Epistemic Uncertainty and Excess Risk in Variational Inference

Futoshi Futami
Osaka University / RIKEN AIP

Abstract

Bayesian inference is widely used in practice due to its ability to assess epistemic uncertainty (EU) in predictions. However, its computational complexity necessitates the use of approximation methods, such as variational inference (VI). When estimating EU within the VI framework, metrics such as the variance of the posterior predictive distribution and conditional mutual information are commonly employed. Despite their practical importance, these metrics lack comprehensive theoretical analysis. In this paper, we investigate these EU metrics by providing their novel relationship to excess risk, which allows for a convergence analysis based on PAC-Bayesian theory. Based on these analyses, we then demonstrate that some existing objective functions of VI regularize EU metrics in different ways leading to different performance in EU evaluation. Finally, we propose a novel objective function for VI that directly optimizes both prediction and EU under the PAC-Bayesian framework. Experimental results indicate that our algorithm significantly improves EU estimation compared to existing VI methods.

1 INTRODUCTION

As machine learning applications continue to expand, understanding the uncertainty in predictions is becoming increasingly critical to enhance the reliability of these algorithms (Bhatt et al., 2021). Uncertainty refers to the variability in predictions resulting from incomplete information, and it can be classified into two categories (Bhatt et al., 2021): aleatoric uncertainty (AU), which arises from inherent noise in the data, and epistemic uncertainty (EU), which is due to a lack of sufficient training data.

Bayesian inference is frequently employed to capture EU, as the posterior distribution, which is updated from a prior distribution, can effectively represent uncertainty stemming from data scarcity (Hüllermeier & Waegeman, 2021). Bayesian approaches integrated with deep learning, known as Bayesian deep learning, are thus utilized in scenarios where EU estimation is crucial, such as dataset shift (Ovadia et al., 2019), adversarial example detection (Ye & Zhu, 2018), active learning (Houlsby et al., 2011), Bayesian optimization (Hernández-Lobato et al., 2014), and reinforcement learning (Janz et al., 2019). Due to the computational infeasibility of exact Bayesian inference in these settings, approximate methods such as variational inference (VI) are commonly used (Bishop, 2006). In VI, the variance of the posterior predictive distribution and conditional mutual information (Kendall & Gal, 2017; Depeweg et al., 2018) are standard metrics for quantifying EU.

Despite their widespread use, the theoretical understanding of these EU metrics under VI remains limited. For example, there is no existing analysis on their convergence rates concerning the training data size, which is a fundamental property of EU metrics. Conventional EU analyses typically assume access to the exact Bayesian posterior and predictive distributions (Fiedler et al., 2021; Lederer et al., 2019) and focus on their asymptotic behavior as sample sizes increase (Clarke & Barron, 1990). However, we rely on variational posterior distributions in VI, making these traditional techniques inapplicable. While the predictive performance of VI has been explored using the PAC-Bayesian framework (Alquier, 2021), EU metrics differ from standard generalization bounds, leaving it unclear how PAC-Bayesian theory can be directly applied to their analysis.

To address these challenges, we investigate these EU metrics by providing their novel relationship to excess risk in VI. Excess risk, formally introduced in Section 2.1, is defined as the difference between the expected test loss and the loss under the optimal parameter, which represents the loss due to insufficient training data. While excess risk is conceptually similar to the EU, it cannot be used as the practical EU metric since we cannot evaluate it numerically in the algorithm for the standard VI settings. Nevertheless, excess risk has been rigorously studied using PAC-Bayesian theory and is well understood theoretically (Alquier, 2021).

Building on this background, our first contribution is es-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

tablishing a new connection between excess risk and the widely used EU metrics in VI (Section 3.1). This connection is derived from the excess risk perspective of the EU under Bayesian decision theory (Xu & Raginsky, 2022). Specifically, we introduce a new joint distribution for the test data, training data, and learned parameters (Eq.(7)) that aligns with the VI framework. With this formulation, we define a new excess risk, termed Bayesian excess risk (BER) (Eq.(8)), and demonstrate that the commonly used EU metrics are equivalent to this new excess risk (Theorem 1).

