Predictive variational Bayesian inference as risk-seeking optimization

Futoshi Futami[1] **Tomoharu Iwata**[1] **Naonori Ueda**[1] **Issei Sato**[2] **Masashi Sugiyama**[2]
[1]NTT Communication Science Laboratories [2]The University of Tokyo

Abstract

Since the Bayesian inference works poorly under model misspecification, various solutions have been explored to counteract the shortcomings. Recently proposed predictive Bayes (PB) that directly optimizes the Kullback Leibler divergence between the empirical distribution and the approximate predictive distribution shows excellent performances not only under model misspecification but also for over-parametrized models. However, its behavior and superiority are still unclear, which limits the applications of PB. Specifically, the superiority of PB has been shown only in terms of the predictive test log-likelihood and the performance in the sense of parameter estimation has not been investigated yet. Also, it is not clear why PB is superior with misspecified and over-parameterized models. In this paper, we clarify these ambiguities by studying PB in the framework of risk-seeking optimization. To achieve this, first, we provide a consistency theory for PB and then present intuition of robustness of PB to model misspecification using a response function theory. Thereafter, we theoretically and numerically show that PB has an implicit regularization effect that leads to flat local minima in over-parametrized models.

1 INTRODUCTION

Bayesian inference is a popular choice in statistics and machine learning as a stochastic modeling tool (Murphy, 2012). In Bayesian inference, we update a prior distribution that represent our assumptions to a Bayesian posterior distribution and obtain a predictive distribution with Bayesian model averaging. However, it has been theoretically and experimentally reported that the Bayesian framework has sub-optimal performance under some conditions.

For example, it is widely known that Bayesian inference works poorly when the model is misspecified (Grünwald, 2012; Masegosa, 2020). Nevertheless, many misspecified models are still useful for understanding the mechanism of the data (Wang et al., 2017). To perform inference in a Bayesian framework even with misspecified models, various methods have been proposed, including the tempered posterior (Grünwald, 2012; van Erven et al., 2015; Heide et al., 2020), which assigns a small weight to the likelihood function, and using maximum mean discrepancy (MMD) instead of the likelihood function (Chérief-Abdellatif and Alquier, 2020). These methods are a kind of “pseudo Bayes” (Bissiri et al., 2016) that aims to resolve the difficulty of model misspecification by changing the likelihood function in some way.

While these methods focus on parameter estimation, a method called predictive Bayes (PB), which focuses on the prediction performance, has been attracting much attention in recent years to address the model misspecification (Masegosa, 2020; Futami et al., 2021; Morningstar et al., 2020). PB minimizes the Kullback Leibler (KL) divergence between the approximate predictive distribution and true data generating distribution based on the PAC Bayesian theory (Germain et al., 2016). The likelihood function in standard Bayesian inference is replaced by the KL divergence between the empirical distribution and the approximate predictive distribution.

Masegosa (2020), Futami et al. (2021), and Morningstar et al. (2020) have theoretically shown that when a model is misspecified, PB shows a better predictive test log-likelihood than the exact Bayesian predictive distribution. In addition, unlike the tempered posterior, PB has shown superior empirical performance to standard Bayesian inference even with over-parametrized models, such as Bayesian neural networks (Masegosa, 2020; Futami et al., 2021).

However, despite these excellent properties, theoretical

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

and practical ambiguities remain in PB, and thus, its application to a wider range of fields is still limited. Following ambiguities remain: i) Although existing work assured a superior performance in terms of the predictive test log-likelihood, the performance in the sense of parameter estimation has not been investigated yet. ii) Although existing work showed that the PB's predictive distribution has a larger variance than the standard Bayesian predictive distribution, this does not explain why PB is robust under model misspecification.

These two understandings are vital in practical applications. As we discuss in Section 3, when the target task is regression, we are interested not only in the predictive test log-likelihood but also the mean squared error (MSE). In this scenario, when a model is misspecified, the predictive test log-likelihood can be maximized by increasing the variance of the predictive distribution, although the MSE remains significantly large. Therefore, applying the solution of PB can be counterproductive in these cases.

As for over-parameterized models, it is unclear why PB outperforms standard Bayesian inference. This question is an important since deep learning has been widely used in modern probabilistic modeling (Johnson et al., 2016). Another limitation of existing work is that applying PB is difficult to such latent variable models as variational autoencoders (Kingma and Welling, 2013), since those models are optimized by maximizing the marginal likelihood, which is different from the objective of PB.

In this paper, to clarify these ambiguities, we analyze PB by comparing its distorted loss function to a standard Bayesian inference. Note that distorted loss functions have been discussed in the field of risk-seeking optimization (Lee et al., 2020), which provides an elaborate analysis to reflect the variability of the loss functions in the problem. By focusing on PB's risk-seeking property, we present its theoretical properties in model misspecified and over-parameterized settings.

First, we show a parameter estimation property of PB by providing a frequentist consistency when the model is both well-specified and misspecified. We show that PB converges to the closest model to the true data generating distribution in the Hellinger distance. Next, we describe our intuition of PB's robustness to model misspecification using a response function theory (Lindsay, 1994). Then we numerically validate the superior performance of PB for regression tasks where standard Bayesian inference fails.

For over-parameterized models, we numerically and theoretically demonstrate that PB has an implicit regularization effect that guides the solution to a flat minimum, which is considered to have a higher generalization ability (Keskar et al., 2016). Finally, we extend PB to latent variable models by providing a novel lower bound of the marginal likelihood, which incorporates its risk-seeking property.

2 RELATED WORK

The tempered posterior has been studied theoretically and numerically as the "safe inference" under model misspecification. Heide et al. (2020) and Alquier and Ridgway (2020) proved its concentration and consistency properties, which are closely related to the parameter estimation performance. The most significant difference between PB and the tempered posterior is the implicit regularization effect, which is useful when the model is over-parametrized as shown in Section 4.5. Unlike the tempered posterior, the cold posterior, which downweights the prior distribution, empirically improves the performance in over-parametrized models (Wenzel et al., 2020).

Several existing works proposed to replace the likelihood in Bayesian inference with different loss functions. Futami et al. (2018) proposed robust variational inference that uses β- and γ-divergences to enhance robustness to the outliers in training data. Chérief-Abdellatif and Alquier (2020) used MMD as a loss function and showed the consistency property, which leads to robustness to model misspecification. We found that entropy-stochastic gradient descent (SGD) (Chaudhari et al., 2019, 2018) is a particular case of PB where a mean-field (MF) Gaussian distribution is used for variational posteriors. Then the loss is equivalent to the convolution of the original log-likelihood with Gaussian distribution. Thus, the surface of the loss function becomes smoother and leads to flat minima in over-parameterized models. They proposed a two-time scale approximation for the optimization and manually tuned the variance of the MF Gaussian distribution. On the other hand, PB uses a multi-sample bound and does not need to manually tune the variance parameter.

Existing risk-seeking methods in machine learning (Lee et al., 2020; Chow et al., 2015) tried to find a high variability solution concerning the training dataset, and their theoretical analysis considered a deterministic hypothesis. On the other hand, PB seeks a solution with high variability concerning the randomness of the posterior distribution.

3 PRELIMINARIES

Here we introduce this paper's notations and settings and briefly describe PB.

3.1 Notations and settings

Assume that a training dataset consists of N identically independently distributed (i.i.d) random variables

$(x_1,\dots,x_N) := x^N$ according to unknown data generating density $\nu(x)$ on $\mathcal{X} \subset \mathbb{R}^{d_X}$. Our goal is to model $\nu(x)$ using a parameterized statistical model $p(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^d$. If there exists parameter $\theta^* \in \Theta$ such that $\nu(x) = p(x|\theta^*)$, then our model is well-specified. If such a θ^* does not exist, the model is misspecified (Alquier and Ridgway, 2020). We express the log-loss as $l_n(\theta) := -\ln p(x_n|\theta)$ and $L_N(\theta) = \sum_{n=1}^N l_n(\theta)$. In Bayesian inference, we incorporate our prior knowledge into a prior distribution $\pi(\theta)$, which is updated to the Bayesian posterior distribution as $p(\theta|x^N) \propto \exp(-L_N(\theta))\pi(\theta)$. Since calculating the exact posterior is computationally infeasible for many practical models, we need to rely on approximation methods. Variational inference (VI) (Attias, 1999; Beal, 2003), which approximates the exact posterior by a predefined parametric distribution, has been widely used for this purpose because of its computational efficiency. We express the approximate posterior distribution as $q(\theta;\phi)$, where ϕ is called a variational parameter. The standard VI optimizes ϕ by minimizing the objective:

$$\mathrm{Obj}_{\mathrm{VI}}(\phi) := \frac{1}{N}\mathbb{E}_{q(\theta;\phi)}[L_N(\theta)] + \frac{\mathrm{KL}(q \mid \pi)}{\alpha N}. \qquad (1)$$

Here $\mathrm{KL}(q|\pi)$ is the KL divergence and α is the temperature. When $\alpha = 1$, $-\mathrm{Obj}_{\mathrm{VI}}(\phi)$ is a lower bound of marginal likelihood $\ln p(x^N) \geq -\mathrm{Obj}_{\mathrm{VI}}(\phi)$. When $\alpha < 1$, the solution is called a tempered posterior distribution. We obtain the variational parameter:

$$\phi^*_{\mathrm{VI}} = \mathrm{argmin}_\phi \mathrm{Obj}_{\mathrm{VI}}(\phi). \qquad (2)$$

Then the predictive distribution is given by $p(x;\phi^*_{\mathrm{VI}}) = \mathbb{E}_{q(\theta;\phi^*_{\mathrm{VI}})}p(x|\theta)$. Note that in the objective function of VI, Eq.(1) is decomposed to loss function $\mathbb{E}_q[L_N(\theta)]$ and regularization term $\mathrm{KL}(q|\pi)$. The loss function in VI corresponds to $\mathrm{KL}(\hat{\nu}(x)|p(x|\theta))$, where $\hat{\nu}(x) := \frac{1}{N}\sum_{n=1}^N \delta_{x_n}(x)$ and $\delta_{x_n}(x)$ is the Dirac mass at x_n.

The theoretical property of $q(\theta;\phi^*_{\mathrm{VI}})$ has widely been investigated. For example, Alquier and Ridgway (2020) clarified that the variational posterior distribution can concentrate on true parameter θ^* in the well-specified case. Thus, when we use the variational posterior as an estimator of the parameter, our statistical model can converge to true model $p(x|\theta^*)$. When the model is misspecified, they also proved that the tempered posterior converges to the parameter that minimizes $\mathrm{KL}(\nu(x)|p(x|\theta))$. In this sense, parameter estimation performance in VI is well understood.

3.2 Predictive Bayes

Masegosa (2020), Futami et al. (2021), and Morningstar et al. (2020) directly minimized the KL divergence between the approximate predictive distribution and data generating distribution:

$$\mathrm{KL}(\nu(x)|\mathbb{E}_{q(\theta;\phi)}p(x|\theta)) = \mathbb{E}_{\nu(x)}[-\ln \mathbb{E}_{q(\theta;\phi)}p(x|\theta)] + \mathrm{Const.}$$

Since the data generating distribution is unknown, we approximate it with a training dataset. Motivated by the PAC Bayesian theory (Germain et al., 2016), they added regularization term $\mathrm{KL}(q|\pi)$. The final objective function is

$$\mathrm{Obj}_{\mathrm{PB}}(\phi) := \frac{1}{N}\sum_{n=1}^N[-\ln \mathbb{E}_{q(\theta;\phi)}p(x_n|\theta)] + \frac{\mathrm{KL}(q \mid \pi)}{\alpha N}. \qquad (3)$$

We refer to this approach as predictive variational Bayesian inference (PB-VI). Note that by using the Jensen inequality, we have

$$\mathbb{E}_{\nu(x)}[-\ln \mathbb{E}_{q(\theta)}p(x|\theta)] \leq \mathbb{E}_{\nu(x),q(\theta)}[-\ln p(x|\theta)]. \qquad (4)$$

Thus the objective Eq.(3) is a tighter bound than Eq.(1). Masegosa (2020), Futami et al. (2021), and Morningstar et al. (2020) claimed that this tightness improves the performance, especially when the model is misspecified. Moreover, Masegosa (2020) and Futami et al. (2021) argued that the tight bound creates the diversity-enhancing term in ensemble learning and particle VI (Wang et al., 2019). To optimize Eq.(3), previous works used a multi-sample bound. Morningstar et al. (2020) directly optimized it by relying on a biased Monte Carlo objective like the importance weighted autoencoder (IWAE) (Burda et al., 2015):

$$-\ln \mathbb{E}_{q(\theta)}p(x|\theta) \leq -\mathbb{E}_{\theta_1\dots\theta_m\sim q(\theta)}\frac{1}{m}\sum_{m'=1}^m \ln p(x|\theta_{m'}). \qquad (5)$$

On the other hand, Masegosa (2020) and Futami et al. (2021) considered the upper bound of Eq.(3) using the second-order Jensen inequality:

$$\mathrm{Obj}_{\mathrm{MV}}(\phi) := \frac{1}{N}\mathbb{E}_q L_N(\theta) - \frac{c}{N}\sum_{n=1}^N \mathrm{Var}_q(l_n(\theta)) + \frac{\mathrm{KL}(q|\pi)}{\alpha N}, \qquad (6)$$

where $\mathrm{Var}_q(l_n(\theta))$ is the variance of $l_n(\theta)$ with respect to q and c is a coefficient of the variance. See Appendix B.1 for the details. We refer to this approach as mean-variance VI (MV-VI).

Masegosa (2020), Futami et al. (2021), and Morningstar et al. (2020) proposed a PAC Bayes generalization error bound for $\mathrm{KL}(\nu|\mathbb{E}_q p(x|\theta))$ and also demonstrated promising numerical performances of PB-VI not only in misspecified settings but also for over-parametrized models as discussed in Section 1.

One major drawback of PB is that its generalization performance is only assured for the predictive test log-likelihood. As mentioned in Section 1, this is problematic for regression tasks under the misspecified setting.

For example, consider $(x,y) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{d_{\mathcal{X}}} \times \mathbb{R}$. Assume that the true data generating distribution is $\nu(x,y) = \mathcal{N}(y|R(x),\rho^2)\nu(x)$, where $\mathcal{N}(y|R(x),\rho^2)$ is the Gaussian distribution with mean $R(x): \mathcal{X} \to \mathbb{R}$ and variance $\rho^2 \in \mathbb{R}$. Assume that our model is $p(y|x,\theta) = \mathcal{N}(y|r_\theta(x),\sigma^2)$. We predict y given x with $y = \mathbb{E}_{q(\theta;\phi)}[r_\theta(x)]$. Under this setting, Watanabe (2009) showed that

$$\mathbb{E}_\nu \mathrm{KL}(\nu(y|x)|\mathbb{E}_q p(y|x,\theta)) = \frac{\mathbb{E}_\nu\|R(x)-\mathbb{E}_q[r_\theta(x)]\|^2}{2\sigma^2} + \frac{1}{2}\ln\frac{\sigma^2}{\rho^2}$$
$$+\mathbb{E}_\nu\frac{\sigma^2-\rho^2}{2\sigma^4}(\mathrm{Var}_q[r_\theta(x)]-1) - \frac{1}{2\sigma^4}\mathrm{Var}_q[r_\theta(x)G(x)+\|r_{\theta^*}-r_\theta\|^2],$$

where $G(x) := (R(x) - r_{\theta^*}(x))^2$. Note that the first term in the right-hand side corresponds to the MSE. When the model is misspecified, e.g., $\rho^2 \neq \sigma^2$ or $R(x) \neq r_{\theta^*}(x)$, we can increase the test log-likelihood by increasing the variance of the predictive distribution, which corresponds to the third and fourth terms in the right-hand side while the MSE is not changed. Since a large variance in PB was numerically reported in Masegosa (2020) and Futami et al. (2021), there is a possibility that PB achieves a high test log likelihood with the large MSE, which is unsuitable for prediction. Thus, it is unclear whether PB really works well in regression tasks.

To guarantee the MSE of PB under model misspefication, the consistency property, which is closely related to parameter estimation performance, should be assured. However, existing work only clarified the PB's performance in the predictive distribution, which is averaged over the approximate posterior distributions.

4 METHOD

Here we present our main theory. First, we point out the relation of PB-VI to risk-seeking optimization (RSO) in Section 4.1 and present the theory for PB-VI in model misspecification in Sections 4.2 and 4.3. Finally, we present the results for over-parameterized settings in Sections 4.5 and 4.6.

4.1 Risk-seeking objective function

When focusing on the loss function, the loss function of VI corresponds to $\sum_n \mathbb{E}_q l_n(\theta)$. On the other hand, the loss function of PB-VI is $\sum_n -\ln \mathbb{E}_q e^{-l_n(\theta)}$. This observation indicates that original loss l_n in VI is exponentially distorted in PB-VI. The idea of distorting the objective function is extensively studied in RSO (Lee et al., 2020). The exponentially distorted loss is known as an entropic-risk in RSO, which is defined as $-\frac{1}{\gamma}\ln \mathbb{E}_{q(\theta;\phi)} e^{-\gamma l_n(\theta)}$, where $\gamma \in \mathbb{R}^+$. We propose using entropic-risk as a loss function in VI for theoretical analysis:

$$\mathrm{Obj}_{\mathrm{Ent}}(\phi) := \frac{1}{N}\sum_{n=1}^N[-\frac{1}{\gamma}\ln \mathbb{E}_{q(\theta;\phi)} e^{-\gamma l_n(\theta)}] + \frac{\mathrm{KL}(q|\pi)}{\alpha N}. \quad (7)$$

We refer to this as entropic-risk VI (Ent-VI), which is equivalent to PB-VI in Eq.(3) when $\gamma = 1$. We denote the the solution of Ent-VI as

$$\phi^*_{\mathrm{Ent}} = \mathrm{argmin}_\phi \mathrm{Obj}_{\mathrm{Ent}}(\phi) \quad (8)$$

and we express the optimized variational posterior distribution as $q^{\mathrm{Ent}}(\theta) := q(\theta;\phi^*_{\mathrm{Ent}})$.

