Appendix: 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
{futami@ms., sato@, sugi@}k.u-tokyo.ac.jp

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,

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

This is empirically approximated as

$$L_n(\theta) = d_\gamma(\hat{p}(x), p(x;\theta)) = -\frac{1}{\gamma}\ln\frac{1}{n}\sum_{i=1}^n p(x_i;\theta)^\gamma dx + \frac{1}{1+\gamma}\ln\int p(x;\theta)^{1+\gamma}dx. \tag{2}$$

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

$$0 = -\frac{\sum_{i=1}^n p(x_i;\theta)^\gamma \frac{\partial}{\partial\theta}\ln p(x_i;\theta)}{\sum_{i=1}^n p(x_i;\theta)^\gamma} + \int \frac{p(x;\theta)^{1+\gamma}}{\int p(x;\theta)^{1+\gamma}dx}\frac{\partial}{\partial\theta}\ln p(x;\theta)dx. \tag{3}$$

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.

Actually above estimator is equivalent to minimizing following expression,

$$L'_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\gamma+1}{\gamma}\frac{p(x_i|\theta)^\gamma}{\left\{\int p(x|\theta)^{1+\gamma}dy\right\}^{\frac{\gamma}{1+\gamma}}} \tag{4}$$

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

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 $p(y|x;\theta)$.

Following Fujisawa and Eguchi [2008], we define the divergence between true distribution and the model by

$$D_\gamma(p^*(y|x), p(y|x;\theta)|p^*(x))$$
$$= \frac{1}{\gamma}\ln\int\left\{\int g^*(y|x)^{1+\gamma}dy\right\}^{\frac{1}{1+\gamma}}p^*(x)dx - \frac{1}{\gamma}\ln\int\left\{\int g^*(y|x)p(y|x;\theta)^\gamma dy/\left(\int g^*(y|x)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}\right\}p^*(x)dx. \tag{5}$$

As discussed in Fujisawa and Eguchi [2008], in the limit where $\gamma \to 0$, this divergence becomes ordinary KL divergence,

$$\lim_{\gamma \to 0} D_\gamma(p^*(y|x), p(y|x;\theta)|p^*(x)) = \int D_{\mathrm{KL}}(p^*(y|x), p(y|x;\theta))p^*(x)dx \tag{6}$$

What we minimize is following γ cross entropy over the distribution $p^*(x)$, Actually, minimizing γ divergence is equivalent to minimizing the second term of Eq.(5). By empirical approximation, what we minimize is following expression,

$$L_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} L_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{p(y_i|x_i;\theta)^\gamma}{\left\{\int p(y|x_i;\theta)^{1+\gamma}dy\right\}^{\frac{\gamma}{1+\gamma}}}. \tag{7}$$

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

$$L_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln p(y_i|x_i;\theta). \tag{8}$$

This is ordinary KL cross entropy.

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 for supervised settings is

$$L_n(\theta) = d_\beta(\hat{p}(y|x), p(y|x;\theta)|\hat{p}(x)) = -\frac{\beta+1}{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n} p(y_i|x_i;\theta)^\beta\right\} + \left\{\frac{1}{n}\sum_{i=1}^{n}\int p(y|x_i;\theta)^{1+\beta}dy\right\}. \tag{9}$$

For the unsupervised setting, the empirical approximation of β cross entropy is

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

C Proof of Eq.(14) in the main paper

From the definition of KL divergence Eq.(2) in the main paper, the cross entropy can be expressed as

$$d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right) = D_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right) + \mathrm{Const.} \tag{11}$$

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

$$\arg\min_{q(\theta)\in\mathcal{P}} L(q(\theta)), \tag{12}$$

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

$$\arg\min_{q(\theta)\in\mathcal{P}} \frac{1}{N}L(q(\theta)). \tag{13}$$

Therefore Eq.(14) is equivalent to Eq.(13)

D Proof of Theorem 1

The objective function is given as

$$L_\beta = \mathbb{E}_{q(\theta)}[D_\beta\left(\hat{p}(x)||p(x|\theta)\right)] + \lambda' D_{KL}\left(q(\theta)||p(\theta)\right) \tag{14}$$

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,

$$\frac{d(L_\beta + \lambda(\int q(\theta)d\theta - 1))}{dq(\theta)} = D_\beta\left(\hat{p}(x)|p(x|\theta))\right] + \lambda' \ln \frac{q(\theta)}{p(\theta)} - (1+\lambda) = 0 \tag{15}$$

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

$$q(\theta) \propto p(\theta) e^{-\frac{1}{\lambda'} d_\beta(\hat{p}(x)|p(x|\theta))} \tag{16}$$

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

$$q(\theta) = \frac{e^{-N d_\beta(\hat{p}(x)|p(x|\theta))} p(\theta)}{\int e^{-N d_\beta(\hat{p}(x)|p(x|\theta))} p(\theta) d\theta}. \tag{17}$$

We can get the similar expression for γ cross entropy.

Interestingly, if we use KL cross entropy instead of β cross entropy in the above discussion, following relation holds,

$$\begin{aligned} q(\theta) \propto p(\theta) e^{-\frac{1}{\lambda'} d_{KL}(\hat{p}(x)|p(x|\theta))} &= p(\theta) e^{-N(-\frac{1}{N}\sum_i \ln p(x_i|\theta))} \\ &= p(\theta) \prod_i p(x_i|\theta) \\ &= p(\theta) p(D|\theta) \end{aligned} \tag{18}$$

The normalizing constant is

$$\int p(\theta) \prod_i p(x_i|\theta) d\theta = p(D). \tag{19}$$

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

$$q(\theta) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{20}$$

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. In this paper we only consider the situation that regularization constant is $\frac{1}{\lambda'} = N$ based on the similarity of Bayes' theorem. 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.