Leveraging this new excess risk, we establish formal connections between this new excess risk and the standard excess risk, showing that our new excess risk provides a lower bound for the standard excess risk (Theorems 2 and 3). This finding implies that the widely used EU metrics serve as a lower bound for the standard excess risk, aligning with the frequently observed underestimation of uncertainty in these metrics. Through this relationship, we also present the first convergence analysis of EU metrics using PAC-Bayesian theory (Theorems 2 and 3).

Building on these theoretical insights, we further examine how different objective functions in VI implicitly regularize the excess risk compared to standard VI, which minimizes the Kullback–Leibler (KL) divergence. Specifically, we focus on the α-divergence-based VI (Li & Gal, 2017; Sheth & Khardon, 2020; Masegosa et al., 2020; Futami et al., 2021) and demonstrate that α-divergence VI applies weaker regularization on Bayesian excess risk (BER). This provides a new interpretation for why epistemic uncertainty (EU) estimates derived from α-divergence VI are often larger than those from KL divergence VI. Motivated by this observation, we introduce a novel VI algorithm that directly optimizes a new excess risk using PAC-Bayesian theory (Eq. (16)). Numerical experiments show that our approach significantly enhances EU evaluation compared to existing VI methods.

2 PRELIMINARIES

In this study, capital letters such as X represent random variables, and lowercase letters such as x represent deterministic values. We summarize the notations and settings in Appendixes A and B.

2.1 Variational Inference and Excess Risk

We consider supervised learning and denote input output pairs by $Z := (X, Y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. We assume that all the data are independent and identically distributed (i.i.d.) from an unknown data distribution $\nu(Z) = \nu(Y|X)\nu(X)$. Learners can access N training data, $\mathbf{Z}^N := (Z_1, \ldots, Z_N)$ with $Z_n := (X_n, Y_n)$, which are generated by $\mathbf{Z}^N \sim \nu(Z)^N$. We express $\nu(Z)^N$ as $\nu(\mathbf{Z}^N)$ and the conditional distribution $\nu(Y|X = x)$ as $\nu(Y|x)$ for simplicity. We aim to estimate the predictor of Y given X from the training dataset. We use the parametric model $p(y|x, \theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ represents the parameters. For the latter purpose, we assume that the mean of Y under this model can be expressed as $\mathbb{E}_{p(Y|x,\theta)}[Y] := m_\theta(x)$, where m_θ is a parametric function. For example, in a Gaussian model for $y \in \mathbb{R}$, we express the model as $p(y|x, \theta) = N(y|m_\theta(x), v^2)$ with the mean $m_\theta(x) \in \mathbb{R}$ and the variance $v^2 \in \mathbb{R}^+$.

To evaluate the performance of supervised learning, we define a loss function ℓ and focus on log loss and squared loss in this work. When using the model $p(y|x, \theta)$, the log loss is defined as $\ell(y, p(\cdot|x, \theta)) = -\ln p(y|x, \theta)$. The squared loss is defined as $\ell(y, m_\theta(x)) = |y - m_\theta(x)|^2$. Note that for the squared loss, the goal is to estimate the conditional mean $\mathbb{E}_{\nu(Y|X=x)}[Y|X = x]$ since the Bayes decision function, which achieves the minimum achievable risk for all measurable functions, is $\mathbb{E}_{\nu(Y|X=x)}[Y|X = x]$ for the squared loss (Steinwart & Christmann, 2008). In general, a loss function is defined as $l : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ is the action space. In this context, for the log loss, $\mathcal{A}$ is the set of all distributions on Y, whereas for the squared loss, $\mathcal{A} = \mathbb{R}$. Following this notation, we can handle log loss and squared loss together and we summarize them as $\ell(y, f_\theta(x))$, where $f_\theta(x) : \Theta \times \mathcal{X} \to \mathcal{A}$ refers to $p_\theta(\cdot|x, \theta)$ for the log loss, and $m_\theta(x)$ for the squared loss.