Interpreting PB as RSO has the intuitive and theoretical advantages. We describe the intuition in Section 4.4 later. As for the theoretical advantages, we can easily analyze PB by utilizing techniques in the field of RSO. In particular, we utilize the fact that objective functions of RSO have dual forms, which are easier to analyze theoretically. Another motivation is that, we can analyze not only Ent-VI but also MV-VI in a unified manner. In particular, the loss in MV-VI in Eq.(6) has also been widely used in RSO and is often called a mean-variance (MV) loss. The MV loss is particularly important in RSO since various risk-seeking objective functions can be upper or lower bounded by it (Lee et al., 2020). Due to page limitations, we analyze MV-VI in Appendix I.

We analyze PB-VI based on its variational dual form (Ben-Tal et al., 1991). For an entropic-risk, we have

$$-\frac{1}{\gamma}\ln\mathbb{E}_{q(\theta)}e^{-\gamma l_n(\theta)} = \inf_{q'} \mathbb{E}_{q'(\theta)}[l_n(\theta)] + \frac{1}{\gamma}\mathrm{KL}(q'(\theta)|q(\theta;\phi)), \quad (9)$$

where the infimum is taken for all the probability measures that are absolutely continuous regarding q. Thus, the Ent-VI problem can be expressed in a dual from

$$\inf_\phi \inf_{q'} \frac{1}{N}\sum_{n=1}^N[\mathbb{E}_{q'(\theta)}[l_n(\theta)] + \frac{1}{\gamma}\mathrm{KL}(q'|q)] + \frac{\mathrm{KL}(q|\pi)}{\alpha N}. \quad (10)$$

This is a double-loop optimization, where the inner optimization addresses the variational problem of Eq.(9) and the outer loop is the optimization concerning variational parameter ϕ. We used Eq.(10) to analyze the convergence properties of Ent-VI in the misspecified model and its implicit regularization effect in the over-parameterized model in Sections 4.2 and 4.5.

4.2 Convergence property of Ent-VI

Here, we analyze the parameter estimation property of PB-VI and the guarantee for the MSE loss as discussed in Section 3.2. First, we study the consistency of the posterior distribution since it provides the frequentist guarantee for the approximate posteriors.

We define the discrepancy of the misspecification using the KL divergence. We define $\theta^* := \mathrm{argmin}_{\theta\in\Theta}\mathrm{KL}(\nu(x)|p(x|\theta))$ and write:

$$\mathrm{KL}(\nu(x)|p(x|\theta)) = \mathrm{KL}(\nu(x)|p(x|\theta^*)) + \mathbb{E}\ln\frac{p(x|\theta^*)}{p(x|\theta)}. \quad (11)$$

The first term corresponds to the error caused by the misspecification. Then, we introduce the following assumptions.

Assumption 1. *Assume there exists $\epsilon_N > 0$ for which there is a variational distribution $q(\theta;\phi_N)$ such that:*

$$\int \mathbb{E}_{\nu(x)}\ln\frac{p(x|\theta^*)}{p(x|\theta)}q(\theta;\phi_N)d\theta \leq \epsilon_N, \quad (12)$$

$$\mathrm{KL}(q(\theta;\phi_N)|\pi) \leq N\epsilon_N. \quad (13)$$

These assumptions, Eqs.(12) and (13) are the same as a previous work (Alquier and Ridgway, 2020). Intuitively, this assumption implies that a prior gives sufficient mass near the true parameter, and a family of $q(\theta;\phi_N)$ contains distributions that are concentrated near the true parameter. We also introduce assumptions about $q^{\mathrm{Ent}}(\theta)$, which is required to control the dual problem:

Assumption 2. *i) The entropy of $q^{\mathrm{Ent}}(\theta)$ is upper-bounded by a positive constant H_0, $\mathrm{E}_{q^{\mathrm{Ent}}(\theta)}[-\ln q^{\mathrm{Ent}}(\theta)] \leq H_0$. ii) There exist a positive constant $\lambda_0, \tilde{\sigma}^2$ such that for all $0<\lambda<\lambda_0$, $\mathrm{E}_{q^{\mathrm{Ent}}(\theta)}e^{\lambda\ln\pi} \leq e^{\lambda^2\tilde{\sigma}^2}$.*

Assumption 2 are satisfied when we choose q and π appropriately. For example, when q and π are both Gaussian distributions, Assumption 2 holds. We further discuss assumptions in Appendix F. With this assumption, we first present the consistency property of Ent-VI for a misspecified setting:

Theorem 1. *For a given $\alpha \in (0,1)$ and $\gamma \in (0, 2N/(4N+\alpha))$, under Assumption 1 and 2, $q^{\mathrm{Ent}}(\theta) := q(\theta;\phi^*_{\mathrm{Ent}})$ satisfies*

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(\nu(x), p(x|\theta))q^{\mathrm{Ent}}(\theta)d\theta$$
$$\leq \mathrm{KL}(\nu(x)|p(x|\theta^*)) + \frac{1+\alpha}{\alpha}\epsilon_N + \frac{2\tilde{\sigma}^2+H_0^2}{N\alpha^2}, \quad (14)$$

where Hel^2 is the Hellinger distance defined in Appendix A.

The proof is shown in Appendix E. When the model is well-specified, the first term disappears. If $\epsilon_N \to 0$ as $N \to \infty$, the Ent-VI is consistent.

Here we cite an example of ϵ_N from Alquier and Ridgway (2020). Assume that q is MF Gaussian $q(\theta;\phi) = N(\theta;\mu,\sigma^2 I_d)$ and $\pi(\theta) = N(\theta;0,\sigma_0^2 I_d)$, where I_d is the d-dimensional identity matrix. Moreover, we assume that there exists a measurable function $M(x)$ for $p(x|\theta)$ that satisfies

$$|\ln p(x|\theta) - \ln p(x|\theta')| \leq M(x)\|\theta-\theta'\| \quad (15)$$

and furthermore we assume that $\mathbb{E}_\nu M(x) := L < \infty$, then we have

$$\epsilon_N = \tfrac{L}{N} \vee \left\{\tfrac{d}{N}\left[\tfrac{1}{2}\ln(\sigma_0^2 N^2 d^{1/2}) + \tfrac{1}{N\sigma_0^2}\right] + \tfrac{\|\theta^*\|}{N\sigma_0^2} - \tfrac{d}{2N}\right\}. \quad (16)$$

This holds for a linear and logistic regression model, see Alquier and Ridgway (2020) for details. Thus, substituting this ϵ_N, the convergence rate is $\mathcal{O}(\ln N/N)$ for the Ent-VI, which shows the same convergence rate as tempered posteriors. This rate is a minmax optimal within a log-factor. See Appendix H for a detailed comparison.

Using this consistency property, we provide a guarantee in the MSE as discussed in Section 3. Our statistical model is $p(y|x,\theta) = N(y|r_{\theta_0}(x), \sigma^2)$, where $\theta = (\theta_0, \sigma^2)$ and $r_{\theta_0}(x)$ is linear in parameters. We define θ^* as $\theta^* := \mathrm{argmin}_{\theta\in\Theta}\mathbb{E}_{\nu(x)}\mathrm{KL}(\nu(y|x)|p(y|x,\theta))$. We do not assume that $\nu(y|x)$ is a Gaussian distribution. Instead, we introduce the assumptions below:

Corollary 1. *Assume Θ and $\mathcal{X}$ are compact, and there exists a constant η such that $\sup_{x\in\mathcal{X}}\mathbb{E}_{\nu(y|x)}e^{\eta|y|} < \infty$, and $\mathbb{E}_{\nu(y|x)}[y] = r_{\theta_0^*}(x)$. Then, under the assumptions in Theorem 1, there exist constants $c_1, c_2 > 0$, which only depends on the parameters of the problem*

$$\mathbb{E}_{\nu(x,y)}\int\left(\frac{\mathbb{E}_{\nu(x)}\|r_{\theta_0^*}(x) - r_{\theta_0}(x)\|^2}{2\sigma^2} + \frac{1}{2}\ln\frac{\sigma^2}{\sigma^{2*}}\right)q^{\mathrm{Ent}}(\theta)d\theta$$
$$\leq (c_1 + c_2\ln N)\left(\frac{1+\alpha}{\alpha}\epsilon_N + \frac{2\tilde{\sigma}^2+H_0^2}{N\alpha^2}\right). \quad (17)$$

This is the direct consequence of Heide et al. (2020), see Appendix G. We remark that this corollary is not restricted to when $p(y|x,\theta)$ is a Gaussian distribution. The similar statement also holds for a generalized linear model, see Appendix G for details. Note that the assumption requires $\mathbb{E}_{\nu(y|x)}[y] = r_{\theta_0^*}(x)$. Intuitively, this corollary provides the guarantee even when the noise function is misspecified. In Section 5, we confirm this by considering a Bernoulli heteroscedastic noise for $\nu(y|x)$ and a Gaussian homoscedastic noise for $p(y|x,\theta)$. With this corollary, we can upper-bound the MSE. Note that if $\epsilon_N \to 0$ as $N \to \infty$, the estimated regression model converges to the optimal model in the MSE. When using Eq.(16), we obtain an $\mathcal{O}((\ln N)^2/N)$ bound, which is only a log factor worse than Theorem 1.

In the previous work, e.g, Masegosa (2020), the performance of PB-VI is guaranteed in $\mathrm{KL}(\nu(x)|\mathbb{E}_{q^{\mathrm{Ent}}(\theta)}p(x|\theta))$, where the model is averaged inside the KL divergence. On the other hand, our Theorem 1 and Corollary 1 guarantee the performance when the expectation with respect to $q^{\mathrm{Ent}}(\theta)$ is outside the Hellinger distance or the MSE. Using the convexity of f-divergences and the MSE, we can move the expectation inside the Hellinger distance and the MSE. Then our theoretical guarantee about the consistency and the MSE holds for the approximate predictive distribution $\mathbb{E}_{q^{\mathrm{Ent}}(\theta)}p(x|\theta)$ similarly.

4.3 Analysis based on response theory

Here we study Ent-VI's the robustness to model misspecification using the associated residuals and the response function (Lindsay, 1994). These concepts are utilized to investigate the robustness of the Hellinger distance, which is appeared in the consistency propoerty of Ent-VI in Theorem 1. First, we define the Pearson residual (PR) at x:

$$\delta(x,\theta) := (\nu(x) - p(x|\theta))p(x|\theta)^{-1}. \tag{18}$$

Lindsay (1994) clarified that the maximum likelihood (ML) estimator of $\min_\theta \mathrm{KL}(\nu(x)|p(x|\theta))$ can be expressed as $\nabla_\theta \mathrm{KL}(\nu(x)|p(x|\theta)) = \int \delta(x,\theta)\nabla p(x|\theta)dx +$ Const. The coefficient of $\nabla p(x|\theta)$ is called the residual adjustment function (RAF), which is defined as a function of δ. Thus, the RAF of ML is $\delta(x,\theta)$ and shows a linear response to PR. The intuition of RAF is that since δ is the relative deviation, model misspecification implies a large δ. Thus, robustness to model misspecification requires a damped response as an increasing δ. Lindsay (1994) argued that since the RAF of the Hellinger distance minimization shows $\sqrt{\delta+1}-1$, it is more robust than the KL divergence. Thus, to investigate the robustness, our goal is comparing the RAFs of the Ent-VI and the standard VI.

Since the loss of the standard VI is $\mathbb{E}_{q(\theta;\phi)}[\mathrm{KL}(\nu(x)|p(x|\theta))]$, we derive its RAF by differentiating it with respect to ϕ. We use a reparameterized gradient with $\theta \sim T(\theta_0;\phi)$, where T is a translation, ϕ is a variational parameter, and θ_0 obeys a simple distribution $p(\theta_0)$ like a Gaussian distribution. More details are discussed in Appendix J. We use an m-sample bound, and we draw $\theta_0^1 \dots \theta_0^m$ from $p(\theta_0)$ and obtain $\theta_{m'} = T(\theta_0^{m'};\phi)$. Then, we have

$$\begin{aligned}&\nabla_\phi \mathbb{E}_{q(\theta;\phi)}\mathrm{KL}(\nu(x)|p(x|\theta))\\&=\mathbb{E}_{\theta_0^1\dots\theta_0^m\sim p(\theta_0)}\int \frac{1}{m}\sum_{m'=1}^m \frac{\nu(x)}{p(x|\theta_{m'})}G(\theta_0^{m'})dx,\end{aligned} \tag{19}$$

where $G(\theta_0^{m'}) := \nabla_\phi T(\theta_0^{m'};\phi)\nabla_\theta p(x|\theta_{m'})$. Since $\nabla_\phi T$ is determined by the choice of the variational distribution, we define the RAF of the standard VI as $\frac{\nu(x)}{p(x|\theta_{m'})} := \delta + 1$. This shows a linear response like ML estimation. As for Ent-VI, we use an m-sample bound as in Eq.(5). To eliminate the summation's non-linearity, we approximate the loss of Ent-VI:

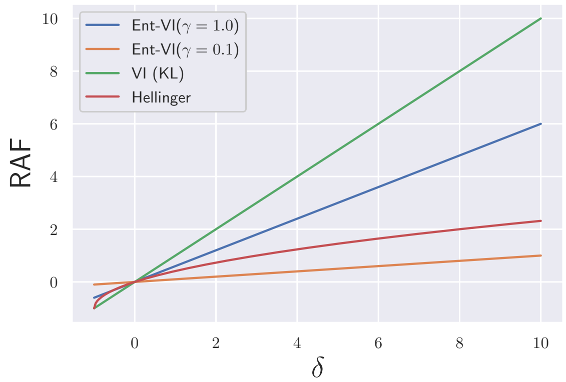

Figure 1: RAF of different methods

$$\ln \sum_{m'=1}^m e^{\gamma \ln w(\theta_{m'})} \approx \ln w(\tilde{\theta}) + \sum_{m'=1}^m \left(\frac{w_{\theta_{m'}}}{w(\tilde{\theta})}\right)^\gamma,$$

where we define $\tilde{\theta} := \mathrm{argmax}_{\theta_{m'}} \ln w(\theta_{m'})$ and $w(\theta) := p(x|\theta)/\nu(x)$. The derivative of loss function is

$$\begin{aligned}&\nabla_\phi \mathbb{E}_{\nu(x)}\left(-\gamma^{-1}\ln \mathbb{E}_{q(\theta;\phi)}e^{\gamma \ln p(x|\theta)}\right)\\&=\mathbb{E}_{\theta_0^1\dots\theta_0^m\sim p(\theta_0)}\int \frac{1}{m}\sum_{m'=1}^m R(x,\theta_{m'})G(\theta_0^{m'})dx,\end{aligned} \tag{20}$$

where an explicit form of $R(x,\theta_{m'})$ is shown in Appendix J. When $\gamma > 1$, we can approximate it as $R(x,\theta_{m'}) \approx \nu(x)/p(x|\tilde{\theta})$. Since $\tilde{\theta}$ is a sample that fits the model best among m drawn samples, the Ent-VI's RAF is much smaller than that of the standard VI, although both show linear responses. In Figure 1, we visualized RAF as a function of δ in the standard VI and the Ent-VI using the model of the toy data experiment in Section 5. The Ent-VI's RAF shows a low response compared to VI and we believe that this is the key mechanism of Ent-VI for a robustness to model misspecification. See Appendix J for the detailed settings.

Response analysis fuels the intuition of Ent-VI as a risk-seeking objective. If an approximate posterior has a large variance, we might draw a sample that fits better than a small variance approximate posterior. Thus, this increases the variance of the approximate posterior since the loss of the best fitting sample only matters to the estimator as shown in Eq.(20) from the response analysis. In this way, PB-VI and MV-VI show large variance as reported in Futami et al. (2021) and Morningstar et al. (2020).

4.4 Intuition of the risk seeking property

Following the response analysis in Section 4.3, here we briefly describe the intuition of the risk-seeking property of PB. By interpreting PB as RSO, it becomes clear why PB has a larger variance in the posterior distribution than ordinary VI or Bayesian inference since the standard VI corresponds to the risk-neutral objective function. This intuition is difficult to obtain from the original PB objective function in Eq. (3).

Moreover, this large PB variance may explain PB's good performance in over-parametrized models and model misspecification. For over-parametrized models, the solution with the larger variance means the flatter local minima, which gives an intuitive explanation for PB's good performance in such a model. We present an analysis for this in Section 4.5. As for the model misspecification setting, we describe its inuition in Appendix K.

4.5 Implicit regularization of Ent-VI

In Section 4.2 and 4.3, we have focused on the parameter estimation and robustness to model misspecification of PB-VI, where we assume that models are rather simple. On the other hand, recent Bayesian models can successfully incorporate deep learning methods (Johnson et al., 2016). In such a situation, we are more interested in finding a local minimum that shows a good generalization performance than finding the global minimum.

It has been hypothesized that local minimums with flat geometry show better generalization performance than sharp local minimums (Keskar et al., 2016). Thus, we hypothesize that the good performance of PB-VI (Ent-VI) compared to the standard VI has a relation to a flat minimum. Intuitively, since a flat minimum can be specified with lower precision than a sharp minimum (Keskar et al., 2016), a local minimum of a large variance obtained by a risk-seeking objective have some relation to a flat minimum.

