordinary VI and existing robust methods for all the datasets.

6 Conclusions

In this work, we proposed outlier-robust variational inference based on robust divergences which allows us to robustify variational inference without changing models. We also compared our proposed method and ordinary variational inference by using the influence function. By using the influence function, we can evaluate how much outliers affect our predictions. Our analysis showed that influence by outliers are bounded in our model, but unbounded by the ordinary variational inference in many cases. Further, experiments showed that our method is robust for both input and output related outliers in both regression and classification setting. In addition, our method outperformed the ordinary VI and existing robust methods on benchmark datasets.

Acknowledgement

FF acknowledges support by JST CREST JPMJCR1403 and MS acknowledges support by KAKENHI 17H00757.

References

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. URL http://www.jstor.org/stable/2337385.

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. doi: https://doi.org/10.1016/j.jmva.2008.02.004. URL http://www.sciencedirect.com/science/article/pii/S0047259X08000456.

Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, Apr 2016. ISSN 1572-9052. doi: 10.1007/s10463-014-0499-0. URL https://doi.org/10.1007/s10463-014-0499-0.

José Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Hernández-Lobato, Thang Bui, and Richard Eric Turner. Black-box α-divergence minimization. 2016.

P.J. Huber and E.M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9781118210338. URL https://books.google.co.jp/books?id=j10hquR_j88C.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13.

Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. *arXiv preprint arXiv:1703.02914*, 2017.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

Kevin P Murphy. Machine learning: a probabilistic perspective. 2012.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with Bayesian data reweighting. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3646–3655, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/wang17g.html.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42 (4):278–280, 1988. ISSN 00031305. URL http://www.jstor.org/stable/2685143.

A γ divergence minimization

A.1 Unsupervised setting

In this section, we explain the γ divergence minimization for unsupervised setting. We denote true distribution as $p^*(x)$. We denote the model by $p(x;\theta)$. We minimize the following γ cross entropy,

$$d_\gamma(p^*(x), p(x;\theta)) = -\frac{1}{\gamma}\ln\int p^*(x)p(x;\theta)^\gamma dx + \frac{1}{1+\gamma}\ln\int p(x;\theta)^{1+\gamma}dx. \tag{9}$$

This is empirically approximated as

$$L_n(\theta) = d_\gamma(\hat{p}(x), p(x;\theta)) = -\frac{1}{\gamma}\ln\frac{1}{n}\sum_{i=1}^{n} p(x_i;\theta)^\gamma dx + \frac{1}{1+\gamma}\ln\int p(x;\theta)^{1+\gamma}dx. \tag{10}$$

By minimizing $L_n(\theta)$, we can obtain following estimation equation,

$$0 = -\frac{\sum_{i=1}^n p(x_i;\theta)^\gamma \frac{\partial}{\partial\theta}\ln p(x_i;\theta)}{\sum_{i=1}^n p(x_i;\theta)^\gamma} + \int \frac{p(x;\theta)^{1+\gamma}}{\int p(x;\theta)^{1+\gamma}dx}\frac{\partial}{\partial\theta}\ln p(x;\theta)dx. \tag{11}$$

This is actually weighted likelihood equation, where the weights are $\frac{p(x_i;\theta)^\gamma}{\sum_{i=1}^n p(x_i;\theta)^\gamma}$. The second term is for the unbiasedness of the estimating equation.

A.2 Supervised setting

In this section, we explain the γ divergence minimization for the supervised setting. We denote the true distribution as $p^*(y,x) = p^*(y|x)p^*(x)$. We denote the regression model by $p(y|x;\theta)$. What we minimize is following γ cross entropy over the distribution $p^*(x)$,

$$\begin{aligned} d_\gamma(p^*(y|x), p(y|x;\theta)|p^*(x)) = -\frac{1}{\gamma}\ln\int\left\{\int p^*(y|x)p(y|x;\theta)^\gamma dy\right\}p^*(x)dx + \\ \frac{1}{1+\gamma}\ln\int\left\{\int p(y|x;\theta)^{1+\gamma}dy\right\}p^*(x)dx. \end{aligned} \tag{12}$$

This is empirically approximated as

$$L_n(\theta) = d_\gamma(\hat{p}(y|x), p(y|x;\theta)|\hat{p}(x)) = -\frac{1}{\gamma}\ln\left\{\frac{1}{n}\sum_{i=1}^n p(y_i|x_i;\theta)^\gamma\right\} + \frac{1}{1+\gamma}\ln\left\{\frac{1}{n}\sum_{i=1}^n\int p(y|x_i;\theta)^{1+\gamma}dy\right\}. \tag{13}$$

By minimizing $L_n(\theta)$, we can obtain following estimation equation.

$$0 = -\frac{\sum_{i=1}^n p(y_i|x_i;\theta)^\gamma \frac{\partial}{\partial\theta}\ln p(y_i|x_i;\theta)}{\sum_{i=1}^n p(y_i|x_i;\theta)^\gamma} + \frac{\sum_{i=1}^n\int p(y|x_i;\theta)^{1+\gamma}\frac{\partial}{\partial\theta}\ln p(y_i|x_i;\theta)dy}{\sum_{i=1}^n\int p(y|x_i;\theta)^{1+\gamma}}. \tag{14}$$

Actually, minimizing $L_n(\theta)$ is equivalent to minimizing following expression,

$$L'_n(\theta) = -\frac{\gamma+1}{\gamma}\frac{1}{n}\sum_{i=1}^n \frac{p(y_i|x_i;\theta)^\gamma}{\left\{\int p(y|x_i;\theta)^{1+\gamma}dy\right\}^{\frac{\gamma}{1+\gamma}}}. \tag{15}$$

As $\gamma \to 0$, above expression goes to

$$L'_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \ln p(y_i|x_i;\theta). \tag{16}$$

This is usual KL cross entropy. In the main paper, we use $L'_n(\theta)$ as γ cross entropy instead of using original expression. The reason is given in Appendix J.