E Pseudo posterior

The expression Eq.(17) is called pseudo posterior in statistics. In general, pseudo posterior is given as

$$q(\theta) = \frac{e^{-\lambda R(\theta)} p(\theta)}{\int e^{-\lambda R(\theta)} p(\theta) d\theta}. \tag{21}$$

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

$$\begin{aligned} q(\theta) &\propto e^{-N\left\{\frac{\beta+1}{\beta}\frac{1}{N}\sum_{i=1}^N p(x_i;\theta)^\beta + \int p(x;\theta)^{1+\beta} dx\right\}} p(\theta) \\ &= \left[\prod_i^N e^{l_\theta(x_i)} p(\theta)\right] \end{aligned} \tag{22}$$

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.(18). In ordinary 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.(18). 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.

Table 1: Cross-entropies for robust variational inference.

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

F Proof of Theorem 2

We consider the situation where the distribution is expressed as

$$G_\varepsilon(x) = (1-\varepsilon)\, G_n(x) + \varepsilon \Delta_z(x) \tag{23}$$

Before going to the detail, we summarize the objective function of VI and proposed method.

First, the objective function of ordinary VI is given by

$$L = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) + N\mathbb{E}_{q(\theta)}\left[N d_{\mathrm{KL}}\left(\hat{p}(x)\|p(x|\theta)\right)\right]. \tag{24}$$

In the same way, objective functions of β-VI and γ-VI are given by

$$L_\beta = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) + N\mathbb{E}_{q(\theta)}\left[N d_\beta\left(\hat{p}(x)\|p(x|\theta)\right)\right], \tag{25}$$

$$L_\gamma = D_{\mathrm{KL}}(q(\theta)\|p(\theta)) + N\mathbb{E}_{q(\theta)}\left[N d_\gamma\left(\hat{p}(x)\|p(x|\theta)\right)\right], \tag{26}$$

where d_β and d_γ are summarized in Table 1. By using these expressions, we will derive the influence functions.

F.1 Derivation of IF for ordinary VI

We start from the first order condition,

$$\begin{aligned} 0 &= \frac{\partial}{\partial m} L\Big|_{m=m^*} \\ &= \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x)\ln p(x|\theta) + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right] \end{aligned} \tag{27}$$

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

$$\begin{aligned} 0 = {} & \nabla_m \int d\theta \frac{\partial m^*(\epsilon)}{\partial \epsilon}\frac{\partial q}{\partial m^*(\epsilon)}\left\{(1-\epsilon)N\int dG_n(x)\ln p(x|\theta) + \epsilon N \ln p(z|\theta) + \ln p(\theta)\right\} \\ & + \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[-N\int dG_n(x)\ln p(x|\theta) + N\ln p(z|\theta)\right] \\ & - \nabla_m \int d\theta \frac{\partial m^*(\epsilon)}{\partial \epsilon}\frac{\partial q}{\partial m^*(\epsilon)}\ln q(\theta;m^*(\epsilon)) - \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[\frac{\partial m^*(\epsilon)}{\partial \epsilon}.\frac{\partial \ln q}{\partial m^*(\epsilon)}\right] \end{aligned} \tag{28}$$

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

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\int dG_n(x)\ln p(x|\theta) - N\ln p(z|\theta)\right]. \tag{29}$$

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

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = \left(\frac{\partial^2 L}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[D_{KL}(q(\theta;m)|p(\theta)) + N\ln p(z|\theta)\right]. \tag{30}$$

F.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.(10) and Eq.(9).

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

$$\begin{aligned}0 =& \left.\frac{\partial}{\partial m} L_\beta\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x)\frac{\beta+1}{\beta}p(x|\theta)^\beta - N\int p(x|\theta)^{1+\beta}dx + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right].\end{aligned} \tag{31}$$

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

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\frac{\beta+1}{\beta}\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\int dG_n(x)p(x|\theta)^\beta - Np\left(z|\theta\right)^\beta\right]. \tag{32}$$

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

$$G_\varepsilon(x,y) = (1-\varepsilon)G_n(x,y) + \varepsilon\Delta_{z=(x',y')}(x,y) \tag{33}$$

The first order condition is,

$$\begin{aligned}0 =& \left.\frac{\partial}{\partial m} L_\beta\right|_{m=m^*} \\ =& \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))}\left[N\int dG_\epsilon(x,y)\frac{\beta+1}{\beta}p(y|x,\theta)^\beta - N\int dG_\epsilon(x)\left\{\int p(y|x,\theta)^{1+\beta}dy\right\} + \ln p(\theta) - \ln q(\theta;m^*(\epsilon))\right].\end{aligned} \tag{34}$$

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

$$\begin{aligned}\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} =& -N\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\frac{\beta+1}{\beta}\left(\int dG_n(y,x)p(y|x,\theta)^\beta - p\left(y'|x',\theta\right)^\beta\right)\right] \\ &+ N\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[\int dG_n(x)\left(\int p(y|x,\theta)^{1+\beta}dy\right) - \int p(y|x',\theta)^{1+\beta}dy\right].\end{aligned} \tag{35}$$

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

F.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.(8). For unsupervised situation, the first order condition is given by,

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

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

$$\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L_\gamma}{\partial m^2}\right)^{-1}\frac{\partial}{\partial m}\mathbb{E}_{q(\theta)}\left[N\frac{\int dG_n(x)p(x|\theta)^\gamma - p(z|\theta)^\gamma}{\left\{\int p(x|\theta)^{1+\gamma}dx\right\}^{\frac{\gamma}{1+\gamma}}}\right]. \tag{37}$$