We elaborate this intuition using the dual problem in Eq.(10). The entire derivation is shown in Appendix L. We focus on a Bayesian neural network (BNN) and consider the MF approximation with $q(\theta;\phi) := N(\theta;\mu_1,\sigma^2 I_d)$. We restrict q' in Eq.(10) to parametric distributions and consider the MF Gaussian distribution: $q'(\theta;\phi') := N(\theta;\mu_2,\sigma^2 I_d)$. Here for simplicity, we assume that the variances of q and q' are identical. We express the difference of the mean of q and q' as $s = \mu_2 - \mu_1$. Then the samples from q' can be written as $s + \mu_1 + \sigma\xi$, where ξ is drawn from $N(\xi|0, I_d)$. Then, Eq.(10) is approximated:

$$\inf_{\mu,\sigma^2,s} \frac{1}{N}\sum_{n=1}^{N}\left[\mathbb{E}_\xi[l_n(\mu_1+s+\sigma\xi)]+\frac{\|s\|^2}{2\gamma\sigma^2}\right]+\frac{\mathrm{KL}(q|\pi)}{\alpha N}. \quad (21)$$

Since we restrict q' to a specific parametric distribution, the objective function of Eq.(10) is upper-bounded by Eq.(21). We first solve the inner problem concerning s and assume that s is small enough. Then using the Taylor expansion, the inner problem can be solved:

$$\begin{aligned}&\inf_s \mathbb{E}_\xi[l_n(\mu_1 + s + \sigma\xi)] + (2\gamma\sigma^2)^{-1}\|s\|^2 \\ &\approx l_n(\mu_1) + \mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2/2 - \nabla l_n(\mu_1)^\top H \nabla l_n(\mu_1), \quad (22)\end{aligned}$$

where $s = H\nabla l_n(\mu_1)$ and $H := [\gamma^{-1}\sigma^{-2}I_d + \nabla^2 l_n(\mu_1)]^{-1}$. On the other hand, the loss of the standard VI can also be expanded by the Taylor theorem:

$$\mathbb{E}_{q(\theta;\phi)}[l_n(\theta)] \approx l_n(\mu_1) + \mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2/2. \quad (23)$$

Comparing Eqs.(22) and (23), Eq.(22) of Ent-VI has additional regularization term $-\nabla l_n(\mu_1)^\top H \nabla l_n(\mu_1)$. This implicit regularization pushes the solution to a flat minimum as follows. Denote the eigenvalues and eigenvectors of H as (λ_i, v_i). Then $-\nabla l_n(\mu_1)^\top H \nabla l_n(\mu_1) = -\sum_i (\nabla l_n^\top v_i)^2 \lambda_i$ and this is a weighted sum of the eigenvalues of H. Thus the additional regularization term in Ent-VI provides a solution such that the weighted sum of λ_i increases. On the other hand, since $H := [\sigma^{-2}I_d + \nabla^2 l_n(\mu_1)]^{-1}$, increasing the weighted sum corresponds to a decrease in the eigenvalues of $\nabla^2 l_n(\mu_1)$, which is the Hessian of our loss function when σ is fixed. Since low eigenvalues of H are related to flat geometry, Ent-VI has implicit regularization that leads to a flatter minimum than the standard VI. In Section 5, we numerically compare the generalization performance of the standard VI and Ent-VI.

4.6 Novel lower bounds of the marginal likelihood

Based on the discussion about the implicit regularization of PB-VI in Section 4.5, introducing those risk-seeking property even when learning latent variable models (LVMs) seems promising to improve the performance. When learning LVMs, we are often interested in maximizing a marginal likelihood, not the distorted risks. For example, many deep latent models optimize the lower bound of the marginal likelihood. Therefore, we present a novel objective that is a valid lower bound of the marginal likelihood and still incorporates the risk-seeking property. We use a multi-sample bound and express z_j as the j-th random variable drawn from latent posterior distribution $q(z)$. We define $w_j = p(x|z_j)\pi(z_j)/q(z_j)$. We regard this $\ln w$ as a risk function and apply an MV loss, as in Eq.(6). We present two novel lower bounds. The first one is

$$\ln p(x) \geq \mathbb{E}_q \frac{1}{J}\sum_{j=1}^{J}\Big(\log w_j + h_x^{w_j}\Big(\ln w_j - \sum_{j=1}^{J}\frac{\ln w_j}{J}\Big)^2\Big), \quad (24)$$

where h_x^w are the weights of the variances, which depend on x and w, see Appendix M for the detailed expression and the proof. We refer to Eq.(24) as VI-VAR since the first term corresponds to the objective of the standard VI. Thus, thanks to the weighted variance term, VI-VAR bound is tighter than the standard VI objective.

The second bound is

$$\ln p(x) \geq \mathbb{E}_q \frac{1}{M}\sum_{m=1}^{M} \Big(\ln \sum_{j=1}^{J} \frac{w_j^m}{J} + h_x^{w_j^m}\Big(\ln \sum_{j=1}^{J} \frac{w_j^m}{J} - \frac{1}{M}\sum_{m=1}^{M} \ln \sum_{j=1}^{J} \frac{w_j^m}{J}\Big)^2\Big), \tag{25}$$

where $w_j^m := p(x|z_j^m)\pi(z_j^m)/q(z_j^m)$ and $\{z_j^m\}$ are JM samples drawn from $q(z)$ and h_x^w are the weights of the variances, which depend on x and w, see Appendix M for the detailed expression and the proof. This bound is tighter than the IWAE bound due to the weighted variance term of the second term. We refer to this as IWAE-VAR.

Thus, these bounds include the weighted variance term, which induces a risk-seeking property like PB. They are rigorous lower bounds of the marginal likelihood and tighter than the standard VI and IWAE bounds. In Section 5, we use them to optimize the variational autoencoder models (VAEs).

5 NUMERICAL EXPERIMENTS

We numerically confirmed the Ent-VI properties in model misspecified and over-parameterized settings. See Appendix N for details of the experiments and additional results.

5.1 Model misspecification

We addressed a toy data regression task and measured its performance in the MSE. Motivated by Heide et al. (2020), we considered a task where the standard VI completely fails. We generated toy-data $(x_i, y_i)_{i=1}^N$ as follows: $x_i = \zeta_i' u_i$ and $y_i = \zeta_i' \zeta_i$ where $u_i \sim \mathrm{Uniform}[-1, 1]$, $\zeta_i \sim N(0, 1)$, and ζ_i' follows a Bernoulli distribution, which takes 0 with a probability 1/2. Thus, the true conditional expectation is $\mathbb{E}[Y|X] = 0$. As our model, we used $N(y|f(x|w), \sigma^2)$ with a linear model of Fourier basis $f(x|w) = \frac{1}{\pi}\sum_{k=0}^{40} w_k^0 \cos(kx) + w_k^1 \sin(kx)$. Following the theory from Section 4.2, our model includes $\mathbb{E}[Y|X] = 0$, although the noise assumption is misspecified. Prior distributions are imposed on w and σ^2. For $\pi(w)$, we assumed a Laplace distribution, which induces sparsity. We used the non-centered parameterization and used MF Gaussian posteriors, see Appendix N for details. For example, when $N = 40$, we visualized the obtained functions in the upper row of Figure 2 and found that VI showed severe over-fitting and Ent-VI ($\gamma = 0.1$) suffers less over-fitting. In the figure, VI-NN is a result of a two-layer ReLU BNN and VI-NN fit the true regression function well in-domain areas. Thus, when a model is misspecified, standard VI fails in this setting. Next, we studied the MSE changing N. The results are shown in Figure 2 in the bottom row. Ent-VI outperformed in the MSE compared to the VI. A strange convergence behavior of VI was reported in a previous work (Heide et al., 2020); Ent-VI converged well. We show additional numerical experiments including the real data experiments in the Appendix N.1.

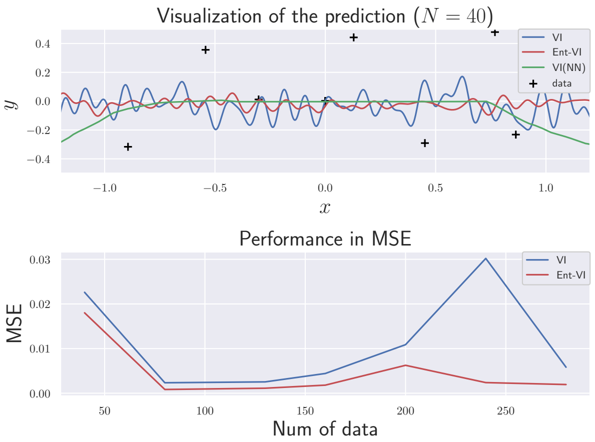

Figure 2: Regression results with a misspecified model

5.2 Flat minimum in deep learning

We numerically validated the implicit regularization of Ent-VI shown in Section 4.5 by following previous work (Masegosa, 2020; Morningstar et al., 2020). We considered a structured prediction task and trained a BNN to predict the bottom half of an image using its top half as input. We used a two-layer MLP with 20 hidden units with ReLU activation and considered an MF Gaussian for the posterior. We used the Fashion MNIST and the CIFAR10 datasets. To observe the flatness of the obtained solutions, under different temperature αs, we measured $\mathrm{KL}(q(\theta;\phi^*)|\pi)$ where ϕ^* is a solution of VI and Ent-VI and the sum of the squared Frobenius norms of each weight matrix. We chose these indicators since they show strong correlations to the generalization performance (Tsuzuku et al., 2020; Jiang et al., 2019). A small $\mathrm{KL}(q(\theta;\phi^*)|\pi)$ with a large test log-likelihood means that an obtained model needs less information in the posterior for a reasonable prediction. This is closely related to the principle of the minimum description length (MDL) theory, which states that a statistical model that requires a fewer rate shows better generalization, and thus, MDL is closely related to the flat minimum. The results are shown in Figure 3. Ent-VI requires a smaller KL divergence and a Frobenius

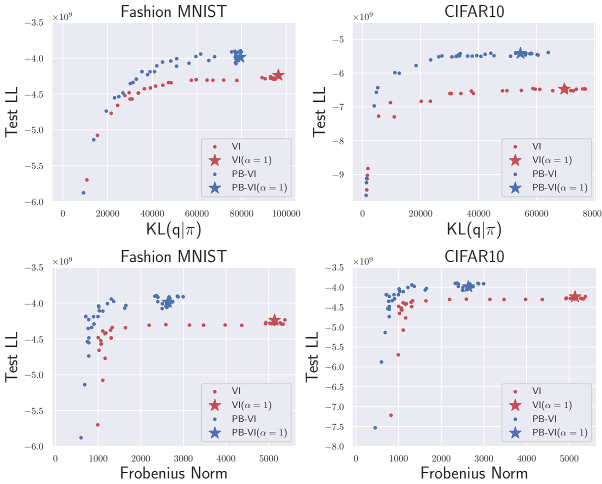

Figure 3: Left column is the result of Fashion MNIST and right column is CIFAR10. LL indicates the log-likelihood. Horizontal line in upper row indicates $\mathrm{KL}(q(\theta;\phi^*)|\pi)$. Low indicates squared sum of Frobenius norm of weight matrices. Stars indicate $\alpha = 1$.

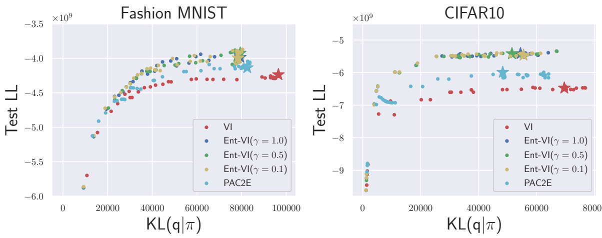

Figure 4: The same experiments as Figure 3 using different PB-VI methods. Stars indicate $\alpha = 1$.

norm to achieve the same test log-likelihood with VI. Next, we compare the relation of the test log-likelihood and $\mathrm{KL}(q(\theta;\phi^*)|\pi)$ using different PB-VI methods in Figure 4. We used different γs for Ent-VI and used the method of Masegosa (2020) as MV-VI, which is denoted by PAC2E in Figure 4. We found that these PB-VI methods consistently outperform the VI. These numerical results support the theoretical findings that Ent-VI has an implicit regularization effect for over-parameterized models.

5.3 Application to variational autoencoder

We applied the bounds developed in Section 4.6 to VAEs and described the detailed network architecture in Appendix N. We used MNIST and the CelebA datasets (Liu et al., 2015). For MNIST, we evaluated the test log-likelihood. For the CelebA dataset, we evaluated the FID score (Heusel et al., 2017) between the real data and the randomly generated data from the models. A smaller FID score means that the distribution of the generated images is closer to the data. The experimental settings, including the network architecture and hyperparameters, are the same as in Shi et al. (2017).

Table 1: MNIST VAE results

Method	IWAE-VAR	IWAE	VI-VAR	VI
Test LL	-88.8	-89.0	-89.7	-89.9

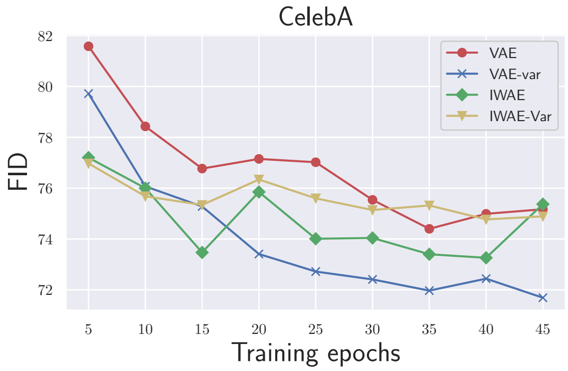

Figure 5: FID score results of CelebA dataset

We trained the VAEs using VI-VAR, IWAE, and a standard VI with $J = 10$, and IWAE-VAR with $J = 2, M = 5$. The MNIST results are shown in Table 1, which shows that a tighter bound shows a better test log-likelihood. The CelebA results are shown in Figure 5 and the VI-VAR showed the best performance. From these experiments, incorporating the risk-seeking property to learn latent variable models seems promising. Perhaps the worse performance of IWAE-VAR was caused by the larger variance of the gradient estimator.

6 CONCLUSION

We analyzed predictive Bayes under model misspecified and over-parameterized settings from the viewpoint of risk-seeking optimization. We provided the consistency and a MSE performance guarantee for PB, both of which are useful when studying parameter estimation. We also clarified that PB has an implicit regularization effect, which induces a better generalization performance. Our future works will replace the loss function in VI with a general risk-seeking objective function and study its theoretical and practical benefits.

Acknowledgements

FF was supported by JST ACT-X Grant Number JPMJAX190R.

References

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 21–30.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom).

Ben-Tal, A., Ben-Israel, A., and Teboulle, M. (1991). Certainty equivalents and information measures: duality and extremal principles. *Journal of Mathematical Analysis and Applications*, 157(1):211–236.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5):1103.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. (2019). Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018.

Chaudhari, P., Oberman, A., Osher, S., Soatto, S., and Carlier, G. (2018). Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):1–30.

Chérief-Abdellatif, B.-E. and Alquier, P. (2018). Consistency of variational bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12(2):2995–3035.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2015). Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in Neural Information Processing Systems*, 28:1522–1530.

Föllmer, H. and Knispel, T. (2011). Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations. *Stochastics and Dynamics*, 11(02n03):333–351.

Futami, F., Iwata, T., Sato, I., Sugiyama, M., et al. (2021). Loss function based second-order jensen inequality and its application to particle variational inference. *Advances in Neural Information Processing Systems*, 34.

Futami, F., Sato, I., and Sugiyama, M. (2018). Variational inference based on robust divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 813–822. PMLR.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). Pac-bayesian theory meets bayesian inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1884–1892.

Gotoh, J.-y., Kim, M. J., and Lim, A. E. (2018). Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4):448–452.

Grünwald, P. (2012). The safe bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer.

Grünwald, P. and Van Ommen, T. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Grünwald, P. D. and Mehta, N. A. (2020). Fast rates for general unbounded loss functions: From erm to generalized bayes. *J. Mach. Learn. Res.*, 21:56–1.

Heide, R., Kirichenko, A., Grunwald, P., and Mehta, N. (2020). Safe-bayesian generalized linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2623–2633. PMLR.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ingraham, J. and Marks, D. (2017). Variational inference for sparse and undirected models. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29:2946–2954.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, J., Park, S., and Shin, J. (2020). Learning bounds for risk-sensitive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13867–13879. Curran Associates, Inc.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The annals of statistics*, 22(2):1081–1114.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Masegosa, A. (2020). Learning under model misspecification: Applications to variational and ensemble methods. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5479–5491. Curran Associates, Inc.

Morningstar, W. R., Alemi, A. A., and Dillon, J. V. (2020). Pac m-bayes: Narrowing the empirical risk gap in the misspecified bayesian regime. *arXiv preprint arXiv:2010.09629*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective.*

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR.

Shi, J., Sun, S., and Zhu, J. (2017). Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119.*

Tsuzuku, Y., Sato, I., and Sugiyama, M. (2020). Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9636–9647. PMLR.

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., and Williamson, R. C. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, Y., Kucukelbir, A., and Blei, D. M. (2017). Robust probabilistic modeling with bayesian data reweighting. In *International Conference on Machine Learning*, pages 3646–3655. PMLR.

Wang, Z., Ren, T., Zhu, J., and Zhang, B. (2019). Function space particle optimization for bayesian neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory.* Number 25. Cambridge university press.

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the bayes posterior in deep neural networks really? In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR.