B β divergence minimization

Until now, we focused on γ divergence minimization. We can also consider supervised setup for β divergence minimization. The empirical approximation of β cross entropy minimization is given by,

$$L_n(\theta) = d_\beta(\hat{p}(y|x), p(y|x;\theta)|\hat{p}(x)) = -\frac{\beta+1}{\beta}\left\{\frac{1}{n}\sum_{i=1}^n p(y_i|x_i;\theta)^\beta\right\} + \left\{\frac{1}{n}\sum_{i=1}^n\int p(y|x_i;\theta)^{1+\beta}dy\right\}. \tag{17}$$

For comparison of unsupervised and supervised setting, we show the empirical approximation of β cross entropy for unsupervised setting,

$$L_n(\theta) = d_\beta(\hat{p}(x), p(x;\theta)) = -\frac{\beta+1}{\beta}\frac{1}{n}\sum_{i=1}^n p(x_i;\theta)^\beta + \int p(x;\theta)^{1+\beta}dx. \tag{18}$$

C Proof of Eq.(4) in the main paper

From the definition of KL divergence, the cross entropy can be expressed as

$$d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right) = D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right) + \mathrm{Const.} \tag{19}$$

By substituting the above expression into the definition of $L(q(\theta))$, we obtain

$$L(q(\theta)) = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) + N\mathbb{E}_{q(\theta)}[D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)] + \mathrm{Const.}$$

What we have to consider is

$$\arg\min_{q(\theta)\in\mathcal{P}} L(q(\theta)), \tag{20}$$

We can disregard the constant term in $L(q(\theta))$, and above optimization problem is equivalent to

$$\arg\min_{q(\theta)\in\mathcal{P}} \frac{1}{N} L(q(\theta)). \tag{21}$$

Therefore Eq.(4) is equivalent to Eq.(20)

D Derivation of Pseudo posterior

In this section, we derive the pseudo posterior in the main text. The objective function is given as

$$L_\beta = \mathbb{E}_{q(\theta)}[D_\beta\left(\hat{p}(x)||p(x|\theta)\right)] + \lambda' D_{KL}\left(q(\theta)||p(\theta)\right) \tag{22}$$

where λ' is the regularization constant. We optimize this with the constraint that $\int q(\theta)d\theta = 1$. We calculate using the method of variations and Lagrange multipliers, we can get the optimal $q(\theta)$ in the following way,

$$\frac{d(L_\beta + \lambda(\int q(\theta)d\theta - 1))}{dq(\theta)} = D_\beta\left(\hat{p}(x)|p(x|\theta)\right)] + \lambda' \ln\frac{q(\theta)}{p(\theta)} - (1+\lambda) = 0 \tag{23}$$

By rearranging the above expression, we can get the following relation,

$$q(\theta) \propto p(\theta) e^{-\frac{1}{\lambda'} d_\beta(\hat{p}(x)|p(x|\theta))} \tag{24}$$

If we set $\frac{1}{\lambda'} = N$ and normalize the above expression, we get the Theorem **??** in the main text,

$$q(\theta) = \frac{e^{-N d_\beta(\hat{p}(x)|p(x|\theta))} p(\theta)}{\int e^{-N d_\beta(\hat{p}(x)|p(x|\theta))} p(\theta) d\theta}. \tag{25}$$

We can get the similar expression for γ cross entropy.

Interestingly, if we use KL cross entropy instead of β cross entropy, following relation holds,

$$\begin{aligned} q(\theta) \propto p(\theta) e^{-\frac{1}{\lambda'} d_{KL}(\hat{p}(x)|p(x|\theta))} &= p(\theta) e^{-N(-\frac{1}{N}\sum_i \ln p(x_i|\theta))} \\ &= p(\theta)\prod_i p(x_i|\theta) \\ &= p(\theta)p(D|\theta) \end{aligned} \tag{26}$$

The normalizing constant is

$$\int p(\theta)\prod_i p(x_i|\theta) d\theta = p(D) \tag{27}$$

Finally, we get the optimal $q(\theta)$

$$q(\theta) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{28}$$

This is the posterior distribution which can be derived by Bayes theorem.

In the above proof, we set regularization constant as $\frac{1}{\lambda'} = N$ to derive the expression. Choosing appropriate regularization constant is difficult in this case. However, as far as we did experiment, the impact of choosing regularization constant to the performance is small compared to the effect of choosing the appropriate β or γ. Therefore, in this paper we only consider the situation that regularization constant is $\frac{1}{\lambda'} = N$. However how to choose the regularization constant should be studied further in the future because which reflects the trade off between prior information and information from data.

Table 3: Cross-entropies for robust variational inference.

	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	$-\frac{1}{N}\frac{\gamma+1}{\gamma}\sum_{i=1}^{N}\frac{p(x_i\|\theta)^\gamma}{\{\int p(x\|\theta)^{1+\gamma}dx\}^{\frac{\gamma}{1+\gamma}}}$	$-\frac{1}{N}\frac{\gamma+1}{\gamma}\sum_{i=1}^{N}\frac{p(y_i\|x_i,\theta)^\gamma}{\{\int p(y\|x_i,\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}$

E Pseudo posterior

The expression Eq.(25) is called pseudo posterior in statistics. In general, pseudo posterior is given as

$$q(\theta) = \frac{e^{-\lambda R(\theta)}p(\theta)}{\int e^{-\lambda R(\theta)}p(\theta)d\theta}. \tag{29}$$

where $p(\theta)$ is prior and $R(\theta)$ expresses empirical risk not restricted to likelihood and not necessarily additive. The is also called Gibbs posterior and extensively studied in the field of PAC Bayes. Our β cross entropy based pseudo posterior is

$$\begin{aligned} q(\theta) &\propto e^{-N\left\{\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^{N} p(x_i;\theta)^\beta + \int p(x;\theta)^{1+\beta}dx\right\}}p(\theta) \\ &= \left[\prod_i^N e^{l_\theta(x_i)}p(\theta)\right] \end{aligned} \tag{30}$$

where $l_\theta(x_i) = \frac{\beta+1}{\beta}p(x_i;\theta)^\beta - \frac{1}{N}\int p(x;\theta)^{1+\beta}dx$.

As discussed in Ghosh and Basu (2016), we can understand the intuitive meaning of above expression by comparing this expression with Eq.(26). In usual Bayes posterior, the prior belief is updated by likelihood $p(x_i|\theta)$ which represents the information from data x_i as shown in Eq.(26). On the other hand, when using β cross entropy, the prior belief is updated by $e^{l_\theta(x_i)}$ which has information about data x_i. Therefore the spirit of Bayes, that is, we update information about parameter based on training data, are inherited to this pseudo posterior.

F Other cross entropies for robust variational inference

Here we summarize the other cross-entropies for robust variational inference in the table 3.

G Influence function

In the main paper, we omit the expression for influence function of supervised version and γ VI. In this section, we list the all expression we derived.

Theorem 1 *When data contamination is given by $G_\varepsilon(x) = (1-\varepsilon)G_n(x) + \varepsilon\delta(x,z)$, IF of ordinary VI is given by*

$$\left(\frac{\partial^2 L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta)) + Nl(z)\right], \tag{31}$$

IF of β-VI is given by

$$\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta)) + Nl_\beta(z)\right], \tag{32}$$

and IF of γ-VI is given by

$$\left(\frac{\partial^2 L_\gamma}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[D_{\mathrm{KL}}(q^*(\theta)\|p(\theta)) + Nl_\gamma(z)\right], \tag{33}$$

where $l(z)$, $l_\beta(z)$, and $l_\gamma(z)$ are defined in Table 4.

Table 4: Influence functions for robust variational inference.