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

$$
\begin{aligned}
0 &= \frac{\partial}{\partial m} L_\gamma \Big|_{m=m^*} \\
&= \nabla_m \mathbb{E}_{q(\theta;m^*(\epsilon))} \left[N \int dG_\epsilon(x,y) \frac{p(y|x,\theta)^\gamma}{\left\{ \int p(y|x,\theta)^{1+\gamma} dy \right\}^{\frac{\gamma}{1+\gamma}}} + \ln p(\theta) - \ln q(\theta; m^*(\epsilon)) \right]. \quad (38)
\end{aligned}
$$

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

$$
\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -N \left(\frac{\partial^2 L_\gamma}{\partial m^2} \right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q(\theta)} \left[\int dG_n(x,y) \frac{p(y|x,\theta)^\gamma}{\left\{ \int p(y|x,\theta)^{1+\gamma} dy \right\}^{\frac{\gamma}{1+\gamma}}} - \frac{p(y'|x',\theta)^\gamma}{\left\{ \int p(y|x',\theta)^{1+\gamma} dy \right\}^{\frac{\gamma}{1+\gamma}}} \right]. \quad (39)
$$

G Other aspects of analysis based on influence function

In the above sections, we considered that outliers are added to the original training dataset. We can consider a other type of contamination, such as training data itself is perturbed, that is, a training point $z = (x, y)$ is perturbed to $z_\epsilon = (x + \epsilon, y)$(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 ordinary VI is given by

$$
\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L}{\partial m^2} \right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q(\theta)} \left[\frac{\partial}{\partial x} \ln p(z|\theta) \right]. \quad (40)
$$

IF of β divergence based VI is given by

$$
\frac{\partial m^*(\varepsilon)}{\partial \varepsilon} = -\left(\frac{\partial^2 L_\beta}{\partial m^2} \right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q(\theta)} \left[\frac{\partial}{\partial x} p(z|\theta)^\beta \right]. \quad (41)
$$

H Another type of γ VI

In the main paper, we used the transformed γ cross entropy, which is given in Eq.(7). 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.(42), we can derive the pseudo posterior using transformed cross entropy,

$$
\begin{aligned}
q(\theta) &\propto e^{N \frac{\gamma+1}{\gamma} \frac{1}{N} \sum_{i=1}^N \frac{p(x_i|\theta)^\gamma}{\{\int p(x|\theta)^{1+\gamma} dy\}^{\frac{\gamma}{1+\gamma}}}} p(\theta) \\
&= \left[\prod_i^N e^{l_\theta(x_i)} p(\theta) \right] \quad (42)
\end{aligned}
$$

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,

$$
q(\theta) \propto e^{N(\frac{1}{\gamma} \ln \frac{1}{N} \sum_i^N p(x_i|\theta)^\gamma dx - \frac{1}{1+\gamma} \ln \int p(x|\theta)^{1+\gamma} dx)} p(\theta) \quad (43)
$$

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

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

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

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

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

I Discussion of Influence function

In this section, we describe the detail discussion of influence function's behavior when using a neural net model for the regression and the classification with logistic loss.

We focus on the influence function of the variational parameter in the approximate posterior distribution. We use mean-field variational inference and Gaussian distribution for approximate posterior. $q(\theta)$ denote the approximate posterior. Since Gaussian distribution is a member of an exponential family, we can parametrize it by its mean value m. In the case of Gaussian distribution, $m = \{\mathbb{E}[\theta], \mathbb{E}[\theta^2]\}$. We can parametrize variational posterior as $q(\theta|m)$. Thus we only analyze the influence function of $m = \mathbb{E}[\theta]$ in this section and m indicates the $m = \mathbb{E}[\theta]$ not $\mathbb{E}[\theta^2]$.

Let us start ordinary variational inference. In Eq.(40), we especially focus on the term, $\frac{\partial}{\partial m} \mathbb{E}_{q(\theta|m)} [\ln p(y|\theta)]$, 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,

$$
\begin{aligned}
\frac{\partial}{\partial m} \mathbb{E}_{q(\theta|m)} [\ln p(y|\theta)] &= \frac{\partial}{\partial m} \left\{ \int q(\theta|m) \ln p(y|\theta) \, d\theta \right\} \\
&= \int \frac{\partial q(\theta|m)}{\partial m} \ln p(y|\theta) \, d\theta \\
&= -\int q(\theta|m) \frac{\partial}{\partial \theta} \ln p(y|\theta) \, d\theta \\
&= -E_{q(\theta|m)} \left[\frac{\partial}{\partial \theta} \ln p(y|\theta) \right]
\end{aligned} \tag{46}
$$

, where We used partial integration for the second line to third line. and also used the following relation which holds for Gaussian distribution

$$
\frac{\partial q(\theta|m)}{\partial m} = \frac{\partial q(\theta|m)}{\partial \theta}. \tag{47}
$$

This relation also holds for the Student-T

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

I.1 Regression

In this subsection, we consider the regression problem by a 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 θs are random variables which obeys approximate posterior $q(\theta|m)$.

We consider the output layer as Gaussian distribution as $p(y|f_\theta(x)) = N(y|f_\theta(x), I)$. From above discussion, what we have to consider is $\dfrac{\partial}{\partial\theta}\ln p\left(y|f_\theta(x)\right)$.

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 ordinary VI,

$$\frac{\partial}{\partial\theta}\ln p\left(y_o|f_\theta(x_o)\right) \propto \left(y_o - f_\theta(x_o)\right)\frac{\partial f_\theta(x_o)}{\partial\theta}. \tag{48}$$

We can see that this term does not bounded when $y_o \to \pm\infty$. And thus IF of ordinary VI is unbounded as output related outlier become large.