We consider Bayesian inference and use variational inference to estimate the parameter θ. We define the expected test and training losses with respect to the approximate posterior distribution $q(\theta|\mathbf{z}^N) \in \mathcal{Q}$, where $\mathcal{Q}$ is a family of distributions over θ, as follows:

$$R_q^\ell(Y|X, \mathbf{Z}^N) := \mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{q(\theta|\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\ell(Y, f_\theta(X)),$$

$$r_q^\ell(\mathbf{Z}^N) := \mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{q(\theta|\mathbf{Z}^N)}\frac{1}{N}\sum_{n=1}^N \ell(Y_n, f_\theta(X_n)).$$

We emphasize the log loss is used by denoting the corresponding test loss as $R_q^{\log}(Y|X)$. Similarly, for squared loss, we denote it as $R_q^{(2)}(Y|X)$. In this way, the corresponding part of ℓ is replaced accordingly in each case for the training loss.

The PAC-Bayesian theory (Alquier, 2021) provides guarantees on the gap between test and training losses. When the loss satisfies σ^2 sub-Gaussian property (See Appendix C for details), the generalization gap is bounded as follows

$$R_q^\ell(Y|X, \mathbf{Z}^N) \le R_q^\ell(\mathbf{Z}^N) + \frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))}{\lambda} + \frac{\lambda\sigma^2}{2N},$$

where λ is any positive constant and $p(\theta)$ is a prior distribution independent of the training data. The connection between PAC-Bayesian theory and VI has been extensively studied (see Germain et al. (2016), for example). Specifically, given training data $\mathbf{Z}^N = \mathbf{z}^N$, an approximate posterior distribution can be obtained by minimizing the upper bound of the generalization gap from the above inequality;

$$\hat{q}(\theta|\mathbf{z}^N) = \underset{q(\theta|\mathbf{z}^N)\in\mathcal{Q}}{\mathrm{argmin}}\ r_q^\ell(\mathbf{z}^N) + \frac{\mathrm{KL}(q(\theta|\mathbf{z}^N)|p(\theta))}{\lambda}. \quad (1)$$

When the log loss is used with $\lambda = N$, this minimization problem is equivalent to the **variational inference (VI)** in Bayesian inference (Germain et al., 2016) and q is called the variational posterior distribution.

In addition to the test loss, PAC-Bayesian theory provides the upper bound of the **excess risk (ER)** defined as

$$\mathrm{ER}^\ell(Y|X,\mathbf{Z}^N,\theta^*) := R_q^\ell(Y|X,\mathbf{Z}^N) - R^\ell(Y|X,\theta^*),$$

where $\theta^* := \mathrm{argmin}_{\theta\in\Theta}\mathbb{E}_{\nu(Z)}\ell(Y, f_\theta(X))$ is the optimal parameter and $R^\ell(Y|X,\theta^*) := \mathbb{E}_{\nu(Z)}\ell(Y, f_{\theta^*}(X))$. We remark that although $\mathrm{ER}^\ell(Y|X,\mathbf{Z}^N,\theta^*)$ should be expressed as $\mathrm{ER}_q^\ell(Y|X,\mathbf{Z}^N,\theta^*)$ since it depends on q, we omit q to simplify the notation. Under additional moderate assumptions, such as the choice of prior and posterior distributions and loss functions (See Appendix C for details), we have

$$\mathrm{ER}^\ell(Y|X,\mathbf{Z}^N,\theta^*) \leq \frac{C}{\sqrt{N}} + \frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))}{\lambda} + \frac{\lambda\sigma^2}{2N}. \quad (2)$$

The constant C depends only on the problem. See Appendix C for details. Since $R^\ell(Y|X,\theta^*)$ cannot be reduced by increasing the training data size, it expresses the fundamental difficulty of learning. Thus, excess risk evaluates the effect of limited training data, as it is the difference between the test risk and the fundamental difficulty of learning.