	Unsupervised	Supervised z=(x',y')
$l(z)$	$\ln p\left(z\middle\vert\theta\right)$	$\ln p\left(y'\middle\vert x',\theta\right)$
$l_\beta(z)$	$\frac{\beta+1}{\beta}p(z\vert\theta)^\beta - \int p(x\vert\theta)^{1+\beta}dx$	$\frac{\beta+1}{\beta}p(y'\vert x',\theta)^\beta - \int p(y\vert x',\theta)^{1+\beta}dy$
$l_\gamma(z)$	$\frac{\gamma+1}{\gamma}\frac{p(z\vert\theta)^\gamma}{\{\int p(x\vert\theta)^{1+\gamma}dx\}^{\frac{\gamma}{1+\gamma}}}$	$\frac{\gamma+1}{\gamma}\frac{p(y'\vert x',\theta)^\gamma}{\{\int p(y\vert x',\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}$

H Proof of Theorem 1

We consider the situation where the distribution is expressed as

$$G_\varepsilon\left(x\right) = \left(1-\varepsilon\right)G_n\left(x\right) + \varepsilon\delta(x,z) \tag{34}$$

H.1 Derivation of IF for usual VI

We start from the first order condition,

$$\begin{aligned} 0 =& \left.\frac{\partial}{\partial m}L\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x)\ln p(x|\theta) + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right] \end{aligned} \tag{35}$$

We differentiate above expression with ϵ, then we obtain following expression,

$$\begin{aligned} 0 =& \nabla_m \int d\theta \frac{\partial m^*(\epsilon)}{\partial \epsilon}\frac{\partial q}{\partial m^*(\epsilon)}\left\{(1-\epsilon)N\int dG_n(x)\ln p(x|\theta) + \epsilon N \ln p(z|\theta) + \ln p(\theta)\right\} \\ &+ \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[-N\int dG_n(x)\ln p(x|\theta) + N\ln p(z|\theta)\right] \\ &- \nabla_m \int d\theta \frac{\partial m^*(\epsilon)}{\partial \epsilon}\frac{\partial q}{\partial m^*(\epsilon)}\ln q(\theta;m^*(\epsilon)) - \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[\frac{\partial m^*(\epsilon)}{\partial \epsilon}.\frac{\partial \ln q}{\partial m^*(\epsilon)}\right] \end{aligned} \tag{36}$$

From above expression, if we take $\epsilon \to 0$, we soon obtain following expression,

$$\frac{\partial m^*\left(\varepsilon\right)}{\partial \varepsilon} = -\left(\frac{\partial^2 L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\int dG_n(x)\ln p\left(x\middle\vert\theta\right) - N\ln p\left(y\middle\vert\theta\right)\right]. \tag{37}$$

Actually, this can be transformed to following expression by using the first order condition,

$$\frac{\partial m^*\left(\varepsilon\right)}{\partial \varepsilon} = \left(\frac{\partial^2 L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[D_{KL}(q(\theta;m)|p(\theta)) + N\ln p\left(z\middle\vert\theta\right)\right]. \tag{38}$$

H.2 Derivation of IF for β VI

Next we consider IF for β VI. To proceed calculation, we have to be careful that empirical approximation of β cross entropy takes different form between unsupervised and supervised setting as shown in Eq.(18) and Eq.(17).

For the unsupervised situation, we can write the first order condition as,

$$\begin{aligned} 0 =& \left.\frac{\partial}{\partial m}L_\beta\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x)\frac{\beta+1}{\beta}p(x|\theta)^\beta - N\int p(x|\theta)^{1+\beta}dx + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right]. \end{aligned} \tag{39}$$

We can proceed calculation in the same way as usual VI. We get the following expression

$$\frac{\partial m^*\left(\varepsilon\right)}{\partial \varepsilon} = -\frac{\beta+1}{\beta}\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\int dG_n(x)p(x|\theta)^\beta - Np\left(z\middle\vert\theta\right)^\beta\right]. \tag{40}$$

Next, we consider the supervised situation. We consider the situation where the contamination is expressed as

$$G_\varepsilon(x,y) = (1-\varepsilon)\, G_n(x,y) + \varepsilon\delta((x,y),(x',y')) \tag{41}$$

The first order condition is,

$$\begin{aligned} 0 =& \left.\frac{\partial}{\partial m} L_\beta\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x,y)\frac{\beta+1}{\beta}p(y|x,\theta)^\beta\right] - \\ & \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x)\left\{\int p(y|x,\theta)^{1+\beta}dy\right\} + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right]. \end{aligned} \tag{42}$$

We can proceed the calculation and derive the influence function as follows,

$$\begin{aligned} \frac{\partial m^*(\varepsilon)}{\partial \varepsilon} =& -N\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\frac{\beta+1}{\beta}\left(\int dG_n(y,x)p(y|x,\theta)^\beta - p\left(y'|x',\theta\right)^\beta\right)\right] \\ & + N\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\int dG_n(x)\left(\int p(y|x,\theta)^{1+\beta}dy\right) - \int p(y|x',\theta)^{1+\beta}dy\right]. \end{aligned} \tag{43}$$

If we take the limit β to 0, the above expression reduced to IF of usual VI.

H.3 Derivation of IF for γ VI

We can derive IF for γ VI in the same way as β VI.

For simplicity, we focus on the transformed cross entropy, which is given Eq.(16). For unsupervised situation, the first order condition is given by,

$$\begin{aligned} 0 =& \left.\frac{\partial}{\partial m} L_\gamma\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x)\frac{p(x|\theta)^\gamma}{\left\{\int p(x|\theta)^{1+\gamma}dx\right\}^{\frac{\gamma}{1+\gamma}}} + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right]. \end{aligned} \tag{44}$$

In the same way as β VI, we can get the IF of γ VI for unsupervised setting as,

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\frac{\int dG_n(x)p(x|\theta)^\gamma - p(z|\theta)^\gamma}{\left\{\int p(x|\theta)^{1+\gamma}dx\right\}^{\frac{\gamma}{1+\gamma}}}\right]. \tag{45}$$

For supervised situation, the first order condition is give by,

$$\begin{aligned} 0 =& \left.\frac{\partial}{\partial m} L_\beta\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x,y)\frac{p(y|x,\theta)^\gamma}{\left\{\int p(y|x,\theta)^{1+\gamma}dy\right\}^{\frac{\gamma}{1+\gamma}}} + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right]. \end{aligned} \tag{46}$$

In the same way as β VI, we can get the IF of γ VI for supervised setting as,

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -N\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\int dG_n(x,y)\frac{p(y|x,\theta)^\gamma}{\left\{\int p(y|x,\theta)^{1+\gamma}dy\right\}^{\frac{\gamma}{1+\gamma}}} - \frac{p(y'|x',\theta)^\gamma}{\left\{\int p(y|x',\theta)^{1+\gamma}dy\right\}^{\frac{\gamma}{1+\gamma}}}\right]. \tag{47}$$