As for the β divergence, we have to treat Eq.(35). Fortunately, when we use Gaussian distribution for output layer, the second term in the bracket of Eq.(35) will be constant by the analytical integration, and thus its derivative will be zero. Therefore the output related term is only the first term. Thanks to this property, the denominator of Eq.(39) 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 the regression. We get the following expression,

$$\begin{aligned}\frac{\partial}{\partial\theta}p\left(y_o|f_\theta(x_o)\right)^\beta &\propto e^{-\frac{\beta}{2}\left(y_o - f_\theta(x_o)\right)^2}\left(y_o - f_\theta(x_o)\right)\frac{\partial f_\theta(x_o)}{\partial\theta}\\ &= \frac{\left(y_o - f_\theta(x_o)\right)}{e^{\frac{\beta}{2}\left(y_o - f_\theta(x_o)\right)^2}}\frac{\partial f_\theta(x_o)}{\partial\theta}\end{aligned} \tag{49}$$

From this expression, we can see that IF of β VI is bounded because Eq.(49) 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. We consider whether Eq.(48) and Eq.(49) are bounded or not 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.(48) is unbouded. As for Eq.(49), the exponential function in the denominator of Eq.(49) plays a crucial role. Thanks to this exponential function,

$$\begin{aligned}\frac{\partial}{\partial W_1}p\left(y_o|f_\theta(x_o)\right)^\beta &\propto \frac{\left(y_o - f_\theta(x_o)\right)}{e^{\frac{\beta}{2}\left(y_o - f_\theta(x_o)\right)^2}}x_o\\ &\xrightarrow[x_o\to\infty]{} 0\end{aligned} \tag{50}$$

From these facts, ordinary 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,

$$\frac{\partial}{\partial W_1}f_\theta(x_o) = W_2 x_o, \quad \frac{\partial}{\partial W_2}f_\theta(x_o) = W_1 x_o + b_1 \tag{51}$$

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 ordinary 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, $f_\theta(x_o) = W_2(relu\left(W_1 x_o + b_1\right)) + b_2$,

$$\frac{\partial f_\theta(x_o)}{\partial W_2} = relu\left(W_1 x_o + b_1\right), \quad \frac{\partial f_\theta(x_o)}{\partial W_1} = \begin{cases} W_2 x_o, & W_1 x_o + b_1 \geq 0 \\ 0, & W_1 x_o + b_1 < 0, \end{cases} \tag{52}$$

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

When we use $tanh$ as a activation function, $f_\theta(x_o) = W_2 tanh\left(W_1 x_o + b_1\right) + b_2$,

$$\frac{\partial f_\theta(x_o)}{\partial W_1} = \frac{W_2 x_o}{cosh^2\left(W_1 x_o + b_1\right)} \xrightarrow[x_o \to \infty]{} 0 \tag{53}$$

The limit of above expression can be easily understand from Fig.1. From this expression, we can

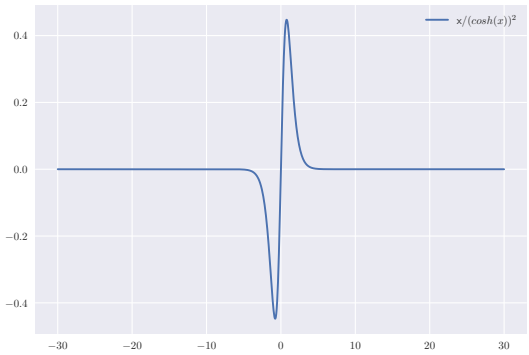


Figure 1: Behavior of $\frac{x}{\cosh^2 x}$

understand IF of W_1 is bounded in both ordinary 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,

$$\frac{\partial f_\theta(x_o)}{\partial W_2} = tanh(W_1 x_o + b_1) \tag{54}$$

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. And thus IF of both ordinary VI and β VI is bounded when we use $tanh$ activation function.

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 ordinary VI will diverge.

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

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,

$$\frac{\partial}{\partial \theta} \ln p\left(y_o | f_\theta(x_o)\right) \propto \frac{\left(y_o - f_\theta(x_o)\right)}{\nu\sigma^2 + \left(y_o - f_\theta(x_o)\right)^2} \frac{\partial f_\theta(x_o)}{\partial \theta} \tag{55}$$

By comparing Eq.(55) with Eq.(48) and Eq.(49), we can confirm that the behavior of IF in the case of Student-t loss in ordinary VI is similar to Gaussian loss model in β VI. First, consider output related outlier,

$$\frac{\partial}{\partial\theta}\ln p\left(y_o|f_\theta(x_o)\right) \xrightarrow[y_o\to\infty]{} 0 \tag{56}$$

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 $\theta = \{W_1, b_1\}$

$$\begin{aligned}
\frac{\partial}{\partial W_1}\ln p\left(y_o|f_\theta(x_o)\right) &\propto \frac{(y_o - f_\theta(x_o))}{\nu\sigma^2 + (y_o - f_\theta(x_o))^2}x_o \\
&= \frac{(y_o - f_\theta(x_o))^2}{\nu\sigma^2 + (y_o - f_\theta(x_o))^2}\frac{x_o}{y_o - f_\theta(x_o)} \\
&= \frac{(y_o - f_\theta(x_o))^2}{\nu\sigma^2 + (y_o - f_\theta(x_o))^2}\frac{f_\theta(x_o) - b_1}{W_1(y_o - f_\theta(x_o))} \\
&\xrightarrow[x_o\to\infty]{} -W_1^{-1}
\end{aligned} \tag{57}$$

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 for Student-t loss, the IF is bounded but finite value remains.

Although the finite value remains in IF when using Student-T loss and its value is W_1, the value is considerably small. Therefore we can disregard the remained influence of Student-T loss in practice.