Xu, A. (2020). Continuity of generalized entropy and statistical learning. *arXiv preprint arXiv:2012.15829.*

Supplementary Material:
Predictive variational Bayesian inference as risk-seeking optimization

Symbolslist

Sign	Description
γ	A temperature parameter in an entropic risk
N	A number of training dataset
m	The number of samples for the multi-sample bound
$\nu(x)$	True data generating distribution
$x^N = (x_1, \ldots, x_N)$	Training data drawn from $\nu(x)$
$p(x\|\theta)$	A model, also called as a likelihood function
$l_n(\theta)$	A log loss, defined as $l_n(\theta) := -\ln p(x_n\|\theta)$
$l_N(\theta)$	The sum of the log loss, defined as $L_N(\theta) := -\sum_n l_n(\theta)$
$\pi(\theta)$	A prior distribution
$p(\theta\|x^N)$	Bayesian posterior distribution, $p(\theta\|x^N) \propto \exp(-L_N(\theta))\pi(\theta)$
$q(\theta;\phi)$	An approximate posterior distribution with a variational parameter ϕ
q_θ^{Ent}	A solution of Ent-VI

A DEFINITIONS OF DIVERGENCES

Here we introduce the definitions of divergences used in the paper. Consider probability distributions P and Q in some measurable space. We define KL divergence between P and Q

$$\mathrm{KL}(P|Q) := \int \ln \frac{dP}{dQ} dP, \tag{26}$$

if Q dominates P, otherwise $\mathrm{KL}(P|Q) = \infty$. We also define α-divergence between p and q as

$$D_\alpha(P|Q) := \frac{1}{\alpha - 1} \ln \int \left(\frac{dP}{dQ}\right)^{\alpha-1} dP, \tag{27}$$

if Q dominates P, otherwise $D_\alpha(P|Q) = \infty$. When $\alpha = 1/2$, D_α is closely related to the Hellinger divergence,

$$\mathrm{Hel}^2(P|Q) := \frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2. \tag{28}$$

B FURTHER PRELIMINARY OF EXISTING WORK

Here we introduce additional results for PB and risk seeking optimization.

B.1 Predivtive bayes

As explained in the main paper, Masegosa (2020), Futami et al. (2021), and Morningstar et al. (2020) directly minimized the KL divergence between the approximate predictive distribution and data generating distribution:

$$\mathrm{KL}(\nu(x)|\mathbb{E}_{q(\theta;\phi)} p(x|\theta)) := \mathbb{E}_{\nu(x)}[-\ln \mathbb{E}_{q(\theta;\phi)} p(x|\theta)] + \mathrm{Const} \tag{29}$$

Table 3: Summary of loss functions

Name	**Loss**	g(t)	dual divergence
VI	$\mathbb{E}_q l_n$	t	–
Ent-VI	$-\ln \mathbb{E}_q e^{-\gamma l_n}$	$\frac{1}{\gamma}e^{\gamma t} - \frac{1}{\gamma}$	KL
MV-VI	$\mathbb{E}_q l_n - \mathrm{Var}_q[l_n]$	$t + ct^2$	χ^2

and motivated by the PAC Bayesian theory (Germain et al., 2016), they added regularization term KL$(q|\pi)$. The final objective function is

$$\mathrm{Obj}_{\mathrm{PB}}(\phi) := \frac{1}{N}\sum_{n=1}^{N}[-\ln \mathbb{E}_{q(\theta;\phi)}.p(x_n|\theta)] + \frac{\mathrm{KL}(q|\pi)}{\alpha N} \tag{30}$$

The exact integral is intractable when optimizing this objective function. Therefore, existing work considered different approximations. As shown in the main paper, Morningstar et al. (2020) considered the multi-sample bound like IWAE Eq.(5). On the other hand, Masegosa (2020) and Futami et al. (2021) applied the second-order Jensen inequality. For simplicity, we only cite the result of Futami et al. (2021). They proved that

$$\mathbb{E}_{q(\theta)} \ln p(x|\theta) \leq \ln \mathbb{E}_{q(\theta)} p(x|\theta) - \underbrace{\mathbb{E}_{q(\theta)} \left(\frac{\ln p(x|\theta) - \mathbb{E}_{q(\theta)} \ln p(x|\theta)}{2h(x,\theta)} \right)^2}_{:= \mathrm{R}(x,h)}, \tag{31}$$

where

$$h(x,\theta)^{-2} = \exp\left(\ln p(x|\theta) + \mathbb{E}_{q(\theta)} \ln p(x|\theta) - 2\max_{\theta} \ln p(x|\theta) \right). \tag{32}$$

And then they proposed the objective function

$$\mathrm{Obj}_{\mathrm{WMV}}(\phi) := \frac{1}{N}\mathbb{E}_q L_N(\theta) - \mathbb{E}_q \frac{1}{N}\sum_{n=1}^{N} R(x_n, h_m) + \frac{\mathrm{KL}(q|\pi)}{\alpha N}, \tag{33}$$

where the second term is the weighted variance. Thus, this is the weighted mean-variance objective function.

B.2 Risk seeking optimization

First, we introduce the general risk-seeking objective function as follows. We define inverted optimized certainty equivalents (I-OCEs)(Lee et al., 2020) as

$$\overline{\mathrm{OCE}}^g(q(\theta;\phi), l_n(\theta)) := \sup_{\lambda \in \mathbb{R}} \left\{ \lambda - \mathbb{E}_{q(\theta;\phi)}[g(\lambda - l_n(\theta))] \right\}, \tag{34}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a non-decreasing convex function that satisfies $g(0) = 0$ and $1 \in \partial g(0)$. This g express a distortion of the original loss function l_n. Important point is that many risk-seeking distorted functions can be derived by choosing appropriate g function. Examples of g are shown in Table 3. For example, by setting $g(t) := \frac{1}{\gamma}e^{\gamma t} - \frac{1}{\gamma}$ where $\gamma \in \mathbb{R}^+$, we get the entropic-risk.

I-OCE risk is called risk-seeking objective function because it tries to find the solution which is more dispersed concerning the distribution used for the expectation in Eq.(34).

As written in the main paper, some OCE risk has dual forms and are discussed in various works (Ben-Tal et al., 1991). For example, the duality of entropic risk is discussed in Föllmer and Knispel (2011). Relation to MV-loss and χ square divergence is discussed in Gotoh et al. (2018):

$$\mathbb{E}_{q(\theta)} l_n(\theta) - \frac{\gamma}{2}\mathrm{Var}_q(l_n) = \inf_{q'} \mathbb{E}_{q'(\theta)}[l_n(\theta)] + \frac{1}{\gamma}\chi^2(q'(\theta)|q(\theta;\phi)), \tag{35}$$

where $\chi^2(q'|q) := \int(\frac{q'}{p}-1)^2 p d\theta$ is the χ square divergence.

As written in the main paper, MV-loss plays an important role in the I-OCE risk framework. Lee et al. (2020) proved that I-OCE risk is upper and lower bounded by the risk functions as

$$l_n - \frac{\mathrm{Lip}_\phi}{2}\sqrt{\mathrm{Var}_q(l_n)} \le \overline{\mathrm{OCE}}^\phi(q, l_n) \le l_n - C_\phi \mathrm{Var}_q(l_n) \tag{36}$$

where Lip_ϕ is the lipscitz constant of ϕ and $C_\phi := \inf_{0<|t|}\frac{\phi(t)-t}{t}$. Thus, studying the I-OCE risk is closely related to studying MV-risk in a sense.

In the case of entropic risk, if a log-loss is bounded, then there exists a constant c such that

$$\frac{1}{N}\sum_{n=1}^{N}[-\frac{1}{\lambda}\ln \mathbb{E}_{q(\theta;\phi)}e^{-\lambda l_n}] \le \frac{1}{N}\sum_{n=1}^{N} l_n - c\mathrm{Var}_q(l_n). \tag{37}$$

Thus, the risk functional of Ent-VB is upper bounded by MV-VI.

C RISK SEEKING OBJECTIVE IN VI

As we introduced in Section 1 and 3, PB uses the distorted loss functions. When focusing on loss function, VI optimizes $\sum_n \mathbb{E}_q l_n(\theta)$, on the other hand. PB-VI optimizes $\sum_n -\ln \mathbb{E}_q e^{-l_n(\theta)}$. This observation indicates that the original loss l_n in VI is exponentially distorted in PB-VI. The idea of distorting the objective function is known as risk-seeking objective (Lee et al., 2020). Motivated by this connection, we can consider a general risk-seeking VI, of which loss functional is replaced with the risk-seeking objective as follows.

$$\mathrm{Obj}_{\mathrm{RS}}(\phi) := \frac{1}{N}\sum_{n=1}^{N}\overline{\mathrm{OCE}}^g(q, l_n(\theta)) + \frac{\mathrm{KL}(q|\pi)}{\alpha N}. \tag{38}$$

We refer to the VI, which uses Eq.(38) as risk-seeking VI (RS-VI). Significantly, by choosing the appropriate g functions, we can recover VI, PB-VI, and MV-VI as shown in Table 3. For example, by setting $\phi(t) := \frac{1}{\gamma}e^{\gamma t} - \frac{1}{\gamma}$ where $\gamma \in \mathbb{R}^+$, get Ent-VI. Also, when we use MV-loss, we get the MV-VI. The intuition of the problem in Eq.(38) is that we enhance the variation of loss function similar to Eq.(6).

We leave it a future work to analyze the theoretical properties of RS-VI and numerical performances.

D Discussion about MSE

Here, we discuss the behavior of MSE when using the standard VI and the PB-VI. First, as we introduced in Section 3.2, the objective of PB-VI is given as

$$\begin{aligned}&\mathbb{E}_\nu \mathrm{KL}(\nu(y|x)|\mathbb{E}_q p(y|x,\theta))\\&= \frac{\mathbb{E}_\nu\|R(x)-\mathbb{E}_q[r_\theta(x)]\|^2}{2\sigma^2} + \frac{1}{2}\ln\frac{\sigma^2}{\rho^2} + \mathbb{E}_\nu\frac{\sigma^2-\rho^2}{2\sigma^4}(\mathrm{Var}_q[r_\theta(x)]-1) - \frac{1}{2\sigma^4}\mathrm{Var}_q[r_\theta(x)I(x) + \|r_{\theta^*}-r_\theta\|^2]\end{aligned} \tag{39}$$

where $I(x) := (R(x) - r_{\theta^*}(x))^2$. When the model is misspecified, e.g., $R(x) \neq r_{\theta^*}(x)$, we can make the objective of PB-VI small by increasing $\mathrm{Var}_q r_\theta(x)$. Even when $R(x) \neq r_{\theta^*}(x)$, we can make the objective of PB-VI small by increasing $\mathrm{Var}_q[\|r_{\theta^*} - r_\theta\|^2]$. This indicates that we can make the objective of PB-VI, $\mathbb{E}_\nu \mathrm{KL}(\nu(y|x)|\mathbb{E}_q p(y|x,\theta))$, small, while the MSE $\frac{\mathbb{E}_\nu\|R(x)-\mathbb{E}_q[r_\theta(x)]\|^2}{2\sigma^2}$ is large by manipulating the dispersity of the function r_θ.

Next, we consider the objective of the standard VI, which is given as

$$\mathbb{E}_\nu\mathbb{E}_{q(\theta)}\mathrm{KL}(\nu(y|x)|p(y|x,\theta)) = \frac{\mathbb{E}_\nu\mathbb{E}_{q(\theta)}\|R(x) - r_\theta(x)\|^2}{2\sigma^2} + \frac{1}{2}\ln\frac{\sigma^2}{\rho^2}. \tag{40}$$

Note that when fixing $q(\theta)$, the optimal σ^2 is

$$\sigma^2 = \mathbb{E}_\nu\mathbb{E}_{q(\theta)}\|R(x) - r_\theta(x)\|^2, \tag{41}$$

which is pointed out in Grünwald and Van Ommen (2017).

On the other hand, the squared loss for prediction is given as $\mathbb{E}_\nu\|R(x)-\mathbb{E}_{q(\theta)}r_\theta(x)\|^2$, which is upper-bounded by

$$\mathbb{E}_\nu\|R(x)-\mathbb{E}_{q(\theta)}r_\theta(x)\|^2 \leq \mathbb{E}_\nu\mathbb{E}_{q(\theta)}\|R(x)-r_\theta(x)\|^2. \tag{42}$$

Thus, compared to PB-VI, the standard VI directly control the squared loss for prediction.

E PROOF OF THEOREM 1 (CONSISTENCY)

Proof. Following Alquier and Ridgway (2020), from the definition of the α-divergence, we obtain

$$\mathbb{E}_{\nu(X)^{\otimes N}}e^{-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)}=1, \tag{43}$$

where

$$r_N(p_\theta,\nu):=\sum_{n=1}^N \ln\frac{\nu(X_n)}{p(X_n|\theta)} \tag{44}$$

and expectation is taken with respect to the draw of the training dataset.

Next, we take the expectation concerning a prior π and, using Fubini's theorem, and we swap the expectation. We obtain

$$\mathbb{E}\left[\int e^{-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)}d\pi(\theta)\right]=1. \tag{45}$$

Then by using the measure change formula in Lemma 2.2 in Alquier and Ridgway (2020), we obtain

$$\mathbb{E}\left[\exp\left\{\sup_\rho\left\{\int\Big(-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)\Big)d\rho(\theta)-\mathrm{KL}(\rho|\pi)\right\}\right\}\right]=1, \tag{46}$$

where the supremum is taken over all the probability distributions on the given measurable space. For completeness, we show the Lemma 2.2 in Alquier and Ridgway (2020)

Lemma 1 (Lemma2.2 in Alquier and Ridgway (2020))**.** *Given a measurable space, for any probability π and any measurable function h that takes real values such that $\int e^h d\pi<\infty$, we have*

$$\ln\int e^h d\pi=\sup_\rho\left[\int hd\rho-KL(\rho|\pi)\right]. \tag{47}$$

We take the log function on both hand-side in Eq.(46), and by applying the Jensen inequality. Then we have

$$\mathbb{E}\left[\sup_\rho\left\{\int\Big(-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)\Big)d\rho(\theta)-\mathrm{KL}(\rho|\pi)\right\}\right]\leq 0. \tag{48}$$

Recall the dual form of the Ent-risk

$$-\frac{1}{\gamma}\ln\mathbb{E}_q e^{-\gamma l_n}=\inf_{q'}\{\mathbb{E}_{q'}[l_n]+\frac{1}{\gamma}\mathrm{KL}(q'(\theta)|q(\theta;\phi))\}, \tag{49}$$

and we express the solution of this q' as q^*. Then, we substitute $\rho=q^*$ in Eq.(48), we obtain

$$\mathbb{E}\left[\int\Big(-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)\Big)q^*d\theta-\mathrm{KL}(q^*|\pi)\right]\leq 0. \tag{50}$$

Then by rearranging the above inequality, we have

$$\mathbb{E}_\nu\int D_\alpha(p_\theta,\nu)q^*d\theta\leq\mathbb{E}_\nu\left[\frac{\alpha}{N(1-\alpha)}\int r_N(p_\theta,\nu)q^*d\theta+\frac{\mathrm{KL}(q^*|\pi)}{N(1-\alpha)}\right]. \tag{51}$$

Next, we lower bound the left-hand side of Eq.(107) by the expectation of $q^{\mathrm{Ent}}(\theta)$. We use the following relation

$$\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \leq \int D_{1/2}(p_\theta,\nu)q^* d\theta + 2\mathrm{KL}(q^*|q^{\mathrm{Ent}}) \leq \frac{1-\alpha}{\alpha}\int D_\alpha(p_\theta,\nu)q^* d\theta + 2\mathrm{KL}(q^*|q^{\mathrm{Ent}}). \tag{52}$$

This can be derived using

$$D_{1/2}(p_\theta,\nu) \leq \frac{1-\alpha}{\alpha}D_\alpha(p_\theta,\nu) \tag{53}$$

and

$$\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \leq -2\ln\mathbb{E}_{q^{\mathrm{Ent}}(\theta)}e^{-1/2D_{1/2}(p_\theta,\nu)} \leq \mathbb{E}_{q^*}[D_{1/2}(p_\theta,\nu)]+2\mathrm{KL}(q^*|_\theta^{\mathrm{Ent}}). \tag{54}$$

Thus, we have

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \leq \mathbb{E}\left[\frac{1}{N}\int\sum_i^N \ln\frac{\nu(x_i)}{p_\theta(x_i)}q^* d\theta + \frac{2}{N}\sum_i^N \mathrm{KL}(q^*|q^{\mathrm{Ent}}) + \Omega(q^{\mathrm{Ent}}(\theta),q^*) + \frac{\mathrm{KL}(q^{\mathrm{Ent}}(\theta)|\pi)}{N\alpha}\right], \tag{55}$$

where

$$\Omega(q^{\mathrm{Ent}}(\theta),q^*) := \frac{1}{N\alpha}\int \ln\frac{q^*}{\pi}q^* d\theta - \frac{1}{N\alpha}\int \ln\frac{q^{\mathrm{Ent}}(\theta)}{\pi}q^{\mathrm{Ent}}(\theta)d\theta. \tag{56}$$

We upper bound this term as follows. From Theorem 2 in Xu (2020), when the exponential integral condition $\mathrm{E}_{q^{\mathrm{Ent}}(\theta)}e^{\gamma\ln\pi} \leq e^{\gamma^2\tilde{\sigma}}$ is satisfied for $0<\gamma<\gamma_0$, we have

$$\mathbb{E}_{q^*}\ln\pi - \mathrm{E}_{q^{\mathrm{Ent}}(\theta)}\ln\pi \leq \sqrt{2\tilde{\sigma}^2\mathrm{KL}(q^*|q^{\mathrm{Ent}})}. \tag{57}$$