I Other aspects of analysis based on influence function

Although in the above sections, we consider outliers as contamination given by Eq.(34), we can other type of contamination, such as training data itself is perturbed, that is, a training point $z = (x, y)$ is

perturbed to $z_\epsilon = (x + \epsilon, y)$ which had proposed in Koh and Liang (2017). We call this type of data contamination as data perturbation. As for data perturbation, following relation holds,

When we consider data perturbation for a training data, IF of usual VI is given by

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\frac{\partial}{\partial x}\ln p(z|\theta)\right]. \tag{48}$$

IF of β divergence based VI is given by

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\frac{\partial}{\partial x} p(z|\theta)^\beta\right]. \tag{49}$$

J Another type of γ VI

In the main paper, we used the transformed γ cross entropy, which is given in Eq.(15). The reason we used the transformed cross entropy instead of original expression is that we can interpret the pseudo posterior when using the transformed cross entropy much easily than when using original cross entropy.

In the same way eq.50, we can derive the pseudo posterior using transformed cross entropy,

$$\begin{aligned} q(\theta) &\propto e^{N\frac{\gamma+1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{p(x_i|\theta)^\gamma}{\{\int p(x|\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}} p(\theta) \\ &= \left[\prod_i^N e^{l_\theta(x_i)} p(\theta)\right] \end{aligned} \tag{50}$$

where $l_\theta(x_i) = \frac{\gamma+1}{\gamma}\frac{p(x_i|\theta)^\gamma}{\{\int p(x|\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}$. In this formulation, it is easy to consider that the information of data x_i is utilized to update the prior information through $e^{l_\theta(x_i)}$.

However, when using original cross entropy, such interpretation cannot be done because the pseudo posterior is given by,

$$q(\theta) \propto e^{N(\frac{1}{\gamma}\ln\frac{1}{N}\sum_i^N p(x_i|\theta)^\gamma dx - \frac{1}{1+\gamma}\ln\int p(x|\theta)^{1+\gamma}dx)} p(\theta) \tag{51}$$

and since the summation is not located in the front, this pseudo posterior has not additivity. Therefore it is difficult to understand how each training data x_i contributes to update the parameter. Moreover it is not straight forward to apply stochastic variational inference framework. Accordingly, we decided to use the transformed cross entropy.

Even thought the interpretation is difficult we can dirive IF in the same way as we discussed. For unsupervised situation, the first order condition is given by

$$\begin{aligned} 0 &= \frac{\partial}{\partial m} L_\gamma\Big|_{m=m^*} \\ &= \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[\frac{N}{\gamma}\ln\int dG_\epsilon(x) p(x|\theta)^\gamma dx - \frac{N}{1+\gamma}\ln\int p(x|\theta)^{1+\gamma}dx + \ln p(\theta) - \ln q(\theta; m^*(\epsilon))\right]. \end{aligned} \tag{52}$$

In the same way as β VI, we can get the IF of γ VI of original cross entropy for unsupervised setting as,

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\frac{N}{\gamma}\left(\frac{\partial^2 L_\gamma}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\frac{\int dG_n(x) p(x|\theta)^\gamma - N p(z|\theta)^\gamma}{\int dG_n(x) p(x|\theta)^\gamma}\right]. \tag{53}$$

For supervised situation, we can derive in the same way.

K Discussion of Influence function

In this section, we describe detail discussion of influence function's behavior when using neural net for regression and classification(logistic regression).

We use mean field variational inference and Gaussian distribution for approximate posterior $q(\theta)$. Gaussian distribution is a member of exponential family, we can parametrize it by mean value m. In the case of Gaussian distribution, $m = \{\mathbb{E}[\theta], \mathbb{E}[\theta^2]\}$. We can parametrize variational posterior as $q(\theta|m)$ by using these parameters. It is well known that the estimation of uncertainty of variational posterior is quite poor, therefore we focus on analyzing $\mathbb{E}[\theta]$ and for simplicity, we denote it by m.

Let us start usual variational inference. In Eq.(48), we especially focus on the term, $\frac{\partial}{\partial m}\mathbb{E}_{q(\theta|m)}\left[\ln p\left(y|\theta\right)\right]$, because this is the only term that is related to outlier. If we assume that approximate posterior is an Gaussian distribution, we can transform this term in the following way,

$$\begin{aligned}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta|m)}\left[\ln p\left(y|\theta\right)\right] &= \frac{\partial}{\partial m}\left\{\int q\left(\theta|m\right)\ln p\left(y|\theta\right)d\theta\right\}\\ &= \int \frac{\partial q\left(\theta|m\right)}{\partial m}\ln p\left(y|\theta\right)d\theta\\ &= -\int q\left(\theta|m\right)\frac{\partial}{\partial\theta}\ln p\left(y|\theta\right)d\theta\\ &= -E_{q(\theta|m)}\left[\frac{\partial}{\partial\theta}\ln p\left(y|\theta\right)\right]\end{aligned} \tag{54}$$

, where we used the following relation,

$$\frac{\partial q\left(\theta|m\right)}{\partial m} = \frac{\partial q\left(\theta|m\right)}{\partial\theta} \tag{55}$$

and partial integration for the second line to third line. This kind of transformation can also be carried out where the approximate posterior is Student-T.

From above expression, it is clear that studying the behavior of $\frac{\partial}{\partial\theta}\ln p\left(y|\theta\right)$ is crucial for analyzing IF. In this case, the behavior of IF in this expression is similar to that of maximum likelihood.

K.1 Regression

In this subsection, we consider the regression problem by neural network. We denote the input to the final layer as $f_\theta(x) \sim p(f|x,\theta)$, where x is the input and θ obeys approximate posterior $q(\theta|m)$. We consider the output layer as Gaussian distribution as $p(y|f_\theta(x)) = N(f_\theta(x), I)$. From above discussion, what we have to consider is $\frac{\partial}{\partial\theta}\ln p\left(y|f_\theta(x)\right)$.

We denote input related outlier as x_o, that means x_o does not follow the same distribution as other regular training dataset. Also, we denote the output related outlier as y_o that it does not follow the same observation noise as other training dataset.