I.2 Classification

In this section, we consider the classification problem. We focus on the 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 the logistic regression model,

$$p(y|f_\theta(x)) = f_\theta(x)^y(1 - f_\theta(x))^{(1-y)} \tag{58}$$

where

$$f_\theta(x) = \frac{1}{1 + e^{-g_\theta(x)}} \tag{59}$$

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

We first assume $g_\theta(x) = Wx + b$, then $\dfrac{\partial g}{\partial W} = x$ and $\dfrac{\partial g}{\partial b} = 1$. We assume prior and posterior distribution of W and b are Gaussian distributions. For IF analysis, we first consider the first term of Eq.(35) and only consider outlier related term inside it. To proceed the calculation, we can use the relation Eq.(46), and what we have to analyze is

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln p(y|f_\theta(x)) &= \frac{\partial}{\partial\theta}\left(y\ln f_\theta(x) + (1-y)\ln(1 - f_\theta(x))\right) \\
&= -y(1-f)\frac{\partial g}{\partial\theta} + (1-y)f\frac{\partial g}{\partial\theta}
\end{aligned} \tag{60}$$

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

$$\frac{\partial}{\partial\theta}\ln p(y = +1|f_\theta(x)) = \frac{1}{1 + e^{g_\theta(x)}}\frac{\partial g}{\partial\theta} \tag{61}$$

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

$$\frac{\partial}{\partial W}\ln p(y = +1|f_\theta(x)) = \frac{1}{1 + e^{Wx+b}}x \tag{62}$$

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 a logistic regression for ordinary VI.

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

$$p(y=+1|f_\theta(x))^\beta \frac{\partial}{\partial \theta} \ln p(y=+1|f_\theta(x)) = \frac{1}{(1+e^{-g_\theta(x)})^\beta} \frac{1}{1+e^{g_\theta(x)}} \frac{\partial g}{\partial \theta} \tag{63}$$

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.(35) for analysis of IF, which is vanish in the regression situation. The second term of Eq.(35) can be written as

$$\begin{aligned} &\left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q(\theta)} \left[N \int p(y|x_o,\theta)^{1+\beta} dy \right] \\ &= N \left(\frac{\partial^2 L_\beta}{\partial m^2}\right)^{-1} \frac{\partial}{\partial m} \mathbb{E}_{q(\theta)} \left[f_\theta(x_o)^{1+\beta} + (1-f_\theta(x_o))^{1+\beta} \right] \end{aligned} \tag{64}$$

To proceed the analysis, we can use the relation Eq.(46). Since the inverse of hessian matrix is not related to outlier, what we have to consider is

$$\begin{aligned} &\int d\theta q(\theta) \frac{\partial}{\partial \theta} f_\theta(x_o)^{1+\beta} + \frac{\partial}{\partial \theta}(1-f_\theta(x_o))^{1+\beta} \\ &= -\int d\theta q(\theta) \left(f_\theta(x_o)^{1+\beta}(1-f_\theta(x_o)) \frac{\partial g}{\partial \theta} + (1-f_\theta(x_o))^{1+\beta} f_\theta(x_o) \frac{\partial g}{\partial \theta} \right) \\ &= -\int d\theta q(\theta) \left\{ (1-f_\theta(x_o))^\beta + f_\theta(x_o)^\beta \right\} (1-f_\theta(x_o)) f_\theta(x_o) \frac{\partial g}{\partial \theta} \end{aligned} \tag{65}$$

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 $(1-f_\theta(x_o)) f_\theta(x_o) \frac{\partial g}{\partial \theta}$.

$$\begin{aligned} (1-f_\theta(x_o)) f_\theta(x_o) \frac{\partial g}{\partial \theta} &= \frac{1}{1+e^{g_\theta}} \frac{1}{1+e^{-g_\theta}} \frac{\partial g}{\partial \theta} \\ &\xrightarrow[x_o \to \infty]{} 0 \end{aligned} \tag{66}$$

Therefore, in the limit discussion, we do not have to consider the behavior of second term of Eq.(35). The behavior of IF is determined by the first term of Eq.(35). 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 use activation function for the final layer, the IF of logistic regression using $relu$ activation function is not bounded when using ordinary 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.

Accordingly, our conclusion is that for the logistic regression, $relu$ activation function is not robust against input related outliers when using ordinal VI, while $tanh$ activation function is robust. As for β VI, it is apparent from Eq.(63) and Eq.(66) 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.(39). To proceed the analysis, we can use the relation Eq.(46). Since the inverse of hessian matrix is not related to outlier, what we have to analyze is,

$$\begin{aligned} &\int d\theta q(\theta) \frac{\partial}{\partial \theta} \frac{p(y'|x')^\gamma}{\{\int p(y|x',\theta)^{1+\gamma} dy\}^{\frac{\gamma}{1+\gamma}}} \\ &= \int d\theta q(\theta) \frac{\{\int p(y|x',\theta)^{1+\gamma} dy\}^{\frac{\gamma}{1+\gamma}} \frac{\partial}{\partial \theta} p(y'|x')^\gamma - p(y'|x')^\gamma \frac{\partial}{\partial \theta} \{\int p(y|x',\theta)^{1+\gamma} dy\}^{\frac{\gamma}{1+\gamma}}}{\{\int p(y|x',\theta)^{1+\gamma} dy\}^{\frac{2\gamma}{1+\gamma}}}. \end{aligned} \tag{67}$$

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.(63). Therefore it is bounded for both relu and tanh. The second term can be analyzed in the same way as Eq.(65), 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 even if using $relu$ activation function.

J Analysis based on influence function under no model assumption

Let us compare the behavior of IF of ordinal VI and our proposed methods intuitively. First we consider ordinary VI. In Eq.(29), 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 focus on that term. 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.

$$\frac{\partial}{\partial m}\ln p\left(z;m\right) \tag{68}$$

This is the ordinary maximum likelihood estimator.

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

$$\frac{\partial}{\partial m}\left(p\left(z;m\right)\right)^{\beta} = \left(p\left(z;m\right)\right)^{\beta}\frac{\partial}{\partial m}\ln p\left(z;m\right) \tag{69}$$

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.