Using the above relation, we have

$$\begin{aligned}\Omega(q^{\mathrm{Ent}}(\theta),q^*) &= \frac{1}{N\alpha}\int \ln\frac{q^*}{\pi}q^* d\theta - \frac{1}{N\alpha}\int \ln\frac{q^{\mathrm{Ent}}(\theta)}{\pi}q^{\mathrm{Ent}}(\theta)d\theta \\ &\leq -\frac{1}{N\alpha}\int \ln\pi q^* d\theta + \frac{1}{N\alpha}\int \ln\pi q^{\mathrm{Ent}}(\theta)d\theta + \frac{1}{N\alpha}H_0 \\ &\leq \frac{1}{N\alpha}\sqrt{2\tilde{\sigma}^2\mathrm{KL}(q^*|q^{\mathrm{Ent}})} + \frac{1}{N\alpha}H_0 \\ &\leq \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{\alpha}{2N}\mathrm{KL}(q^*|q^{\mathrm{Ent}}) + \frac{1}{N\alpha}H_0. \end{aligned} \tag{58}$$

where we also used the assumption that the entropy of variational posterior $q^{\mathrm{Ent}}(\theta)$ is upper bounded by H_0, $\mathrm{E}_{q^{\mathrm{Ent}}(\theta)}[-\ln q^{\mathrm{Ent}}(\theta)] \leq H_0$. Then we have

$$\begin{aligned}&\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \\ &\leq \mathbb{E}\left[\frac{1}{N}\sum_i^N\left(\int \ln\frac{\nu(x_i)}{p_\theta(x_i)}q^* d\theta + (2+\alpha/(2N))\mathrm{KL}(q^*|q^{\mathrm{Ent}})\right) + \frac{\mathrm{KL}(q^{\mathrm{Ent}}(\theta)|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right]. \end{aligned} \tag{59}$$

Then by using the definition of the dual form of Ent-risk,

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \leq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N -\frac{1}{\gamma}\ln\mathbb{E}_{q^{\mathrm{Ent}}(\theta)}e^{-\gamma l_i} + \frac{\mathrm{KL}(q^{\mathrm{Ent}}(\theta)|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right], \tag{60}$$

where $\gamma = 2N/(4N+\alpha)$ Then, by definition $q^{\mathrm{Ent}}(\theta)$ is the solution of the problem, we can write above as

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \leq \mathbb{E}\inf_\rho\left[\frac{1}{N}\sum_i^N -\frac{1}{\gamma}\ln\mathbb{E}_{q^{\mathrm{Ent}}(\theta)}e^{-\gamma l_i} + \frac{\mathrm{KL}(\rho|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right]. \tag{61}$$

Then, applying the dual form of Ent-risk again, we obtain We substitute $\rho = \phi_n$

$$\begin{aligned}&\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta\\ &\leq \mathbb{E}\inf_{\rho,\rho'}\left[\frac{1}{N}\sum_i^N\left(\int \ln\frac{\nu(x_i)}{p_\theta(x_i)}\rho' d\theta+(2+\alpha/(2N))\mathrm{KL}(\rho'|\rho)\right)+\frac{\mathrm{KL}(\rho|\pi)}{N\alpha}+\frac{\tilde{\sigma}^2}{N\alpha^2}+\frac{1}{N\alpha}H_0\right].\end{aligned} \tag{62}$$

Then from the assumption, we assumed that $q(\theta;\phi_N)$ such that

$$\int \mathrm{KL}(\nu(x)|p(x|\theta))q(\theta;\phi_N)d\theta \leq \epsilon_N, \tag{63}$$

$$\mathrm{KL}(q(\theta;\phi_N)|\pi) \leq N\epsilon_N. \tag{64}$$

So, we substitute this to $\rho' = q(\theta;\phi_n)$ and $\rho' = \rho$, we obtain

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta)d\theta \leq \left[\frac{1+\alpha}{\alpha}\epsilon_N+\frac{\tilde{\sigma}^2}{N\alpha^2}+\frac{1}{N\alpha}H_0\right]. \tag{65}$$

So far we focused on model is well specified setting. As for the model misspecification, we just replace the log loss in Eq.(107) by

$$\mathrm{KL}(\nu(x)|p(x|\theta))=\mathrm{KL}(\nu(x)|p(x|\theta^*))+\mathbb{E}\ln\frac{p(X_n|\theta^*)}{p(X_n|\theta)}. \tag{66}$$

Then, the proof goes almost in the same way, and we add model misspecified term $\mathrm{KL}(\nu(x)|p(x|\theta^*))$. □

F DISCUSSION ABOUT THE ASSUMPTIONS

Recall the assumptions

$$H[q^{\mathrm{Ent}}(\theta)] := -\int \ln q^{\mathrm{Ent}}(\theta)q^{\mathrm{Ent}}(\theta)d\theta \leq H_0, \tag{67}$$

and there exists a γ_0 and $\tilde{\sigma}^2$ such that for all $0 < \gamma < \gamma_0$

$$\mathrm{E}_{q^{\mathrm{Ent}}(\theta)}e^{\gamma\ln\pi} \leq e^{\gamma^2\tilde{\sigma}}. \tag{68}$$

Assume that an approximate posterior distribution is a Gaussian distribution. Then the entropy condition is equivalent to the condition that the log determinant of the covariance matrix of the approximate posterior distribution is finite. This condition will not be satisfied only when the covariance matrix has 0 or ∞ as an eigenvalue. Thus, the finite entropy condition is satisfied when we restrict the domain of the variational parameter ϕ.

The exponential integrable condition with respect to $\ln\pi$ is satisfied because Gaussian distribution is subGaussian distribution. Note that the exponential condition is closely related to subExponential condition of $q^{\mathrm{Ent}}(\theta)$.

G PROOF OF COLLORARY 1 (Excess risk bound)

G.1 The outline of the proof

To prove Collorary 1, we introduce two assumptions used in the previous work in Grünwald and Mehta (2020); Heide et al. (2020). Here, we express $l_\theta := -\ln p(x|\theta)$ and $l_{\theta^*} := -\ln p(x|\theta^*)$ and $L_\theta = l_\theta - l_{\theta^*}$. Here θ^* indicates the parameter that minimizes the KL divergence between $\nu(x)$ and $p(x|\theta)$. We introduce several definitions to use the results of Grünwald and Mehta (2020); Heide et al. (2020).

Definition 1. *Given η, we say that $(\nu(x), L_\theta)$ satisfies the η-strong central condition if*

$$\mathbb{E}e^{-\eta L_\theta} \leq 1 \tag{69}$$

is satisfied.

The model that satisfies Assumption 1 is, for example, the generalized linear model, in which the linear regression model is included (Grünwald and Mehta, 2020; Heide et al., 2020). Another assumption used in Grünwald and Mehta (2020); Heide et al. (2020) is the witness condition

Definition 2. *We say that (u, c) witness condition holds for a constant $u > 0$ and $c \in (0, 1]$ if*

$$\mathbb{E}_\nu[L_\theta \cdot \mathbf{1}_{L_\theta \leq u}] \geq c\mathbb{E}_\nu L_\theta \tag{70}$$

holds. We say that for a function $\tau : \mathbb{R}^+ \to [1, \infty)$ and constant $c \in (0, 1]$, we say that (τ, c) witness condition holds if $\mathbb{E}L_\theta < \infty$ and

$$\mathbb{E}_\nu[L_\theta \cdot \mathbf{1}_{L_\theta \leq \tau(\mathbb{E}_\nu L_\theta)}] \geq c\mathbb{E}_\nu L_\theta. \tag{71}$$

Then, Lemma 13 in Grünwald and Mehta (2020) show that

Lemma 2 (Lemma 13 in Grünwald and Mehta (2020))**.** *when η-strong central condition in Assumption 1 and (τ, c) witness condition in Definition 2 are satisfied, then, for all $\lambda > 0$ we have,*

$$\mathbb{E}_\nu L_\theta \leq \lambda \vee \left(c_{\tau(\lambda)} \frac{1}{\eta'} \left(1 - \mathbb{E}_\nu e^{-\eta' L_\theta} \right) \right) \tag{72}$$

where

$$c_{\tau(\lambda)} := \frac{1}{c} \frac{\eta'\tau(\lambda) + 1}{1 - \frac{\eta'}{\eta}}. \tag{73}$$

and η' is a arbitrary constant such that $0 < \eta' < \eta$.

Our strategy is that:

1. we prove that PB-VI under our assumptions satisfies η-strong and (τ, c) witness conditions
2. We upper bound $\frac{1}{\eta'}\left(1 - \mathbb{E}_\nu e^{-\eta' L_\theta}\right)$ in Eq.(72) of PB-VI.
3. Combine i) and ii) and using Eq.(72), we upper bound the RMSE.

G.2 Proof

First, we consider step 1) in the above. We check the η-strong and (τ, c) witness conditions. We use the following lemma:

Lemma 3 (Lemma 16 in Grünwald and Mehta (2020))**.** *If there exists a κ that satisfies*

$$\sup_\theta \mathbb{E}_\nu e^{\kappa L_\theta} < \infty, \tag{74}$$

then the (τ, c) witness condition holds under $c = 1/2$ and

$$\tau(x) = 1 \vee \kappa^{-1} \ln \frac{2M_\kappa}{\kappa x}, \tag{75}$$

where

$$M_\kappa = \sup_\theta \mathbb{E}_\nu e^{\kappa L_\theta} < \infty. \tag{76}$$

To analyze the condition of Eq.(74), we use Proposition 1 in Heide et al. (2020). They studied that under what condition the generalized linear model (GLM) satisfies the condition of Eq.(74).

To state that condition, we introduce the definitions of a GLM:

$$p(y|x, \theta) := \exp\left(x^\top \theta y - F(\theta) + r(y)\right). \tag{77}$$

Here, given $x \in \mathcal{X} \subset \mathbb{R}^d$ and the mean value parameter is given by $g^{-1}(x^\top \theta)$ where g is the link function. F is the normalizing constant and r is the reference measure. With this setting, $\mathbb{E}_{p(y|x,\theta)}[y|x,\theta] = g^{-1}(x^\top \theta)$.

Under this definition, Proposition 1 and lemma 1 in Heide et al. (2020) states that i) Θ and $\mathcal{X}$ is restricted on a compact domain, ii) for some η $\sup_{x\in\mathcal{X}} \mathbb{E}_{\nu(Y|X)}[e^{\eta|Y|}|X = x] < \infty$, and iii) there exists a true mean parameter $\mathbb{E}_{\nu(Y|X)}[Y|X] = g^{-1}(x^\top \theta^*)$, there exists $\eta > 0$ that only depending on the parameters of the problem and satisfies η-strong central condition. Moreover the condition of Eq.(74) will be satisfied in GLM, which means (u, c) witness condition is satisfied.

Since we assumed these conditions, η-strong and (τ, c) witness conditions are satisfied for our settings.

Next, we consider the step 2), that is derive the upper bound $\frac{1}{\eta'}\left(1 - \mathbb{E}_\nu e^{-\eta' L_\theta}\right)$. We use the following relation: Using Proposition 1 in Grünwald and Mehta (2020), for all $0 < \eta' < 1/2$ we have

$$\frac{1}{\eta'}\left(1 - \mathbb{E}_\nu e^{-\eta' L_\theta}\right) \leq 2\left(1 - \mathbb{E}_\nu e^{-\frac{1}{2}L_\theta}\right). \tag{78}$$

Thus, we define the right hand side as

$$\mathrm{Hel}^{2'}(p_\theta, p_\theta^*) := 2\left(1 - \mathbb{E}_\nu e^{-\frac{1}{2}\ln\frac{p(x|\theta^*)}{p(x|\theta)}}\right), \tag{79}$$

We need to bound this Hellinger like metric function.

For that purpose, we need to slightly change the result of in Theorem 1. Recall that Theorem 1 states that

$$\mathbb{E}_{\nu(x)} \int \mathrm{Hel}^2(p_\theta, \nu) q^{\mathrm{Ent}}(\theta) d\theta \leq \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right]. \tag{80}$$

So when we want to use Lemma 13 in Grünwald and Mehta (2020), we need to change the Hellinger divergence between ν and $p(x|\theta)$ to $p(x|\theta^*)$ and $p(x|\theta)$ given as $\mathrm{Hel}^{2'}(p_\theta, p_\theta^*)$.

Corollary 2. *Under the assumption as theorem 1, we have*

$$\mathbb{E}_{\nu(x)} \int \mathrm{Hel}^{2'}(p_\theta, p_\theta^*) q^{\mathrm{Ent}}(\theta) d\theta \leq \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right] \tag{81}$$

Proof. We proceed the proof in the same way as the Theorem 1 except for the beginning.

At the beginning of the proof of Theorem 1, we use the following results: From the definition of the Hellinger divergence, setting $\alpha = 1/2$, we have the relation,

$$\mathbb{E}_{\nu(X)^{\otimes N}} e^{-\alpha r_N(p_\theta, p_\theta^*) + (1-\alpha)ND'_\alpha(p_\theta, p_\theta^*)} = 1, \tag{82}$$

where

$$r_N(p_\theta, p_\theta^*) := \sum_{n=1}^{N} \ln\frac{p(X_n|\theta^*)}{p(X_n|\theta)}, \tag{83}$$

and we define

$$D'_\alpha(p_\theta, p_\theta^*) := \frac{1}{\alpha - 1}\ln\int\left(\frac{p(x|\theta^*)}{p(x|\theta)}\right)^{\alpha-1}\nu(x)dx. \tag{84}$$

With these new notation, the proof is same as the Theorem 1. □

Thus, for all $0 < \eta' < 1/2$ we have

$$\frac{1}{\eta'}\left(1 - \mathbb{E}_\nu e^{-\eta' L_\theta}\right) \leq \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right]. \tag{85}$$

From Lemma 13 in Grünwald and Mehta (2020) for all λ and for all $0 < \eta' < \eta \vee 1/2$we have,

$$\mathbb{E}_\nu L_\theta \leq \lambda \vee \left(c_{\tau(\lambda)} \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0 \right] \right) \tag{86}$$

where

$$c_{\tau(\lambda)} := \frac{1}{c}\frac{\eta'\tau(\lambda)+1}{1-\frac{\eta'}{\eta}}. \tag{87}$$

with $c = 1/2$.

Since λ can take an arbitrary positive value, we set $\lambda = 1/N$. Then we have

$$c_{\tau(1/N)} := \frac{1}{c}\frac{\eta\tau(1/N)+1}{1-\frac{1-\alpha}{\eta}}. \tag{88}$$

and

$$\tau(1/N) = 1 \vee \kappa^{-1} \ln \frac{2M_\kappa N}{\kappa}. \tag{89}$$

Thus, in conclusion, we have

$$\mathbb{E}_{\nu(x)} \int \mathbb{E}_\nu L_\theta q^{\mathrm{Ent}}(\theta) d\theta \leq c_{\tau(1/N)} \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0 \right]. \tag{90}$$

From the definition of $c_{\tau(1/N)}$, there exists a constant c_1 and c_2, s.t.

$$c_{\tau(1/N)} = c_1 + c_2 \ln N \tag{91}$$

and c_1 and c_2 only depend on the parameters of the problem. Thus, we have

$$\mathbb{E}_{\nu(x)} \int \mathbb{E}_\nu L_\theta q^{\mathrm{Ent}}(\theta) d\theta \leq (c_1 + c_2 \ln N) \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0 \right]. \tag{92}$$

Then if we assume that the likelihood is the Gaussian distribution, we proved the corollary.

Remark 1. *Note that Eq.*(92) *can be expressed as*

$$\begin{aligned} &\mathbb{E}_{\nu(y,x)} \int \mathbb{E}_{\nu(x)} \mathrm{KL}(\nu(y|x)|p(y|x,\theta^*)) q_\theta^{\mathrm{Ent}} d\theta \\ &\leq \mathbb{E}_{\nu(x)} \mathrm{KL}(\nu(y|x)|p(y|x,\theta^*)) + (c_1 + c_2 \ln N) \left(\frac{1+\alpha}{\alpha}\epsilon_N + \frac{2\tilde{\tilde{\sigma}}^2 + H_0^2}{N\alpha^2} \right) \end{aligned} \tag{93}$$

which is known as the excess risk bound. In Theorem 1, the metric is the Hellinger divergence. On the other hand, above bound uses KL divergence, which is stronger than the Hellinger divergence. By controlling KL divergence, we can bound other useful metrics, such as Wasserstein distances.

Remark 2. *Here we discuss the condition ii) of Lemma 1 and Proposition 1 in Heide et al. (2020), that is, ii) for some η $\sup_{x\in\mathcal{X}} \mathbb{E}_{\nu(Y|X)}[e^{\eta|Y|}|X=x] < \infty$. For example, when the true data generating distribution is Gaussian distribution used in Section 3 in the main paper, this holds since*

$$\begin{aligned} \mathbb{E}_{\nu(Y|X)}[e^{\eta|Y|}|X=x] &= \mathbb{E}_{\nu(Y|X)}[e^{\eta|Y|} \cdot \mathbf{1}_{|Y|\leq 1}|X=x] + \mathbb{E}_{\nu(Y|X)}[e^{\eta|Y|} \cdot \mathbf{1}_{|Y|>1}|X=x] \\ &\leq e^\eta + \mathbb{E}_{\nu(Y|X)}[e^{\eta|Y|^2} \cdot \mathbf{1}_{|Y|>1}|X=x]. \end{aligned} \tag{94}$$

Since the distribution of Y^2 is sub-exponential distribution (Wainwright, 2019), from the definition of sub-exponential distribution, there exists a η such that $\mathbb{E}[e^{\eta|Y|^2}] < \infty$. Thus the condition is satisfied.