Output related outlier

Since we consider the model that output layer is Gaussian distribution, following relation holds for IF of usual VI,

$$\frac{\partial}{\partial\theta}\ln p\left(y_o|f_\theta(x_o)\right) \propto \left(y_o - f_\theta(x_o)\right)\frac{\partial f_\theta(x_o)}{\partial\theta} \tag{56}$$

As for the β divergence, we have to treat Eq.(43). Fortunately, when we use Gaussian distribution for output layer, the second term in the bracket of Eq.(43) will be constant, hence its derivative will be zero. Therefore the output related term is only the first term. Thanks to this property, the denominator of Eq.(47) will also be a constant. Therefore IF of β VI and γ VI behaves in the same way. Therefore, we only consider β VI for regression tasks. By using the same transformation as Eq.(54), following relation holds

$$\begin{aligned}\frac{\partial}{\partial\theta}p\left(y_o|f_\theta(x_o)\right)^\beta &\propto e^{-\frac{\beta}{2}\left(y_o - f_\theta(x_o)\right)^2}\left(y_o - f_\theta(x_o)\right)\frac{\partial f_\theta(x_o)}{\partial\theta}\\ &= \frac{\left(y_o - f_\theta(x_o)\right)}{e^{\frac{\beta}{2}\left(y_o - f_\theta(x_o)\right)^2}}\frac{\partial f_\theta(x_o)}{\partial\theta}\end{aligned} \tag{57}$$

From Eq.(56) and Eq.(57), we can see that IF of usual VI is unbounded as output related outlier become large. On the other hand β VI is bounded. Actually, eq.(57) goes to 0 as $y_o \to \pm\infty$. This means that the influence of this contamination will become zero. This is the desired property for robust estimation.

Input related outlier

Next, we consider input related outlier, that is, we consider whether Eq.(56) and Eq.(57) are bounded or not even when $x_o \to \pm\infty$.

To proceed the analysis, we have to specify models. We start from the most simple case, $f_\theta(x_o) = W_1 x_o + b_1$, where $\theta = \{W_1, b_1\}$. This is the simple linear regression. In this case $\dfrac{\partial f_\theta(x_o)}{\partial W_1} = x_o$ and $\dfrac{\partial f_\theta(x_o)}{\partial b_1} = 1$. When $x_o \to \pm\infty$, $f_\theta(x_o) \to \pm\infty$.

From these fact, we can soon find that Eq.(56) is unbouded. As for Eq.(57), the exponential function in the denominator of eq.(57) plays a crucial role. Thanks to this exponential function,

$$\begin{aligned}\frac{\partial}{\partial W_1} p\left(y_o | f_\theta(x_o)\right)^\beta &\propto \frac{(y_o - f_\theta(x_o))}{e^{\frac{\beta}{2}(y_o - f_\theta(x_o))^2}} x_o \\ &\xrightarrow[x_o \to \infty]{} 0 \end{aligned} \tag{58}$$

From these facts, usual VI is not robust against input related outliers, however β VI is robust.

Next we consider the situation that there is a hidden layer, that is $f_\theta(x_o) = W_2(W_1 x_o + b_1) + b_2$, where $\theta = \{W_1, b_1, W_2, b_2\}$. At this point, we do not consider activation function. Following relations hold,

$$\frac{\partial}{\partial W_1} f_\theta(x_o) = W_2 x_o, \quad \frac{\partial}{\partial W_2} f_\theta(x_o) = W_1 x_o + b_1 \tag{59}$$

From these relations, the behavior of IF in the case of $x_o \to \pm\infty$ is actually as same as the case where there is no hidden layers. Therefore, IF of input related outlier is bounded in β VI and that is unbounded in usual VI. Even if we add more layers the situation does not change in this situation where no activation exists.

Next, we consider the situation that there exists activation function. We consider $relu$ and $tanh$ as activation function. In the situation that there is only one hidden layers, $f_\theta(x_o) = W_2(relu\left(W_1 x_o + b_1\right)) + b_2$,

$$\frac{\partial f_\theta(x_o)}{\partial W_2} = relu\left(W_1 x_o + b_1\right), \quad \frac{\partial f_\theta(x_o)}{\partial W_1} = \begin{cases} W_2 x_o, & W_1 x_o + b_1 \geq 0 \\ 0, & W_1 x_o + b_1 < 0, \end{cases} \tag{60}$$

Actually, this is almost the same situation as above situation where there are no activation functions, because there remains possibility that IF will diverge in usual VI, while IF in β VI is bounded..

In the situation that $f_\theta(x_o) = W_2 tanh\left(W_1 x_o + b_1\right) + b_2$,

$$\frac{\partial f_\theta(x_o)}{\partial W_1} = \frac{W_2 x_o}{cosh^2\left(W_1 x_o + b_1\right)} \xrightarrow[x_o \to \infty]{} 0 \tag{61}$$

The limit of above expression can be easily understand from Fig.1. From this expression, we can understand IF of W_1 is bounded in both usual estimator and β estimator, when we consider the model, $f_\theta(x_o) = tanh\left(W_1 x_o + b_1\right)$. As for W_2,

$$\frac{\partial f_\theta(x_o)}{\partial W_2} = tanh(W_1 x_o + b_1) \tag{62}$$

In this expression, even if input related outlier goes to infinity, the maximum of above expression is 1. Accordingly, the IF of W_2 is bounded in any case.

Up to now, we have seen the model which has a hidden model. The same discussion can be held for the model which has much more hidden layers. If we add layers, above discussion holds and there remains possibility that IF using relu in usual VI will diverge.

We can say that usual VI is not robust to output related outliers and input related outliers. The exception is that using tanh activation function makes the IF bounded. In β VI, the IF of parameters are always bounded.

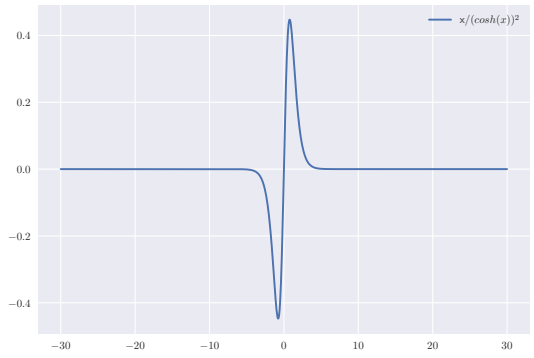


Figure 1: Behavior of $\frac{x}{\cosh^2 x}$

Using Student-T output layer

We additionaly consider the property of Student-t loss in terms of IF. When we denote degree of freedom as ν, and the variance as σ^2, following relation holds,

$$\frac{\partial}{\partial \theta} \ln p\left(y_o | f_\theta(x_o)\right) \propto \frac{\left(y_o - f_\theta(x_o)\right)}{\nu\sigma^2 + \left(y_o - f_\theta(x_o)\right)^2} \frac{\partial f_\theta(x_o)}{\partial \theta} \tag{63}$$

By comparing Eq.(63) with Eq.(56) and Eq.(57), we can confirm that the behavior of IF in the case of Student-t loss in usual VI is similar to Gaussian loss model in β VI. First, consider output related outlier,

$$\frac{\partial}{\partial \theta} \ln p\left(y_o | f_\theta(x_o)\right) \xrightarrow[y_o \to \infty]{} 0 \tag{64}$$

From above expression, we can find that Student-T loss is robust to output related outlier. This is the desiring property of Student-T.