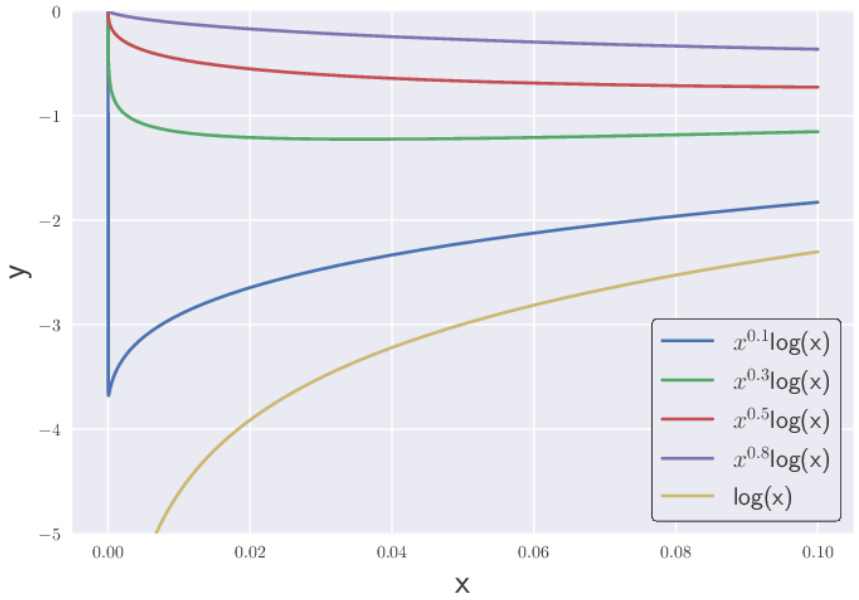


Figure 2: Behavior of $y = \log x$ and $y = x^{\beta}\log x$. As x become small, $y = \log x$ diverges to $-\infty$, on the other hand $y = x^{\beta}\log x$ is bounded.

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.

K Comparison of β VI and γ VI

In this section, we compare the proposed β VI and γ VI theoretically. Although β VI and γ VI have robustness in based on the influence function analysis, their robustness property have significant

difference if the proportion of contamination is large. If the proportion of contamination is large the assumption of discussion of influence function does not hold because we assumed that the ϵ is near zero to derive the influence function.

If the proportion of contamination is not small, other kinds of discussion is needed. Such a discussion is given in in Fujisawa and Eguchi [2008], therefore we review it and use it for our variational objectives.

Following the notation in Fujisawa and Eguchi [2008], $g(x)$ denotes the contaminated probability density function,

$$g(x) = (1-\epsilon)f(x) + \epsilon\delta(x), \tag{70}$$

where $f(x)$ is the underlying true probability density function, $\delta(x)$ denotes the contamination probability density function, and ϵ is the contamination proportion.

We assume that when a data point x^* is an outlier $f(x^*)$ is sufficiently small. We express this assumption by saying that the following quantity is sufficiently small for an appropriate large $\gamma_0 > 0$,

$$\nu_f = \left\{ \int \delta(x) f(x)^{\gamma_0} dx \right\}^{1/\gamma_0}. \tag{71}$$

This means that $\delta(x)$ exists on the tail of $f(x)$. If $\delta(x)$ is the Dirac function at x^*, $\nu_f = f(x^*)$, and above assumption simply means when a data point x^* is an outlier $f(x^*)$ is sufficiently small.

Under this assumption, following lemma and theorem holds (this is lemma3.1 and theorem 3.2 in Fujisawa and Eguchi [2008]) that

Lemma 1 *Suppose that the positive function h satisfies the above assumption, where f is replaced by h. It then holds*

$$\begin{aligned} d_\gamma(g,h) &= d_\gamma((1-\epsilon)f,h) + O(\epsilon\nu_h^\gamma) \\ &= d_\gamma(f,h) - \frac{1}{\gamma}\log(1-\epsilon) + O(\epsilon\nu^\gamma) \end{aligned} \tag{72}$$

Theorem 1 *Suppose that the positive function h satisfies the above assumption, where f is replaced by h. Let $\nu = \max\{\nu_f, \nu_h\}$. Then, the Pythagorean relation among g, f, and h approximately holds:*

$$\Delta(g,f,h) = D_\gamma(g,h) - D_\gamma(g,f) - D_\gamma(f,h) = O(\epsilon\nu^\gamma) \tag{73}$$

This theorem means that the minimizing divergence from the model h to contaminated density g is approximately equivalent to minimizing the divergence h to true distribution f and its order of error is given by $O(\epsilon\nu^\gamma)$.

Recall that the objective function of our proposed is given by

$$L_\gamma(q(\theta)) = \int q(\theta) \left(N d_\gamma \left(g(x) \| p(x|\theta) \right) \right) d\theta + D_{\mathrm{KL}}(q(\theta)\|p(\theta)), \tag{74}$$

where $g(x)$ is the contaminated distribution and $p(x|\theta)$ is the model we prepared. By using the Pythagorean relation, we can rewrite the above expression in the following way by using the true underlying distribution,

$$L_\gamma(q(\theta)) = \int q(\theta) \left(N d_\gamma(f(x)\|p(x|\theta)) - \frac{1}{\gamma}\log(1-\epsilon) + O(\epsilon\nu^\gamma) \right) d\theta + D_{\mathrm{KL}}(q(\theta)\|p(\theta)). \tag{75}$$

This equation means that by using the γ cross entropy, we can utilize the γ cross entropy between true distribution to our model. We optimized the objective function by using the black-box variational inference method and optimize the variational parameters by gradient decent, and thus the constant terms inside the integral are neglected.

This relation is obtained under the assumption of Eq. (71). The assumption is not the assumption that we used in the influence function that contamination proportion of ϵ is small. Therefore even if the contamination proportion is large, we can obtain the Actually, the robustness of β divergence is assured by the influence function (Basu et al. [1998]) and thus it is not guaranteed if the contamination proportion is not sufficiently small. Following this observation, γ divergence based method is superior to β divergence method.