H FURTHER DISCUSSION ABOUT THE RELATED WORK

Here we discuss the relationship between our theoretical results with existing works in more detail.

H.1 Comparison with tempered posterior

In the analysis of tempered posterior distributions Alquier and Ridgway (2020), we only need Assumption 1 to get the consistency result. On the other hand, ours requires Assumption 2 in addition to Assumption 1 to control the dual problem of Ent-VI. Replacing Assumption 2 with weaker ones should be addressed in the future work.

As for the convergence speed Alquier and Ridgway (2020) showed that

$$\mathbb{E}_{\nu(x)}\int D^{\alpha}(\nu(x),p(x|\theta))q(\theta;\phi*_{\mathrm{VI}})d\theta \leq \frac{\alpha}{1-\alpha}\mathrm{KL}(\nu(x)|p(x|\theta^*))+\frac{1+\alpha}{1-\alpha}\epsilon_N, \tag{95}$$

where D^{α} is the α-divergence. Compared to our result, the difference is twofold. First, ours has additional term $\frac{2\tilde{\sigma}^2+H_0^2}{N\alpha^2}$ in the right-hand side. This is the term to control the dual problem. If $\epsilon=\mathcal{O}(1/N)$, ours show the same order as the result of Alquier and Ridgway (2020). The second difference is that Alquier and Ridgway (2020) showed the consistency in α-divergence for $\alpha<1$, on the other hand, we showed the consistency in the Hellinger distance, which is the special case of the α-divergence when $\alpha=1/2$. It is widely known that for any distribution P and Q, $D^{1/2}(p,Q)\leq\frac{1-\alpha}{\alpha}D^{\alpha}(P,Q)$, we can transform the result of Alquier and Ridgway (2020) into a Hellinger distance.

Grünwald and Mehta (2020); Heide et al. (2020) also worked on the tempered posterior distributions, and they focused on the exact posterior distribution. On the other hand, ours and Alquier and Ridgway (2020) focused on the variational posterior distributions. Alquier and Ridgway (2020) focused on the log-loss by considering the standard variational inference, and we focused on the distorted log-loss by considering the risk-seeking objective function. On the other hand, Grünwald and Mehta (2020) considered the more general risk function and derived the excess risk bound in various assumptions.

H.2 Some examples of ϵ_N

In the main paper, we introduced the ϵ_N for logistic regression model. The result holds for more general lipschitz log-loss models. From Alquier and Ridgway (2020), assume that q is MF Gaussian $q(\theta;\phi)=N(\theta|\mu,\sigma^2I_d)$ and $\pi(\theta)=N(\theta|0,\sigma_0^2I_d)$, where I_d is the d-dimensional identity matrix. Then we assume that for any $\theta,\theta'\in\times$ there exists a function $M(x)$ where $x\in\mathcal{X}$ that satisfies

$$|\ln p(x|\theta)-\ln p(x|\theta')|\leq M(x)\|\theta-\theta'\|_2, \tag{96}$$

and $\mathbb{M}(x)\leq L$. Then we have

$$\epsilon_N=\frac{L}{N}\vee\left\{\frac{d}{N}\left[\frac{1}{2}\ln(\sigma_0^2N^2d^{1/2})+\frac{1}{N\sigma_0^2}\right]+\frac{\|\theta^*\|}{N\sigma_0^2}-\frac{d}{2N}\right\}. \tag{97}$$

Other than this, Alquier and Ridgway (2020) provided examples of ϵ_N for a matrix completion problem and non-parametric regression problem. In Chérief-Abdellatif and Alquier (2018), the examples of ϵ_N for various mixture models are derived.

I THEORY FOR MV-VI

Here present the theory for MV-VI. Recall that the MV-VI is given as

$$\mathrm{Obj}_{\mathrm{MV}}(\phi):=\frac{1}{N}\left(\mathbb{E}_qL_N(\theta)-\gamma\sum_{n=1}^{N}\mathrm{Var}_q(l_n(\theta))+\beta\mathrm{KL}(q|\pi)\right), \tag{98}$$

and we denote the solution of $\arg\min_{\phi}\mathrm{Obj}_{\mathrm{MV}}(\phi)$ as ϕ_{MV} and $q_\theta^{\mathrm{MV}}:=q(\theta;\phi_{\mathrm{MV}})$. Then we have the following theorem,

Theorem 2. *For a given $\alpha\in(0,1)$ and $\gamma\in(0,(2+\alpha/2)^{-1}$, under assumption 1 and 2, q_θ^{MV} satisfies*

$$\mathbb{E}_{\nu(x)}\int\mathrm{Hel}^2(\nu(x),p(x|\theta))q_\theta^{\mathrm{MV}}d\theta\leq\mathrm{KL}(\nu(x)|p(x|\theta^*))+\frac{1+\alpha}{\alpha}\epsilon_N+\frac{2\tilde{\sigma}^2+H_0^2}{N\alpha^2}, \tag{99}$$

where Hel^2 is the Hellinger distance defined in Appendix A.

Proof. The proof goes almost same as the Ent-VI. From Eq. 59 and the dual form of MV risk T Following Alquier and Ridgway (2020), from the definition of the α-divergence, we obtain

$$\mathbb{E}_{\nu(X)^{\otimes N}} e^{-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)} = 1, \tag{100}$$

where

$$r_N(p_\theta,\nu) := \sum_{n=1}^{N} \ln \frac{\nu(X_n)}{p(X_n|\theta)} \tag{101}$$

and expectation is taken with respect to the draw of the training dataset.

Then we take the expectation with respect to a prior π and using Fubini's theorem we swap the expectation, we have

$$\mathbb{E}\left[\int e^{-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)} d\pi(\theta)\right] = 1. \tag{102}$$

Then by using the measure change formula in Lemma 2.2 in Alquier and Ridgway (2020), we obtain

$$\mathbb{E}\left[\exp\left\{\sup_\rho\left\{\int\Big(-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)\Big)d\rho(\theta)-\mathrm{KL}(\rho|\pi)\right\}\right\}\right]=1, \tag{103}$$

where the supremum is taken over all the probability distributions on the given measurable space. Then by applying the Jensen inequality, we have

$$\mathbb{E}\left[\sup_\rho\left\{\int\Big(-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)\Big)d\rho(\theta)-\mathrm{KL}(\rho|\pi)\right\}\right]\leq 0. \tag{104}$$

Recall the dual form of the MV-risk

$$\mathbb{E}_{q(\theta)}l_n(\theta)-\frac{\gamma}{2}\mathrm{Var}_q(l_n) = \inf_{q'}\mathbb{E}_{q'(\theta)}[l_n(\theta)] + \frac{1}{\gamma}\chi^2(q'(\theta)|q(\theta;\phi)), \tag{105}$$

and we express the solution of this q' as q^*. Then, we substitute $\rho = q^*$ in Eq.(104), we obtain

$$\mathbb{E}\left[\left\{\int\Big(-\alpha r_N(p_\theta,\nu)+(1-\alpha)ND_\alpha(p_\theta,\nu)\Big)q^*d\theta-\mathrm{KL}(q^*|\pi)\right\}\right]\leq 0. \tag{106}$$

Then by rearranging the above inequality, we have

$$\mathbb{E}_\nu\int D_\alpha(p_\theta,\nu)q^*d\theta \leq \mathbb{E}_\nu\left[\frac{\alpha}{N(1-\alpha)}\int r_N(p_\theta,\nu)q^*d\theta+\frac{\mathrm{KL}(q^*|\pi)}{N(1-\alpha)}\right]. \tag{107}$$

As for the left-hand side, we have

$$\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}}d\theta \leq \int D_{1/2}(p_\theta,\nu)q^*d\theta + 2\mathrm{KL}(q^*|q_\theta^{\mathrm{MV}}) \leq \frac{1-\alpha}{\alpha}\int D_\alpha(p_\theta,\nu)q^*d\theta+2\mathrm{KL}(q^*|q_\theta^{\mathrm{MV}}), \tag{108}$$

where we used the relation

$$D_{1/2}(p_\theta,\nu)\leq\frac{1-\alpha}{\alpha}D_\alpha(p_\theta,\nu) \tag{109}$$

and

$$\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}}d\theta \leq -2\ln\mathbb{E}_{q_\theta^{\mathrm{MV}}}e^{-1/2D_{1/2}(p_\theta,\nu)} \leq \mathbb{E}_{q^*}[D_{1/2}(p_\theta,\nu)] + 2\mathrm{KL}(q^*|_\theta^{\mathrm{MV}}). \tag{110}$$

Thus, we have

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}}d\theta \leq \mathbb{E}\left[\frac{1}{N}\int\sum_i^N \ln\frac{\nu(x_i)}{p_\theta(x_i)}q^*d\theta+\frac{2}{N}\sum_i^N\mathrm{KL}(q^*|q_\theta^{\mathrm{MV}})+\Omega(q_\theta^{\mathrm{MV}},q^*)+\frac{\mathrm{KL}(q_\theta^{\mathrm{MV}}|\pi)}{N\alpha}\right], \tag{111}$$

where

$$\Omega(q_\theta^{\mathrm{MV}}, q^*) := \frac{1}{N\alpha}\int \ln\frac{q^*}{\pi}q^* d\theta - \frac{1}{N\alpha}\int \ln\frac{q^{\mathrm{Ent}}(\theta)}{\pi}q_\theta^{\mathrm{MV}} d\theta. \tag{112}$$

We upper bound this term as follows. From Theorem 2 in Xu (2020), when the exponential interal condition $\mathrm{E}_{q_\theta^{\mathrm{MV}}} e^{\gamma \ln\pi} \leq e^{\gamma^2\tilde{\sigma}}$ is satisfied we have

$$\mathbb{E}_{q^*}\ln\pi - \mathrm{E}_{q_\theta^{\mathrm{MV}}}\ln\pi \leq \sqrt{2\tilde{\sigma}^2 \mathrm{KL}(q^*|q_\theta^{\mathrm{MV}})}. \tag{113}$$

From the assumption of the upper bound of the entropy, we have

$$\begin{aligned}
\Omega(q_\theta^{\mathrm{MV}}, q^*) &= \frac{1}{N\alpha}\int \ln\frac{q^*}{\pi}q^* d\theta - \frac{1}{N\alpha}\int \ln\frac{q_\theta^{\mathrm{MV}}}{\pi}q_\theta^{\mathrm{MV}} d\theta \\
&\leq -\frac{1}{N\alpha}\int \ln\pi q^* d\theta + \frac{1}{N\alpha}\int \ln\pi q_\theta^{\mathrm{MV}} d\theta + \frac{1}{N\alpha}H_0 \\
&\leq \frac{1}{N\alpha}\sqrt{2\tilde{\sigma}^2 \mathrm{KL}(q^*|q_\theta^{\mathrm{MV}})} + \frac{1}{N\alpha}H_0 \\
&\leq \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{\alpha}{2N}\mathrm{KL}(q^*|q_\theta^{\mathrm{MV}}) + \frac{1}{N\alpha}H_0.
\end{aligned} \tag{114}$$

Then we have

$$\begin{aligned}
&\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}} d\theta \\
&\leq \mathbb{E}\left[\frac{1}{N}\sum_i^N\left(\int \ln\frac{\nu(x_i)}{p_\theta(x_i)}q^* d\theta + (2+\alpha/(2N))\mathrm{KL}(q^*|q_\theta^{\mathrm{MV}})\right) + \frac{\mathrm{KL}(q_\theta^{\mathrm{MV}}|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right].
\end{aligned} \tag{115}$$

Since $\mathrm{KL}(q^*|q_\theta^{\mathrm{MV}}) \leq \chi^2(q^*|q_\theta^{\mathrm{MV}})$, and using the definition of the dual form of MV-risk,

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q^{\mathrm{Ent}}(\theta) d\theta \leq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{q_\theta^{\mathrm{MV}}} l_i(\theta) - \gamma\mathrm{Var}_q l_i(\theta) + \frac{\mathrm{KL}(q^{\mathrm{Ent}}(\theta)|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right], \tag{116}$$

where $\gamma = 2N/(4N+\alpha)$. Then, by definition q_θ^{MV} is the solution of the problem, we can write above as

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}} d\theta \leq \mathbb{E}\inf_\rho\left[\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\rho_\theta} l_i(\theta) - \gamma\mathrm{Var}_\rho l_i(\theta) + \frac{\mathrm{KL}(\rho|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right]. \tag{117}$$

Then, applying the dual form of Ent-risk again, we obtain We substitute $\rho = \phi_n$

$$\begin{aligned}
&\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}} d\theta \\
&\leq \mathbb{E}\inf_{\rho,\rho'}\left[\frac{1}{N}\sum_i^N\left(\int \ln\frac{\nu(x_i)}{p_\theta(x_i)}\rho' d\theta + (2+\alpha/(2N))\chi^2(\rho'|\rho)\right) + \frac{\mathrm{KL}(\rho|\pi)}{N\alpha} + \frac{\tilde{\sigma}^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right].
\end{aligned} \tag{118}$$

Then from the assumption, we assumed that $q(\theta;\phi_N)$ such that

$$\int \mathrm{KL}(\nu(x)|p(x|\theta))q(\theta;\phi_N)d\theta \leq \epsilon_N, \tag{119}$$

$$\mathrm{KL}(q(\theta;\phi_N)|\pi) \leq N\epsilon_N. \tag{120}$$

So, we substitute this to $\rho' = q(\theta;\phi_n)$ and $\rho' = \rho$, we obtain

$$\mathbb{E}_{\nu(x)}\int \mathrm{Hel}^2(p_\theta,\nu)q_\theta^{\mathrm{MV}} d\theta \leq \left[\frac{1+\alpha}{\alpha}\epsilon_N + \frac{\sigma^2}{N\alpha^2} + \frac{1}{N\alpha}H_0\right]. \tag{121}$$

So far we focused on model is well specified setting. As for the model misspecification, we just replace the log loss in Eq.(107) by

$$\mathrm{KL}(\nu(x)|p(x|\theta))=\mathrm{KL}(\nu(x)|p(x|\theta^*))+\mathbb{E}\ln\frac{p(X_n|\theta^*)}{p(X_n|\theta)}. \tag{122}$$

Then, the proof goes almost in the same way but we add model misspecified loss $\mathrm{KL}(\nu(x)|p(x|\theta^*))$. □

J DISPARITIES AND ROBUSTNESS

The robustness of Hellinger distance has been analyzed by the disparity distance and associated residuals (Lindsay, 1994). Here we present the model misspecification property of Ent-VI from the disparity analysis similarly to Hellinger distance.

We define the Pearson residual at x as

$$\delta(x)=\frac{\nu(x)-p(x|\theta)}{p(x|\theta)}. \tag{123}$$

We then define a disparity measure between $\nu(x)$ and $p(x|\theta)$ as

$$\rho(\nu,p_\theta):=\int G(\delta(x))p(x|\theta)dx, \tag{124}$$

where $G(\delta(x))$ is the strictly convex function. By choosing appropriate G, ρ corresponds to the family of power density divergences

$$D_{\mathrm{PD}}(\nu,p(\theta)):=\int\frac{(1+\delta)^{\lambda+1}-1}{\lambda(\lambda+1)}p(x|\theta)dx, \tag{125}$$

where $\lambda=0$ correspodns to the KL divergence and $\lambda=-1/2$ corresponds to the Hellinger divergence. We differentiate ρ concerning a parameter and obtain the estimating equation:

$$\nabla_\theta\rho=\int(G'(\delta)(\delta+1)-G(\delta))\nabla p(x|\theta)dx=0. \tag{126}$$

We then define a response function as

$$R(\delta(x)):=G'(\delta)(\delta+1)-G(\delta). \tag{127}$$

This measures how much the estimating equation changes under the difference of the Pearson disparity function. This response function is called the residual adjustment function (RAF). The intuition of RAF is that since $\delta(x)$ is the relative deviation of the model and data generating distribution and model misspecification implies large $\delta(x)$. Thus, robustness to model misspecification requires to have damped reponse as increase $\delta(x)$ and put small $R(\delta(x))$ for large δ For example, when ρ corresponds to the maximum likelihood estimation, $R(\delta)=\delta$, thus, this is efficient but not robust to model misspecification. On the other hand, Hellinger distance shows $\sqrt{\delta+1}-1$. Thus it is robust compared to KL divergence.

We extend above concepts in a Bayesian way. Since a model $p(x|\theta)$ is generated from approximate posterior $q(\theta;\phi)$. We focus on the objective function of VI. We consider using multi-sample bound. We define a disparity measure for each sample as

$$\delta(x,\theta_{m'})=\frac{\nu(x)-p(x|\theta_{m'})}{p(x|\theta_{m'})}. \tag{128}$$

Since the loss function of the standard VI is written as $\mathbb{E}_{q(\theta;\phi)}[\mathrm{KL}(\nu(x)|p(x|\theta))]$, we define it as the disparity function:

$$\rho(\nu,p_\theta,q):=\mathbb{E}_{q(\theta;\phi)}\mathrm{KL}(\nu(x)|p(x|\theta)). \tag{129}$$

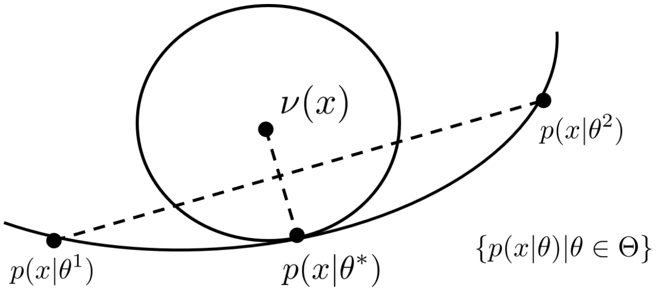

Figure 6: The illustration of risk-seeking property under model misspecification. Under model misspecification, the predictive distribution obtained from PB can get closer to the true distribution $\nu(x)$ than a single model $p(x^*)$.