Next consider input related outlier. We consider the model, $f_\theta(x_o) = W_1 x_o + b_1$, where $\theta = \{W_1, b_1\}$

$$\begin{aligned}
\frac{\partial}{\partial W_1} \ln p\left(y_o | f_\theta(x_o)\right) &\propto \frac{\left(y_o - f_\theta(x_o)\right)}{\nu\sigma^2 + \left(y_o - f_\theta(x_o)\right)^2} x_o \\
&= \frac{\left(y_o - f_\theta(x_o)\right)^2}{\nu\sigma^2 + \left(y_o - f_\theta(x_o)\right)^2} \frac{x_o}{y_o - f_\theta(x_o)} \\
&= \frac{\left(y_o - f_\theta(x_o)\right)^2}{\nu\sigma^2 + \left(y_o - f_\theta(x_o)\right)^2} \frac{f_\theta(x_o) - b_1}{W_1(y_o - f_\theta(x_o))} \\
&\xrightarrow[x_o \to \infty]{} -W_1^{-1}
\end{aligned} \tag{65}$$

This is an interesting result that in β VI, the effect of input related outlier goes to 0 in the limit, on the other hand on Student-t loss, the IF is bounded but finite value remains.

Although the finite value remains in IF, the value is W_1, that is considerably small, therefore we can disregard this influence.

K.2 Classification

In this subsection, we consider the classification problem. We focus on binary classification, and output y can take $+1$ or 0. We only consider the input related outlier for the limit discussion because the influence caused by label misspecification is always bounded.

As the model, we consider logistic regression model,

$$p(y|f_\theta(x)) = f_\theta(x)^y(1-f_\theta(x))^{(1-y)} \tag{66}$$

where

$$f_\theta(x) = \frac{1}{1+e^{-g_\theta(x)}} \tag{67}$$

where $g_\theta(x)$ is input to sigmoid function. We consider neural net for $g_\theta(x)$ later.

We first assume $g_\theta(x) = Wx + b$, then $\frac{\partial g}{\partial W} = x$ and $\frac{\partial g}{\partial b} = 1$. We assume prior and posterior distribution of W and b are Gaussian distributions. For IF analysis, we first consider the first term of Eq.(43) and only consider outlier related term inside it. To proceed the calculation, we can use the relation Eq.(54), and what we have to analyze is

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln p(y|f_\theta(x)) &= \frac{\partial}{\partial\theta}\left(y\ln f_\theta(x) + (1-y)\ln(1-f_\theta(x))\right) \\ &= -y(1-f)\frac{\partial g}{\partial\theta} + (1-y)f\frac{\partial g}{\partial\theta}\end{aligned} \tag{68}$$

Let us consider, for example $y = +1$

$$\frac{\partial}{\partial\theta}\ln p(y=+1|f_\theta(x)) = \frac{1}{1+e^{g_\theta(x)}}\frac{\partial g}{\partial\theta} \tag{69}$$

As for $\theta = b$, this is always bounded. As for $\theta = W$,

$$\frac{\partial}{\partial W}\ln p(y=+1|f_\theta(x)) = \frac{1}{1+e^{Wx+b}}x \tag{70}$$

In above expression, if we take limit $x \to +\infty$, and if $Wx \to -\infty$, above expression can diverge. If $Wx \to \infty$ when $x \to +\infty$, above expression goes to 0. From this observation, it is clear that there is a possibility that IF for input related outlier diverges in simple logistic regression for usual VI.

As for β VI, we have to consider the following term,

$$p(y=+1|f_\theta(x))^\beta\frac{\partial}{\partial\theta}\ln p(y=+1|f_\theta(x)) = \frac{1}{(1+e^{-g_\theta(x)})^\beta}\frac{1}{1+e^{g_\theta(x)}}\frac{\partial g}{\partial\theta} \tag{71}$$

This expression converges to 0 when $x_o \to \pm\infty$. In addition, we have to consider the behavior of the second term in Eq.(43) for analysis of IF, which is vanish in the regression situation. The second term of Eq.(43) can be written as

$$\begin{aligned}&\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\int p(y|x_o,\theta)^{1+\beta}dy\right] \\ &= N\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[f_\theta(x_o)^{1+\beta} + (1-f_\theta(x_o))^{1+\beta}\right]\end{aligned} \tag{72}$$

To proceed the analysis, we can use the relation Eq.(54). Since the inverse of hessian matrix is not related to outlier, what we have to consider is

$$\begin{aligned}&\int d\theta q(\theta)\frac{\partial}{\partial\theta}f_\theta(x_o)^{1+\beta} + \frac{\partial}{\partial\theta}(1-f_\theta(x_o))^{1+\beta} \\ &= -\int d\theta q(\theta)\left(f_\theta(x_o)^{1+\beta}(1-f_\theta(x_o))\frac{\partial g}{\partial\theta} + (1-f_\theta(x_o))^{1+\beta}f_\theta(x_o)\frac{\partial g}{\partial\theta}\right) \\ &= -\int d\theta q(\theta)\left\{(1-f_\theta(x_o))^\beta + f_\theta(x_o)^\beta\right\}(1-f_\theta(x_o))f_\theta(x_o)\frac{\partial g}{\partial\theta}\end{aligned} \tag{73}$$

Since in the logistic regression situation, f_θ is bounded under from 0 to 1, the term $(1-f_\theta(x_o))^\beta + f_\theta(x_o)^\beta$ cannot goes to zero. Therefore, what we have to consider is the term $(1-f_\theta(x_o))f_\theta(x_o)\frac{\partial g}{\partial\theta}$.

$$\begin{aligned}(1-f_\theta(x_o))f_\theta(x_o)\frac{\partial g}{\partial\theta} &= \frac{1}{1+e^{g_\theta}}\frac{1}{1+e^{-g_\theta}}\frac{\partial g}{\partial\theta} \\ &\xrightarrow[x_o\to\infty]{} 0\end{aligned} \tag{74}$$

Therefore, in the limit discussion, we do not have to consider the behavior of second term of Eq.(43). The behavior of IF is determined by the first term of Eq.(43). Accordingly, IF of logistic regression when using β VI is bounded.

Consider the case where there exists activation functions such as $relu$ or $tanh$. Since we do not usually activation function for the final layer, the IF of logistic regression using $relu$ activation function is not bounded when using usual VI because there remains a possibility that $g_\theta(x) \to -\infty$ as $x \to \pm\infty$. In such a case, our analyzing term can diverge. When using $tanh$ activation function, as we discussed in regression setup, IF are always bounded. In the above discussion about $relu$, what is important for the limit is the sign of $g_\theta(x)$. Therefore, even if we add layers, there remains a possibility that IF will diverge. Accordingly, our conclusion is that for logistic regression, $relu$ activation function is not robust against input related outliers even using neural net, while $tanh$ activation function is robust. As for β VI, it is apparent from Eq.(71) and Eq.(74) that IF is bounded for both relu and tanh even using neural net.