Table 2: RMSE of VI and β=0.1 VI for toy data.

Outliers	KL(Gaussian)	$\beta = 0.1$(Gaussian)
No outliers	0.01	0.01
Outlier exists	0.69	0.01

Table 3: Accuracy of VI and β=0.4 VI for toy data.

Outliers	KL(logistic)	$\beta = 0.4$(logistic)
No outliers	0.97	0.97
Outlier exists	0.95	0.97

L Experimental detail and results

In the numerical experiment, all the prior distributions are standard Gaussian distributions. We used mean-field variational inference and we used Gaussian distributions as approximate posteriors.

L.1 Toy experiment

For the regression task, we generated the toy data by using (x,y) by $y \sim w^\top x + \epsilon$, where $w^\top = (-0.5, -0.1), x \sim N(0, 1)$ and $\epsilon \sim N(0, 0.1)$. We generated 1000 data points. Outliers are generated by $x \sim N(-15, 1)$ which we considered the measurement error. We generate 24 outliers, which is 2.4% of the regular dataset. We used the linear regression, $p(y|x) = N(y|f_\theta(x), 1), f_\theta(x) = Wx+b$.

For the binary classification, the toy data are generated with the probability $p(x|y = +1) = N(x|\mu_1, \sigma_1), p(x|y = -1) = N(x|\mu_2, \sigma_2)$, where $\mu_1^\top = (-1, -1),\ \mu_2^\top = (1, 1), \sigma_1 = I,\ \sigma_2 = 4I$, where I is identity matrix. We generate 1000 data for each class, and in total 2000 regular points. As outliers we generate 30 outliers by using $p(x|y = +1) = N(x|\mu_o, \sigma_o)$, where $\mu_o^\top = (7, 0),\ \sigma_2 = 0.1I$. The outliers are shown by stars in the picture in the main paper. For binary classification, we use logistic regression, where $p(y = +1|x) = logit(f_\theta(x)), f_\theta(x) = Wx + b$. We prepare priors and posteriors in the same way with binary classification.

The performance of ordinary VI estimation and our proposing methods are shown in Table.2. Apparently, the performance of ordinary VI significantly deteriorates when adding outliers. On the other hand, the performance of our proposing method is not affected by outliers.

The illustrative results are shown in the main text. We also show the performance on this toy experiment in Table 2 and Table 3. Those tables show that ordinary VI is heavily affected by outliers, while our method is not affected so much.

L.2 Influence function

Based on the discussion of Appendix I, the dominant term in IF of γ VI behaves similarly with β VI, therefore we also expected that the perturbation of predictive distribution by outliers in γ VI behaves in the same way as β VI. And thus, we numerically studied the perturbation of predictive distribution only about ordinary VI and β VI. In each calculation, we used 200MC samples to get stable curves.

Regression

We investigated three cases where there is only input related outliers and only output related outliers and both outliers exist.

For an easy visualization and computational savings, we only contaminated the chosen single feature of the input. Since inputs have four dimensional features, $x \in \mathbb{R}^4$, we chose the first feature x_1 to contaminate. To investigate the how predictive distribution depends on the contamination of the input, we chose randomly a single data point from the training data and moved the value of the first feature of the chosen data from $-\infty$ to ∞.

For the output related outlier setting, we chose randomly a single data point from the training data and moved the output value of chosen data from $-\infty$ to ∞.

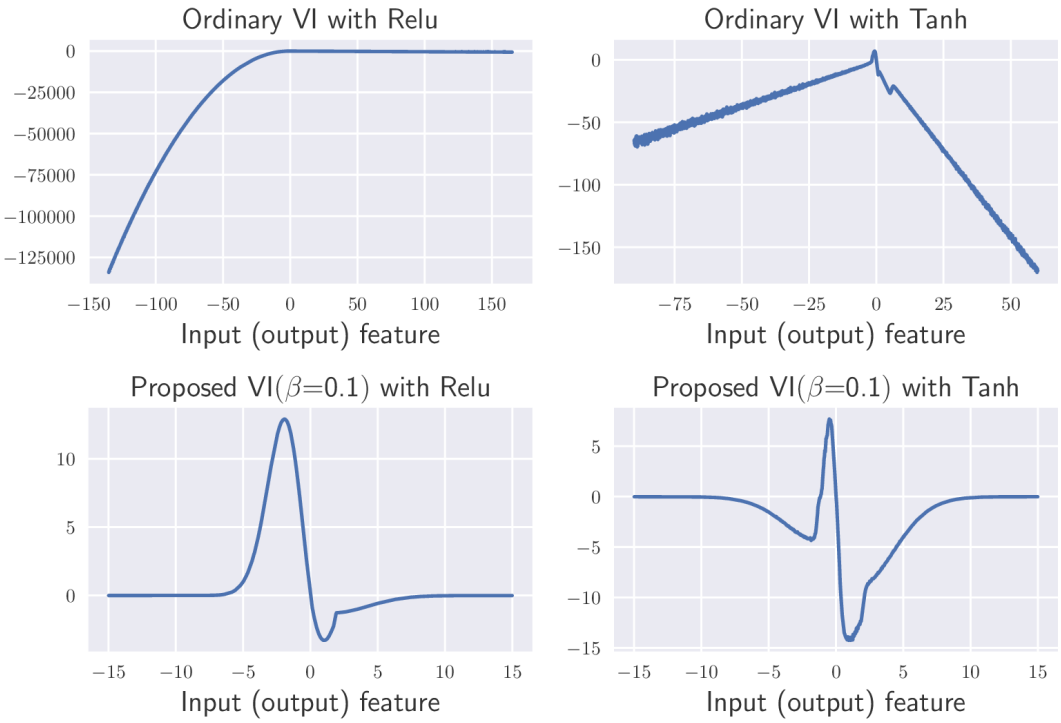


Figure 3: Perturbation on test log-likelihood for neural net regression.