We then differentiate the objective as follows. Assume that we use a reparametrized gradient where $\theta \sim T(\theta_0; \phi)$. Here, T is a translation, ϕ is a variational parameter, and θ_0 obeys some simple distribution. Then we calculate the derivative by considering a multi-sample bound:

$$\nabla_\phi \mathbb{E}_{q(\theta;\phi)} \mathrm{KL}(\nu(x)|p(x|\theta)) = \mathbb{E}_{\theta_0^1 \dots \theta_0^m \sim p(\theta_0)} \int \frac{1}{m} \sum_{m'=1}^{m} \frac{\nu(x)}{p(x|\theta_{m'})} \nabla_\phi T(\theta_0^{m'}; \phi) \nabla_\theta p(x|\theta_{m'}) dx. \tag{130}$$

Next we consider the loss of Ent-VI. We upper bound the objective by using the Jensen inequality:

$$-\frac{1}{\gamma} \ln \mathbb{E}_{\theta_0^{(1)} \dots \theta_0^{(m)}} \frac{1}{m} \sum_{m'=1}^{m} e^{\gamma \ln w(\theta_{m'})} \leq -\mathbb{E}_{\theta_0^{(1)} \dots \theta_0^{(m)}} \frac{1}{\gamma} \ln \frac{1}{m} \sum_{m'=1}^{m} e^{\gamma \ln w(\theta_{m'})}. \tag{131}$$

Then we calculate the gradient and we obtain

$$\begin{aligned}
\nabla_\theta \frac{1}{\gamma} \ln \frac{1}{m} \sum_{m'=1}^{m} e^{\gamma \ln w(\theta_{m'})} = \frac{e^{\gamma \ln w(\theta_{m'})}}{\sum_{m'=1}^{m} e^{\gamma \ln w(\theta_{m'})}} \nabla_\theta \ln w(\theta) &= \frac{e^{\gamma \ln w(\theta_{m'}) - \gamma \ln w(\theta_{\tilde{m}})}}{\sum_{m'=1}^{m} e^{\gamma \ln w(\theta_{m'}) - \gamma \ln w(\theta_{\tilde{m}})}} \nabla_\theta \ln w(\theta) \\
&= \frac{e^{\gamma \ln w(\theta_{m'}) - \gamma \ln w(\theta_{\tilde{m}})}}{1 + \sum_{m'=1, \neq \tilde{m}}^{m} e^{\gamma \ln \frac{w_{\theta'}}{w(\theta_{\tilde{m}})}}} \nabla_\theta \ln w(\theta)
\end{aligned} \tag{132}$$

where we define $\theta_{\tilde{m}} := \mathrm{argmax}_{\theta_{m'}} \ln w(\theta_{m'})$. With this expression, we have

$$\begin{aligned}
&\nabla_\phi \mathbb{E}_{\nu(x)} \left(-\gamma^{-1} \ln \mathbb{E}_{q(\theta;\phi)} e^{\gamma \ln \frac{p(x|\theta)}{\nu(x)}} \right) \\
&= \mathbb{E}_{\theta_0^{(1)} \dots \theta_0^{(m)}} \int \left[\sum_{m'=1}^{m} \nabla_\phi T(\theta_0^{m'}; \phi) \frac{e^{\gamma \ln w(\theta_{m'}) - \gamma \ln w(\theta_{\tilde{m}})}}{1 + \sum_{m'=1, \neq \tilde{m}}^{m} e^{\gamma \ln \frac{w_{\theta'}}{w(\theta_{\tilde{m}})}}} \frac{\nu(x)}{p(x|\theta_{m'})} \nabla_\theta p(x|\theta)|_{\theta_{m'} = T(\theta_0^{m'}; \phi)} \right] dx.
\end{aligned} \tag{133}$$

Thus, when $\gamma > 1$, the weights of the gradient other than the maximum become very small. Thus the RAF is given as

$$\frac{e^{\gamma \ln w(\theta_{m'}) - \gamma \ln w(\theta_{\tilde{m}})}}{1 + \sum_{m'=1, \neq \tilde{m}}^{m} e^{\gamma \ln \frac{w_{\theta'}}{w(\theta_{\tilde{m}})}}} \frac{\nu(x)}{p(x|\theta_{m'})}. \tag{134}$$

This shows much smaller response than the RAF of the standard VI. We numerically check the RAF as shown in the main paper, Section 4.3.

K INTUITION OF RISK SEEKING PROPERTY UNDER MODEL MISSPECIFICATION

Here we discuss the intuition of risk-seeking property under model misspecification. Fig. 6 depicts the settings under model misspecification. In the figure, the area below the curve is the set of parameterized models $p(x|\theta)$, and θ^* is the closest point to true distribution $\nu(x)$ in KL divergence. The usual variational inference makes the posterior distribution around θ^*. Under appropriate assumptions, the set of probability distributions $p(x|\theta)$ may not be a convex set (Grünwald and Van Ommen, 2017). For example, the average of Gaussian distributions results in a Gaussian mixture, which is outside the set of Gaussian distributions. Therefore, a mixture of models $p(x|\theta)$ can be closer to $\nu(x)$ than $p(x|\theta^*)$. Thus, when an approximate posterior $q(\theta)$ which is the solution of PB generates samples θ_1 and θ_2, the predictive distribution $\sum_{i=1,2} w_i p(x|\theta^i)$ can be closer to $\nu(x)$ than the standard Bayesian predictive distribution if the weights $\{w_i\}$ are appropriately chosen. RSO may provide such a dispersed approximate posterior as a solution. Thus, PB might be more capable of dealing with model misspecification since it uses risk-seeking optimization.

L DERIVATION OF THE TAYLOR EXPANSION OF PB

Here we present the derivation of Eqs.(21), (22), and (23) in the main paper.

Recall the definitions. We focus on deep Bayesian models and consider the MF approximation with $q(\theta;\phi) := N(\theta|\mu_1, \sigma^2 I_d)$. We restrict q' in Eq.(10) to the parametric distributions, and consider the MF Gaussian distribution: $q'(\theta;\phi') := N(\theta|\mu_2, \sigma^2 I_d)$. Here for simplicity, we assume that the variances of q and q' are the same. We express the difference of the mean as $s = \mu_2 - \mu_1$. Then samples from q' can be written as $s + \mu_1 + \sigma\xi$, where ξ is drawn from $N(0, I_d)$.

Then, first by using the formula of the KL divergence between Gaussian distributions, from Eq.(10), we obtain

$$\inf_{\mu,\sigma^2,s} \frac{1}{N}\sum_{n=1}^{N}\left[\mathbb{E}_\xi[l_n(\mu_1+s+\sigma\xi)]+\frac{\|s\|^2}{2\gamma\sigma^2}\right]+\frac{\beta\mathrm{KL}(q|\pi)}{N}. \tag{135}$$

Before solving the inner problem concerning s, we consider to expand the loss of the standard VI by the Taylor theorem as

$$\mathbb{E}_{q(\theta;\phi)}[l_n(\theta)] \approx l_n(\mu_1) + \mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2. \tag{136}$$

This expansion was introduced Tsuzuku et al. (2020) and approximation error is discussed there. To derive the above expansion, we assume that loss is twice differentiable, then by the taylor expnasion, we expand $l_n(\theta')$ around θ_1. Then there exists a constant $t \in (0,1]$ such that

$$l_n(\theta') = l_n(\theta_1) + \nabla_\theta l_n(\theta)|_{\theta_1}(\theta'-\theta_1) + \frac{1}{2}(\theta'-\theta_1)^\top \nabla_\theta^2 l_n(\theta)|_{\alpha(\theta'-\theta_1)+\theta_1}(\theta'-\theta_1). \tag{137}$$

Assume that the Hessian matrix satisfies M lipschitzness

$$\frac{\|\nabla_\theta^2 l_n(\theta)|_{\theta'} - \nabla_\theta^2 l_n(\theta)|_{\theta_1}\|_2}{\|\theta'-\theta_1\|_2} \leq M. \tag{138}$$

Then we have

$$l_n(\theta') \leq l_n(\theta_1) + \nabla_\theta l_n(\theta)|_{\theta_1}(\theta'-\theta_1) + \frac{1}{2}(\theta'-\theta_1)^\top \nabla_\theta^2 l_n(\theta)|_{\theta_1}(\theta'-\theta_1) + \frac{1}{2}M\|\theta'-\theta_1\|_2^3. \tag{139}$$

Then we take the average and set $\theta_1 = \mu_1$, we obtain

$$\mathbb{E}_{q(\theta;\phi)}[l_n(\theta)] \leq l_n(\mu_1) + \mathbb{E}_{q(\theta;\phi)}[\nabla_\theta l_n(\theta)|_{\theta_1}(\theta-\mu_1)] + \frac{1}{2}\mathbb{E}_{q(\theta;\phi)}(\theta-\mu_1)^\top \nabla_\theta^2 l_n(\theta)|_{\mu_1}(\theta-\theta_1) + \frac{1}{2}M\mathbb{E}_{q(\theta;\phi)}\|\theta-\mu_1\|_2^3. \tag{140}$$

Then by using the trace and Gaussian relation, we obtain

$$\mathbb{E}_{q(\theta;\phi)}[l_n(\theta)] \leq l_n(\mu_1) + \frac{1}{2}\mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2 + M\frac{\sqrt{2}\Gamma\left(\frac{n+3}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sigma^3. \tag{141}$$

Thus, in the main paper, we drop the third term for simplicity.

With this expansion, first, we solve the inner problem concerning s. We further assume that s is small. This is a kind of parametric assumption on q'.

Then by using the Taylor expansion, we can expand the inner problem of Eq.(135) using Eq.(140), and by dropping the third order term for simplicity, we obtain

$$\begin{aligned}\inf_s \mathbb{E}_\xi[l_n(\mu_1+s+\sigma\xi)]+\frac{\|s\|^2}{2\gamma\sigma^2} &\approx \inf_s l_n(\mu_1)+\frac{1}{2}\mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2+\nabla_\theta l_n(\mu_1)s+\frac{1}{2}s^\top\nabla^2 l_n(\mu_1)s+\frac{\|s\|^2}{2\gamma\sigma^2}\\ &= l_n(\mu_1)+\mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2-\nabla l_n(\mu_1)^\top H\nabla l_n(\mu_1), \end{aligned}\tag{142}$$

Note that from the first line to the second line, since the objective function is the quadratic function of $\nabla_\theta l_n(\mu_1)s+\frac{1}{2}s^\top\nabla^2 l_n(\mu_1)s+\frac{\|s\|^2}{2\gamma\sigma^2}$ with respect to s, we solved it analytically as follows:

$$\inf_s \mathbb{E}_\xi[l_n(\mu_1+s+\sigma\xi)]+\frac{\|s\|^2}{\gamma\sigma^2}\approx l_n(\mu_1)+\frac{1}{2}\mathrm{Tr}[\nabla^2 l_n(\mu_1)]\sigma^2-\nabla l_n(\mu_1)^\top H\nabla l_n(\mu_1), \tag{143}$$

where $s=H\nabla l_n(\mu_1)$ and $H:=[\gamma^{-1}\sigma^{-2}I_d+\nabla^2 l_n(\mu_1)]^{-1}$.

M DERIVATION OF THE LOWER BOUND OF THE MARGINAL LIKELIHOOD

We use the second order Jensen inequality developed in Futami et al. (2021). We cite their theorem

Theorem 3. *When $p(x|\theta)<\infty$ for all x and θ, we have*

$$\mathbb{E}_{q(\theta)}\ln p(x|\theta)\le \ln\mathbb{E}_{q(\theta)}p(x|\theta)-\mathbb{E}_{q(\theta)}\left(\frac{\ln p(x|\theta)-\mathbb{E}_{q(\theta)}\ln p(x|\theta)}{2h(x,\theta)}\right)^2, \tag{144}$$

where

$$h(x,\theta)^{-2}=\exp\left(\ln p(x|\theta)+\mathbb{E}_{q(\theta)}\ln p(x|\theta)-2\max_\theta \ln p(x|\theta)\right). \tag{145}$$

They call this the loss function based second order Jensen inequality. We apply this to the marginal likelihood.

Define

$$w_j=\frac{p(x|z_j)p(z_j)}{p(z_j)}, \tag{146}$$

where z_j is the j-th drawn sample from the approximate posterior distribution. We also define a empirical distribution of z_j:

$$\rho_E(z):=\frac{1}{J}\sum_{j=1}^J\delta_{z_j}(z) \tag{147}$$

Next, We express a marginal log likelihood by using the multi-sample of z_j as

$$\begin{aligned}\ln p(x)=\ln\int p(x|z)p(z)dz &= \ln\left[\mathbb{E}_{(z_1,\dots,z_J)\sim q(z)}\mathbb{E}_{\rho_E}w_j\right]\\ &= \ln\left[\mathbb{E}_{(z_1,\dots,z_J)\sim q(z)}\frac{1}{N}\sum_{j=1}^J w_j\right]\end{aligned}\tag{148}$$

First, we derive a bound of IWAE-VAR. We apply the second order Jensen inequality assuming that $q(\theta)=q(z)$

and $p(x|\theta) = \frac{1}{J}\sum_{j=1}^{J} w_j$ in Theorem 3, and then we obtain

$$
\begin{aligned}
&\ln p(x) \\
&= \ln \left[\mathbb{E}_{(z_1,\dots,z_J)\sim q(z)} \frac{1}{J}\sum_{j=1}^{J} w_j \right] \\
&\geq \mathbb{E}_{\{(z_1^m,\dots,z_J^m)\}_{m=1}^M \sim q(z)} \frac{1}{M}\sum_{m=1}^{M} \ln\left[\frac{1}{J}\sum_{j=1}^{J} w_j^m\right] \\
&+ \mathbb{E}_{\{(z_1^m,\dots,z_J^m)\}_{m=1}^M \sim q(z)} \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{(2h(x,\{z^m\}))^2}\left(\ln\frac{1}{J}\sum_{j=1}^{J} w_j^m - \frac{1}{M}\sum_{m=1}^{M}\ln\frac{1}{J}\sum_{j=1}^{J} w_j^m\right)^2\right],
\end{aligned} \tag{149}
$$

where

$$
h(x,\{z^m\})^{-2} = \exp\left(\ln\frac{1}{J}\sum_{j=1}^{J} w_j^m + \frac{1}{M}\sum_{m=1}^{M}\ln\frac{1}{J}\sum_{j=1}^{J} w_j^m - 2\max_m \ln\frac{1}{J}\sum_{j=1}^{J} w_j^m\right). \tag{150}
$$

Here we consider that we prepare M sets of empirical distribution $\{\rho_E(z^m)\}_{m=1}^M$ and applied the second order Jensen inequality. Thus, this is the multi-sample bound that requires NM samples.

This bound is similar to the bound developed in Rainforth et al. (2018), when we dropped the weighted variance term. The first term in

Next we derive VI-VAR bound. From a multi sample bound, we first apply a normal Jensen inequality with respect to $q(z)$ to Eq.(148), we obtain

$$
\ln p(x) \geq \mathbb{E}_{(z_1,\dots,z_J)\sim q(z)} \ln\left[\mathbb{E}_{\rho_E} w_j\right]. \tag{151}
$$

We then apply the second order Jensen inequality with resepct to ρ_E, then obtain

$$
\ln p(x) \geq \mathbb{E}_{(z_1,\dots,z_J)\sim q(z)}\left[\frac{1}{J}\sum_{j=1}^{J}\ln w_j + \frac{1}{J}\sum_{j=1}^{J}\frac{1}{(2h(x,z_j))^2}\left(\ln w_j - \frac{1}{J}\sum_{j=1}^{J}\ln w_j\right)^2\right], \tag{152}
$$

where

$$
h(x,z_j)^{-2} = \exp\left(\ln w_j + \frac{1}{J}\sum_{j=1}^{J}\ln w_j - 2\max_j \ln w_j\right). \tag{153}
$$

N EXPERIMENTAL SETTINGS AND ADDITIONAL EXPERIMENTAL RESULTS

Here we present the detailed settings of the experiments in the main paper and present additional experimental results.

N.1 Misspecified settings

In this section, we present the settings and additional results for the misspecified models.

N.1.1 Toy data experiments

The experimental settings are similar to Heide et al. (2020). The difference is that they used MCMC to approximate the posterior distributions. We approximate the posterior by VI. As our model, we used $N(y|f(x|w),\sigma^2)$ with a linear model of Fourier basis $f(x|w) = \frac{1}{\pi}\sum_{k=0}^{40} w_k^0\cos(kx) + w_k^1\sin(kx)$.