Next, we consider the case of γ VI, and what we have to analyze is the second term of Eq.(47). To proceed the analysis, we can use the relation Eq.(54). Since the inverse of hessian matrix is not related to outlier, what we have to analyze is,

$$\begin{aligned}
&\int d\theta q\left(\theta\right) \frac{\partial}{\partial\theta} \frac{p(y'|x')^\gamma}{\{\int p(y|x',\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}} \\
&= \int d\theta q\left(\theta\right) \frac{\{\int p(y|x',\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}} \frac{\partial}{\partial\theta} p(y'|x')^\gamma - p(y'|x')^\gamma \frac{\partial}{\partial\theta}\{\int p(y|x',\theta)^{1+\gamma}dy\}^{\frac{\gamma}{1+\gamma}}}{\{\int p(y|x',\theta)^{1+\gamma}dy\}^{\frac{2\gamma}{1+\gamma}}}.
\end{aligned} \tag{75}$$

In the above expression, what we have to consider is the numerator. The analysis of first term can be done in the same way as Eq.(71). Therefore it is bounded for both relu and tanh. The second term can be analyzed in the same way as Eq.(73), we do not have to consider it in the limit. From above discussion, the behavior of IF for γ VI is the same as that for β VI in the limit, accordingly, it is bounded for neural net even if using relu activation function.

L Analysis based on influence function under no model assumption

Let us compare the behavior of IF of usual VI and our proposing methods intuitively. First we consider usual VI. In Eq.(37), since the term which depends on contamination is $\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\ln p\left(z|\theta\right)\right]$, we only have to treat it for analysis of IF. It is difficult to deal with this expression directly, we focus on typical value of $q\left(\theta\right)$, the mean value m. In such a simplified situation, what we have to consider is following expression.

$$\frac{\partial}{\partial m}\ln p\left(z;m\right) \tag{76}$$

This is the usual maximum likelihood estimator.

Let us consider the unsupervised β VI. What we consider is,

$$\frac{\partial}{\partial m}\left(p\left(z;m\right)\right)^\beta = \left(p\left(z;m\right)\right)^\beta \frac{\partial}{\partial m}\ln p\left(z;m\right) \tag{77}$$

To proceed the analysis, it is necessary to specify a model $p(z;\theta)$, otherwise we cannot evaluate differentiation. Here for intuitive analysis, we simply consider the behavior of $\ln p(z;m)$ and $p(z;m)^\beta \ln p(z;m)$, and in the case of z is outlier, that is $p(z;m)$ is quite small.

Fig.1 shows that $\ln p(z;m)$ is unbounded, on the other hand $p(z;m)^\beta \ln p(z;m)$ is bounded. This means that β divergence VI is robust to outliers.

M Experimental detail and results

In numerical experiments, we used two hidden layer neural network with 20 units for regression and logistic regression for classification. As shown in table 3, the objective function of our method is a

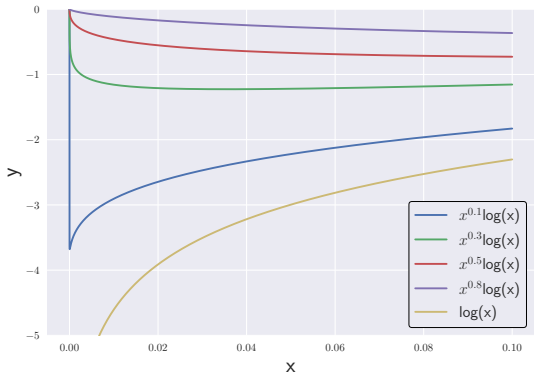


Figure 2: Behavior of $y = \log x$ and $y = x^\beta \log x$. As x become small, $y = \log x$ diverges to $-\infty$, on the other hand $y = x^\beta \log x$ is bounded.

summation over data points and therefore we can employ a stochastic optimization method. We used Adam for the optimizer. Moreover, we optimize objective function by using the re-parameterization trick. More specifically, to separate the randomness to generate θ from variational parameter m, we generate randomness by $p(\omega)$ independently of m and use a deterministic map $\theta = f_q(\omega; m)$. In our implementation, we estimate the gradient of the objective function Eq.(6) by Monte Carlo sampling. In this work, we used Gaussian re-parameterization, for example, if $q(\theta; m) = N(\theta; \mu, \Sigma)$, then $\theta = \mu + \Sigma^{\frac{1}{2}}\omega$, where $\omega \sim N(0, I)$. For the gradient estimation, we used 5 Monte Carlo samples in Sec M.1, and 10 Monte Carlo samples in Sec M.2.

M.1 Influence to predictive distribution

We numerically studied the influence to the predictive distribution by outliers to proceed the analysis of our method.

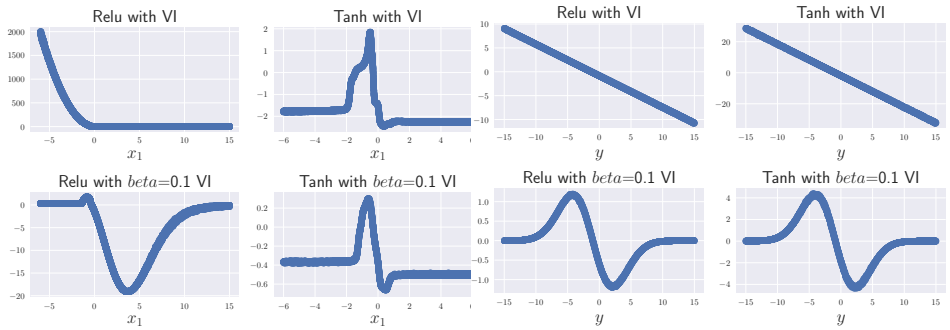
Regression

We used "power plant" dataset in UCI which has four features for input. As a input related outlier, we moves one chosen input feature x_1 and moves it from small value to large value. As a output related outlier, we simply choose the output y. Since it is difficult to plot the behavior of perturbation of predictive distribution, we plot how log-likelihood of a test point is perturbed by a outlier. We compared ordinary VI and VI (β=0.1).

The results are shown in Fig 3, where the vertical axis indicates the value of $\frac{\partial}{\partial \epsilon}\mathbb{E}_{q^*(\theta)}\left[\ln p(x_{\text{test}}|\theta)\right]$, and horizontal axis indicates the value of feature x_1 of a outlier.