For both input and output related outlier setting, we chose randomly a single data point from the training data and moved the first feature of the input and the output of chosen data from $-\infty$ to ∞.

In the main paper, the figure of input and output related outlier settings are shown. Here we show the both the input and output related outlier situation and the graph is the case when the first feature of the input and the output value increase simultaneously.

From this figure, we confirmed again that in this situation, the perturbation on ordinary VI is not unbounded and the perturbation on our proposed method is bounded.

Classification

In the classification problem, first we considered how predictive distribution depends on the input related outlier. The method is as same as the regression problem. Since inputs have 14 dimensional features, $x \in \mathbb{R}^{14}$, we chose the third feature x_3 to move.

In the label misspecification experiment, we flipped one label of training dataset and measured how log-likelihood of test data are changed. From this experiment, we measured how label misspecification by chosen training data influences the prediction. We repeated this procedure for every training data point and took average. By this experiment, we measured how one flip of training data would influence the prediction on average.

The results shown in the main paper indicates that ordinary VI causes larger minus test log-likelihood change compared to β VI. Base on the fact that decrease of log likelihood is almost equivalent to the increase of loss, the label misspecification causes larger perturbation to prediction in ordinary VI compared to proposed VI.

Calculation of the Hessian

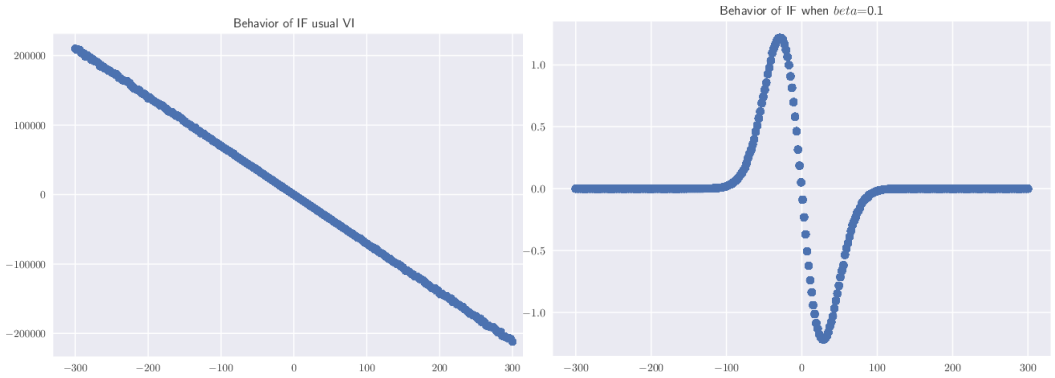
In the above calculation, we have to evaluate the Hessian of ELBO. To save the computational cost we used following method,

$$\frac{\partial^2 L_\beta}{\partial m^2} v = \arg\min_t \frac{1}{2} t^\top \frac{\partial^2 L_\beta}{\partial m^2} t - v^\top t \tag{76}$$

This is the technique that instead of calculating the Hessian directly, we can calculate the product of the Hessian and a vector by solving the second order optimization problem. In our case, we consider $t = \frac{\partial}{\partial m}\mathbb{E}_{q^*(\theta)}\left[\ln p(x_{\text{test}}|\theta)\right]$ and solve above optimization problem.

Influence function of the parameter of the neural network

In this section, we show the IF of parameters. Figure 4 shows the plot of $IF(x_1, W, G)$ where W is a chosen one affine parameter in the case of $relu$ activation function. Figure 4(a) shows the case of ordinary VI, which diverges as absolute value of x_1 become large. This means outliers have unlimited influence to the estimated static. On the other hand, Fig 4(b) shows the case of proposed method and the influence is bounded, that is the effect of outliers goes to zero. These results are compatible our theoretical analysis in the previous section.



(a) VI (b) Proposed method

Figure 4: IF of one affine parameter in Bayesian neural net.

However this is not sufficient analysis because what we want to be robust is the predictive distribution not parameters. Accordingly, it is necessary to study whether prediction is robust against outliers. For the analysis of prediction, we simulated the test log-likelihood. Actually, if the test log-likelihood has affected so much by an outlier, that is prediction on the test point is affected so much. Accordingly, such a model is not robust even under contamination of one outlier.

L.3 Bench mark dataset

In this experiment, we determined β and γ by cross validation. For both regression and classification settings, the range of β and γ are from 0.1 to 0.9.

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 Rényi 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 Student-t distribution, we chose the degree of freedom from 3 to 10 by cross-validation.

In both of the regression and classification problem, we artificially increased the percentage of both input and output related outliers in the training dataset.

To make the input related outliers, we first specified which features of the input we would contaminate. In this experiment, for regression tasks, since input dimension is not so large, we contaminated all the input features. For classification tasks, if the training data has D dimensional features, we randomly chose $D/2$ dimensions to contaminate. Next we randomly chose the data points we contaminate from training dataset. We contaminated the features by adding the Gaussian noise. Since the input data had been preprocessed by standardization, the noise we use is the Gaussian distribution which

follows $\epsilon \sim N(0, 6)$. From the numerical calculation of influence function, we considered that the noise which has the value of "6" sigma variance can be regarded as an outlier.

For output related outlier, in the same way as the input related outlier, we randomly chosen the point which we would contaminate and add the Gaussian noise which follows $\epsilon \sim N(0, 6)$.

We optimized by using Adam and reparameterization trick. The learning rate of Adam was set to 0.01 and MC samples was 5 except for covertype dataset. For the covertype dataset, the learning rate of Adam was set to 0.001 and we used 20 MC samples. The minibatch size was set to 128.

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.

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.

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.