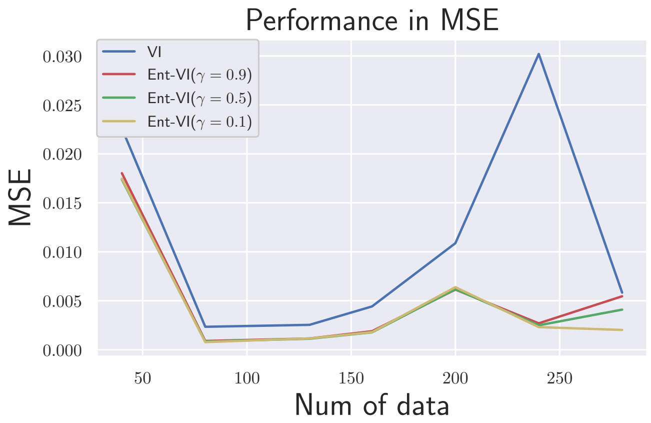

Figure 7: MSE using sparse priors under different γs

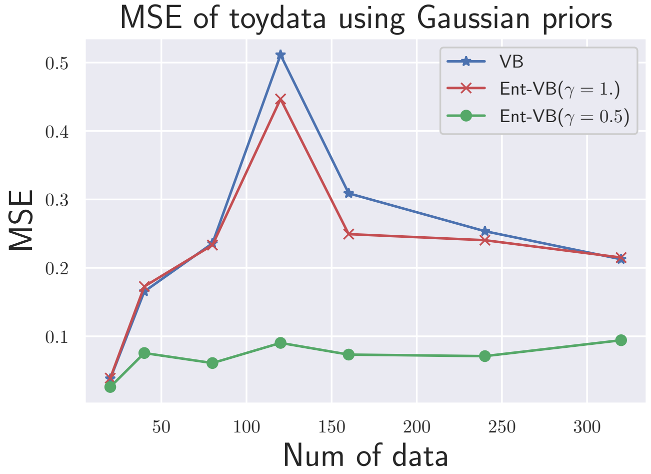

Figure 8: MSE using Gaussian priors

Prior distributions are imposed on w and σ^2. For σ^2, we put on an inverse Gamma prior distribution. For w, we considered using Gaussian and Laplace distribution priors. In the main paper, we only showed the result of a Laplace distribution setting. By assuming a Laplace distribution, we expect that it induces a sparsity and w become close to 0, which is the true model. We used the non-centered parameterization (Papaspiliopoulos et al., 2007) to approximate the posterior distribution by mean-field VI. For example, Ingraham and Marks (2017) proposed to approximate sparse priors using Gaussian MF posteriors using the on-centered parametrization. We use their approach to parametrize the approximate posteriors. To draw a posterior sample of w, we draw two Gaussians z_1 and z_2 from two different approximate posteriors of Gaussian distribution. Then we regard $\tilde{w} = z_1 \odot z_2$ as the sample of approximate posterior of w.

Under this setting, we trained the objective of Ent-VI using the multi-sample bound $m = 1000$ and $\gamma = 0.1$ in the main paper. We also use 1000 samples for the standard VI.

Here, we show additional results for different γ.

Next, we present the result when the prior is Gaussian distribution. We trained the objective of Ent-VI using the multi-sample bound $m = 100$. The results are shown in Fig.8. We found that using $\gamma = 1.$ seems not enough to enhance the robustness.

Table 4: MSE real dataset using sparse prior

Dataset	VI	Ent-VI($\gamma = 0.1$)	Ent-VI($\gamma = 0.5$)	Ent-VI($\gamma = 1.0$)	Ent-VI($\gamma = 2.0$)
Variable star	2.0 ± 0.9	2.0 ± 0.8	$\mathbf{1.8 \pm 0.6}$	1.9 ± 0.7	25.0 ± 12.9
(ppm)^2	155 ± 20	153 ± 22	149 ± 21	$\mathbf{129 \pm 16}$	255 ± 23

N.1.2 Additional toy data experimetns

We addressed an additional toy data regression task and measured its performance in the MSE. We generated toydata $(x_i, y_i)_{i=1}^N$ as follows: $x_i = u_i$ and $y_i = -x + 1 + \epsilon + 2 * \epsilon' \zeta_i'$ where $u_i \sim \text{Uniform}[-2, 2]$, $\epsilon, \epsilon' \sim N(0, 1)$, and ζ_i' follows a Bernoulli distribution, which takes 0 with a probability 1/2. Here ϵ and ϵ' are independent.

Thus, the true conditional expectation is $\mathbb{E}[Y|X] = -x + 1$. As our model, we used $p(y|x, \theta) = N(y|f(x; w, b), \sigma^2)$ with $f(x; w, b) = wx + b$ where $\theta = \{w, b, \sigma\} \in \mathbb{R}$. Following the theory from Section 4.2, our model includes $\mathbb{E}[Y|X]$, although the noise assumption is misspecified. Prior distributions are imposed on w, b and σ^2. For $\pi(w, b) = N(w|0, 1)N(b|0, 1)$.As for σ^2, we put on a inverse Gamma distribution (InvGamma(1., 1.)). As for the posterior distribution, $q(w, b) = N(w|w_\mu, w_\sigma)N(b|b_\mu, b_\sigma)$ and $q(\sigma^2) = \text{InvGamma}(\alpha, \beta)$.

We observed i) Excess risk, $\mathbb{E}_{\nu,q} \ln p(y|x, \theta)$, which is controlled by Corollary 1, ii) Test LL, $\mathbb{E}_\nu \ln \mathbb{E}_q p(y|x, \theta)$, iii) MSE, $\mathbb{E}_\nu |\mathbb{E}[Y|X] - f(x; w, b)|^2$, iv) variational parameters $\alpha, \beta, w_\mu, b_\mu, w_\sigma, b_\sigma$, and v) the mean of the posterior variance given as the mean of the InvGamma distribution, $\beta/(\alpha - 1)$.

We measured those values by changing the number of training datasets from 20 to 500. We optimized the objective function by GD by approximating the posterior expectation using 3000 samples drawn from q. We compare the standard VI and Ent-VI using different γs. The result is shown in Figure 9.

We found that MSEs are almost the same in Ent-VI and the standard VI methods, and w_μ and b_μ seem to converge to optimal values, $w_\mu = -1$ and $b_\mu = 1$. On the other hand, the noise parameter σ, which is related to the model misspecification, behaves differently in Ent-VI and the standard VI. Especially, α and β are significantly different in all the methods. When comparing the mean of the InvGamma distribution, $\beta/(\alpha - 1)$, it is much larger in the standard VI than in Ent-VI. Since $\beta/(\alpha - 1)$ corresponds to the mean of σ^2, this means that the standard VI method fails to estimate σ^2 especially when the number of the training dataset is small. On the other hand, Ent-VI methods estimate it appropriately. Thus, the better estimation of σ in Ent-VI results in better performance in the test LL and the excess risk of Ent-VI methods.

N.1.3 Real data experiments

Next, following Heide et al. (2020), we consider a linear regression task on real datasets. The settings are almost the same as the toy data experiments. We trained the objective of Ent-VI using the multi-sample bound $m = 100$ for all the experiments.

The first experiment is the variable star dataset. The data is available in R-package, **asta**. We consider a one-dimensional input and one-dimensional output regression task similar to the toy dataset experiment. The input is the time, and the output is the feature of stars in the dataset. We randomly select 500 data points as training data and 100 data points as the test data. We used the 51 number Fourier basis.

The second experiment is the London air pollution dataset. This data is also available in R-package, **openair**. We use the data from a monitoring station at London N. Kensington (code is KC1). Our task is that given one-dimensional time data, we predict the concentration value of NO2. As the training dataset, we used the time series air pollution data starting from Monday, January 7, 2013, at midnight until the final date of January 2013. We also have data from Monday, January 6, 2014, at midnight until the final date of January 2014. As the test data, we predict the NO2 value starting from Monday, January 5, 2015, at midnight until the final date of January 2015. We used the 201 number Fourier basis for the air pollution dataset. These results are shown in Tables 4 and 5 under using sparse priors and Gaussian priors.

From these experiments, we found that using $\gamma \leq 1$ seems promising for these simple tasks. On the other hand, using $\gamma \geq 1$ should be avoided.

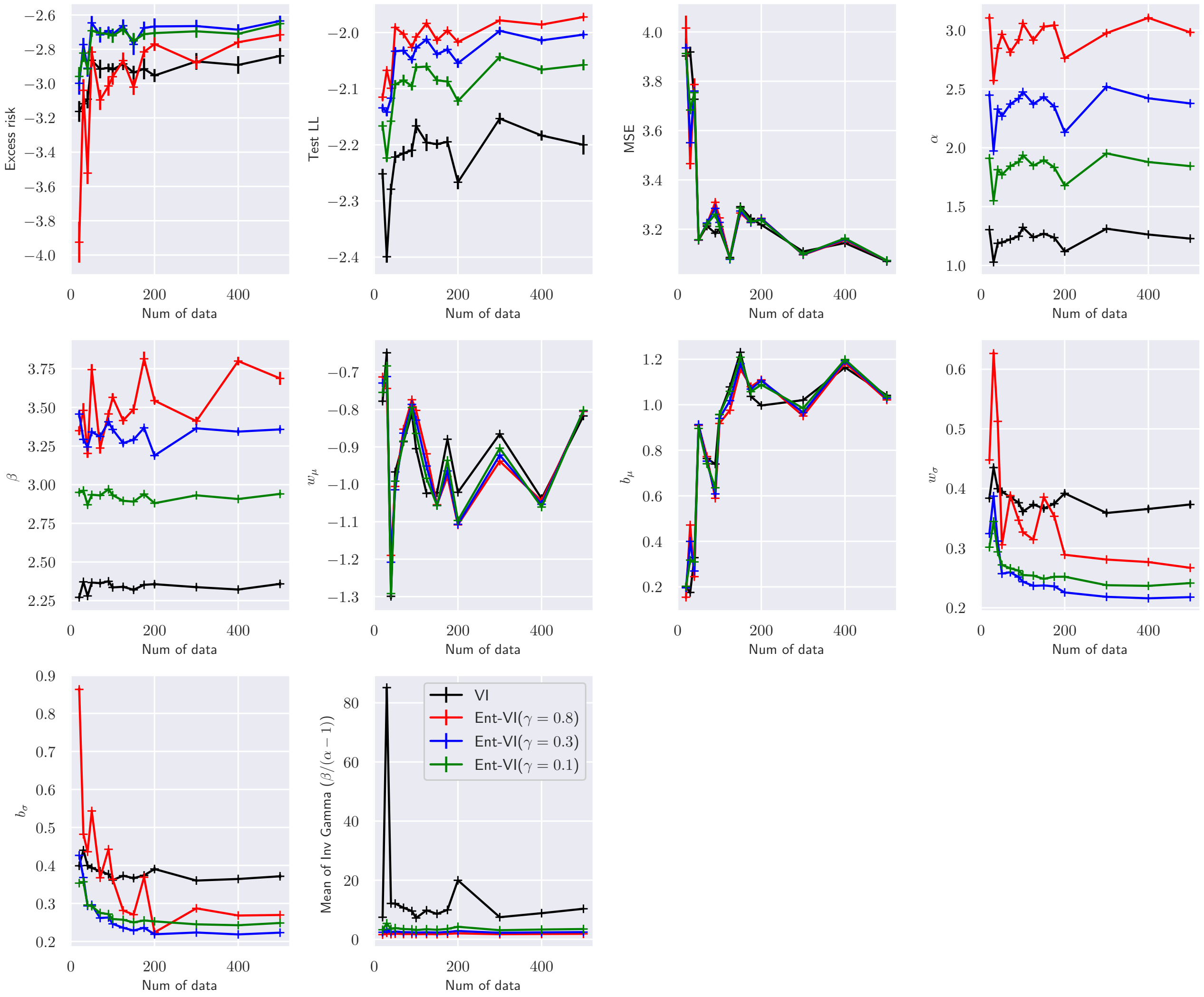

Figure 9: Comparison of Ent-VI using different γs and the standard VI

N.2 Flat minimam in deep learning

Experimental settings are almost the same as previous works (Masegosa, 2020; Morningstar et al., 2020). We set $m = 4$ for the multi-sample bounds in the main paper.

Here we show additional numerical experiments. We changed the number of samples in Ent-VI. We measured at $m = 4, 8, 16$. As for the standard VI, we used $m = 4$. The result is shown in Figure 10. We found that using large m results in better generalization error behavior. Since when we increase m, the bound becomes tighter. This indicates that using a tighter bound might be useful for obtaining better generalization ability.

N.3 Variational autoencoder

The experimental settings, including the network architecture and hyperparameters, are the same as in Shi et al. (2017). As for the MNIST experiments, we used two hidden ReLU layers with 500 units, and the latent space dimension is 8. We used $J = 10$ for VI-VAR and VI and IWAE. As for IWAE-VAR, we need to split the multi-samples for J and M. We used $J = 2$ and $M = 5$. The test likelihood is calculated by using the annealed importance sampling.

Next, we also present different results for changing the number of particles we use in the experiment. We used

Table 5: MSE real dataset using Gaussian prior

Dataset	VI	Ent-VI($\gamma = 0.1$)	Ent-VI($\gamma = 0.5$)	Ent-VI($\gamma = 1.0$)	Ent-VI($\gamma = 2.0$)
Variable star	5.5 ± 2.7	5.5 ± 3.1	$\mathbf{4.4 \pm 3.2}$	6.0 ± 3.2	18.0 ± 5.5
(ppm)^2	129.72 ± 25	$\mathbf{123 \pm 30}$	130 ± 30	129.72 ± 29	260 ± 25

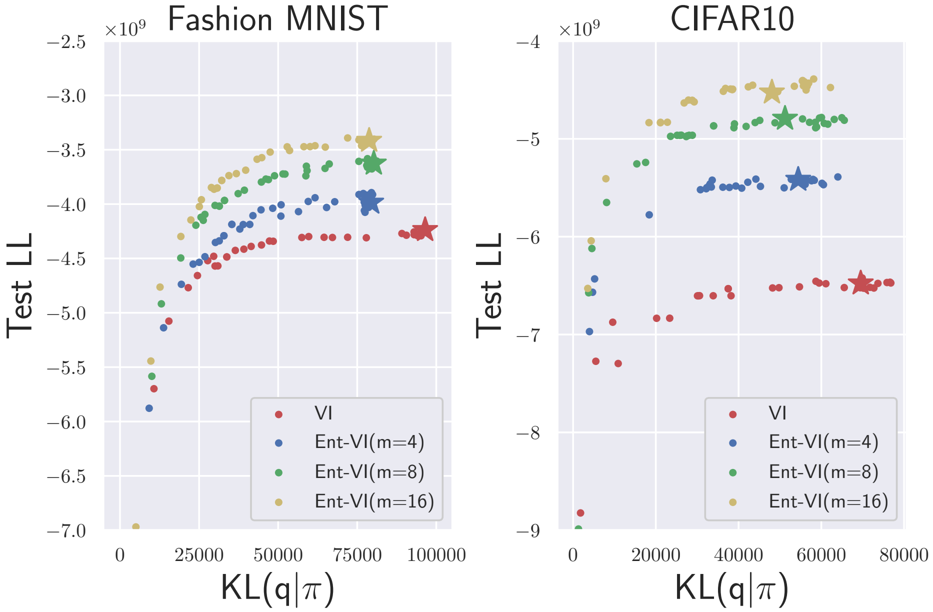

Figure 10: The results of changing the number of samples in the multi-sample bound. LL indicates the log-likelihood.

$J = 20, 40$ for VI-VAR and VI and IWAE. As for IWAE-VAR, we show the best result in the table from the combination of $(J, M) = (2, 10), (4, 5)$ for $J = 20$ and $(J, M) = (4, 10), (4, 10), (2, 10)$ for $J = 40$.

Table 6: MNIST VAE results

Method	IWAE-VAR	IWAE	VI-VAR	VI
Test LL ($J = 10$)	-88.8	-89.0	-89.7	-89.9
Test LL ($J = 20$)	-88.0	-88.1	-89.6	-88.9
Test LL ($J = 40$)	-87.9	-87.8	-88.6	-88.8

As for the CelebA experiments, we used DCGAN structure for the decoder, and the latent dimension is 32. As for the encoder, we used the symmetric structure to decoder except for the final layer, which is flattened and added Gaussian noise. We used $J = 10$ for VI-VAR and VI and IWAE. As for IWAE-VAR, we need to split the multi-samples for J and M and used $J = 2$ and $M = 5$.

N.4 Contextual bandit

Motivated from the previous work of Futami et al. (2021), we applied Ent-VI and MV-VI methods to the contextual bandit tasks. The experimental settings are precisely the same as the previous work of Futami et al. (2021).

We place a prior for a reward depending on the context and action. This prior distribution is updated to a posterior distribution. At each step, Thompson sampling selects the action, and then the corresponding posterior is updated by the observed reward.

We consider a neural network where the input is the context and the dimension of the output is the same as the action space, and we assume a prior on parameters of the network. We approximate the posterior of the neural network and express the uncertainty by the approximate posterior distribution. Following Futami et al. (2021), we consider a ensemble framework to approximate posteriors and we tired $\mathrm{PAC}^2_\mathrm{E}$ in Masegosa (2020) and VAR in Futami et al. (2021) as MV-VI. We consider Ent-VI with $\gamma = 1$ and approximate the objective using the multi-sample bound in Eq.(5) with ensembles. We considered 20 ensembles for all experiments. The result is shown in Table 7. We found that Ent-VI shows inferior performance in all the experiments. We conjectured that since $\mathrm{PAC}^2_\mathrm{E}$ and VAR are MV-VI methods, thus there exist repulsion forces, which pushes away ensembles with each other. This leads to better performance in ensemble learning. On the other hand, since Ent-VI with Eq.(5) does not have such a diversity enhancing term in the objective function. Thus Ent-VI showed inferior performance as ensemble learning.

Table 7: Results of contextual bandit

Dataset	$\mathrm{PAC}^2_\mathrm{E}$	VAR	Ent-VI
Mushroom	0.033±0.010	**0.029±0.004**	0.410±0.060
Financial	0.191±0.028	**0.152±0.027**	0.685±0.051
Statlog	0.032±0.0027	**0.006±0.0005**	0.397±0.040
CoverType	0.390±0.005	**0.289±0.003**	0.801±0.040