The results in Fig 3 shows the model using ReLU activation under ordinary VI can be affected infinitely by input related outliers, while the perturbation is bounded under our method. We can also confirm that the perturbation under our method is smaller that that of ordinary VI even in the case of tanh. As for output related outliers, models under ordinary VI are infinitely perturbed, while perturbation of our method is bounded. From those results, we can see our method is robust for both input and output related outliers in the sense that test point prediction is not influenced infinitely by contaminating one training point.

The difference compared to the influence function analysis in Sec. 4.2 is that the perturbation by input related outliers under tanh activation function model does not converges to zero even when using proposed method in the limit. This might be due to the fact that as the absolute value of input data goes to large, the input to next layer goes to ± 1 when using tanh activation function. For the next layer, the input which has value ± 1 might not be so strange compared to regular data, and not regarded as outliers. Therefore, during the optimization, likelihood of input related outliers is not downweighted so much by robust divergence property and the influence of outliers remains finite. .



(a) Influence by input related outlier (b) Influence by output related outlier

Figure 3: Perturbation on test log-likelihood for neural net regression.

Table 5: Average of test log-likelihood change

	usual VI	VI ($\beta = 0.1$)
ReLU	-1.65e-3	-3.29e-5
tanh	-2.3e-3	-3.49e-4

Classification

We considered binary classification and used logistic regression. We used "eeg" dataset in UCI which has 14 features for input. As a input related outlier, we choose one feature and move it. The result of how test log-likelihood is perturbed is given in Fig. 4. For ordinary VI, using ReLU activation function causes unbounded perturbation, while our method makes the perturbation bounded. We can also confirm that the perturbation under our method is smaller than that of ordinary VI even in the case of tanh.

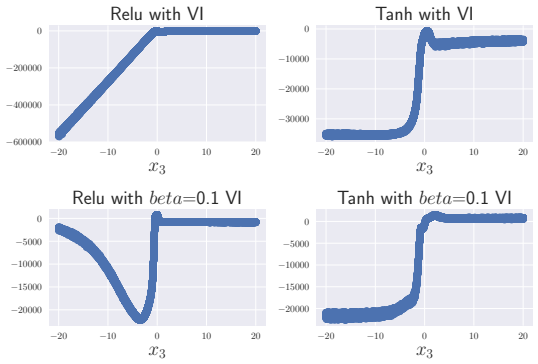


Figure 4: Perturbation on test log-likelihood by input related outlier for logistic regression

As output related outlier, we studied influence of label misspecification. We flip one label of training point and observe how the test log-likelihood change. By assuming that $\epsilon = \frac{1}{N}$, where N is the number of training dataset, we calculated $\frac{1}{N}\frac{1}{N}\sum_i \frac{1}{N_{\text{test}}}\sum_j \frac{\partial}{\partial \epsilon_i}\mathbb{E}_{q^*(\theta)}\left[\ln p(y^j_{\text{test}}|x^j_{\text{test}},\theta)\right]$, which represents the averaged amount of change in test log-likelihood, and the term inside the sum of j means the change of log-likelihood at test data j caused by flipping a label of training data i. Without IF, this amount is difficult to calculate because we have to retraining neural network with flipped data and this is computational heavy. The results are shown in Table 5 that the change of test log-likelihood under our method is smaller than that of ordinary VI. This implies that our method is robust against label misspecification. From these case studies, we can see that our method is robust for both input and output related outliers for both regression and classification setting in the sense that the prediction is less perturbed by adding one outliers.

Table 6: Logistic Regression results, Accuracy(%)

Dataset	Outliers	KL	KL(ϵ)	WL	Réyni	BB-α	β	γ
spam	0%	93.2(0.8)	93.4(1.0)	94.1(0.8)	93.5(0.8)	93.5(0.8)	**94.2(2.0)**	93.3(1.0)
N=4601	10%	92.9(0.74)	93.0(0.9)	93.4(1.1)	93.4(0.72)	93.5(0.8)	**94.2(0.2)**	93.1(0.9)
D=57	20%	92.9(1.1)	92.9(0.7)	93.1(1.1)	93.4(0.97)	93.3(0.9)	**94.2(2.6)**	93.1(0.9)
covertype	0%	65.7(10.2)	60.7(13.6)	**68.5(12.6)**	65.0(9.9)	65.6(9.2)	68.1(9.1)	66.5(9.5)
N=581012	10%	61.7(9.9)	60.0(13.9)	54.5(14.3)	63.9(11.3)	65.4(9.2)	65.6(8.6)	**65.8**(10.1)
D=54	20%	63.7(11.4)	57.2(11.4)	51.2(10.3)	63.2(10.2)	64.4(12.1)	65.8(6.9)	**66.1**(8.1)
eeg	0%	75.6(4.5)	73.4(6.2)	71.2(8.1)	76.6(4.4)	76.6(3.9)	**79.6(2.5)**	79.2(3.1)
N=14890	10%	70.9(7.1)	67.7(7.8)	57.9(1.9)	75.3(4.3)	74.7(4.7)	74.6(3.8)	**77.9(4.0)**
D=14	20%	68.4(7.8)	65.5(6.8)	56.1(2.4)	73.2(4.5)	72.3(4.7)	73.3(6.6)	**77.3(3.6)**

M.2 Bench mark dataset

In this experiment, we determined β and γ by cross validation. We increase the value of β and γ from 0.1 to 0.9 by 0.1.

For Student-t distribution, we chosen the degree of freedom from 3 to 10 by cross-validation. For WL(weighted likelihood proposed in Wang et al. [2017]), we considered Beta distribution for the prior of the weights and we used the method of ADVI for the optimization. For Réyni VI, we chosen α from the set of $\{-1.5, -1.0, -0.5, 0.5, 1.0, 1.5\}$ by cross-validation. For BB-α, we chosen α from the set of $\{0, 0.25, 0.5, 0.75, 1.0\}$ by cross-validation.

For outliers, we both consider input and output related outliers for both regression and classification setting. In the case of regression, if the input are D dimension, we randomly selected $D/2$ dimensions as the dimension which is contaminated and for input related outlier, we first calculate the mean μ and standard deviation σ of the inputs and add the noise to the selected inputs which follows $\epsilon \sim N(\mu, 2\sigma)$. For output related outlier, in the same way to the input related outlier, we first calculate the mean μ and standard deviation σ of the outputs and add the noise to the output which follows $\epsilon \sim N(\mu, 2\sigma)$. The results are shown in the main paper.

In the case of classification, we add noise to input $D/4$ features which are randomly selected. The noise are generated in the same way as the regression setup. As output related noise, we flip the label of data which is chosen randomly. The results are shown in the table 6. KL means the logistic regression using ordinary VI. KL(ϵ) is the case where we used robust likelihood, $p(y = 1|g(x, \theta)) = \epsilon + (1 - 2\epsilon)\sigma(g(x, \theta))$, where σ is sigmoid function and $g(x, \theta)$ is the input to final layer and ϵ is the probability that the target value has been flipped to the wrong value.