Although Eq. (2) focuses on the average test loss over the posterior distribution, we commonly use the predictive distribution in the Bayesian inference (Sheth & Khardon, 2017). We define the **prediction risk (PR)**:

$$\mathrm{PR}_q^\ell(Y|X) := \mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\ell(Y, \mathbb{E}_{q(\theta|\mathbf{Z}^N)}f_\theta(X)). \quad (3)$$

When the log loss is used,

$$\mathrm{PR}_q^{\log}(Y|X,\mathbf{Z}^N) = -\mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\log p^q(Y|X,\mathbf{Z}^N),$$

where $p^q(y|x,\mathbf{z}^N) := \mathbb{E}_{q(\theta|\mathbf{z}^N)}p(y|x,\theta)$ is the approximate predictive distribution. Thus, $\mathrm{PR}_q^{\log}(Y|X,\mathbf{Z}^N)$ represents the log loss of the predictive distribution, which is commonly used in the analysis of the Bayesian inference (Watanabe, 2009, 2018). Similarly to the excess risk in Eq. (2), we define the **prediction excess risk (PER)** using Eq. (3), which plays a crucial role in our analysis:

$$\mathrm{PER}^\ell(Y|X,\mathbf{Z}^N) := \mathrm{PR}_q^\ell(Y|X,\mathbf{Z}^N) - R^\ell(Y|X,\theta^*). \quad (4)$$

PER has also been studied using the PAC-Bayesian theory (Sheth & Khardon, 2017). Given $(x, \mathbf{z}^N)$, we express the conditional PER as $\mathrm{PER}^\ell(Y|x,\mathbf{z}^N)$, which is formally defined in Appendix D.2.

2.2 Widely Used Epistemic Uncertainty Metrics

Although ER and PER represent the effect of the insufficient training data, they are impractical for evaluating EU since Eqs. (2) and (4) require both $\nu(Y|x)$ and θ^*, which are not available in practice. In practice, we would like to evaluate the EU of the prediction at the test input x using $q(\theta|\mathbf{z}^N)$ and model $p(\cdot|x,\theta)$.

In practice, the following types of EU metrics are widely used. For the log loss, conditioned on $(X,\mathbf{Z}^N) = (x,\mathbf{z}^N)$, approximate mutual information has been widely used for EU evaluation (Depeweg et al., 2018):

$$I_q(\theta;Y|x,\mathbf{z}^N) := H[p^q(Y|x,\mathbf{z}^N)] - \mathbb{E}_{q(\theta|\mathbf{z}^N)}H[p(Y|x,\theta)], \quad (5)$$

where $H[p^q(Y|x,\mathbf{z}^N)] := -\mathbb{E}_{p^q(Y|x,\mathbf{z}^N)}\log p^q(Y|x,\mathbf{z}^N)$ is the entropy of the approximate predictive distribution, and $\mathbb{E}_{q(\theta|\mathbf{z}^N)}H[p(Y|x,\theta)]$ is the conditional entropy. $I_q(\theta;Y|x,\mathbf{z}^N)$ has been used in Bayesian experimental design (Foster et al., 2019) and reinforcement learning (Depeweg et al., 2018). Taking the expectation, we have $I_q(\theta;Y|X,\mathbf{Z}^N) = \mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}I_q(\theta;Y|x,\mathbf{z}^N)$.

In the case of the squared loss, the posterior variance of the model's prediction is often used for EU. This is a common practice in VI, Monte Carlo dropout (Kendall & Gal, 2017), and deep ensemble methods (Lakshminarayanan et al., 2017). Conditioned on $(X,\mathbf{Z}^N) = (x,\mathbf{z}^N)$, it is defined as

$$\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)] := \mathbb{E}_{q(\theta|\mathbf{z}^N)}|m_\theta(x) - \mathbb{E}_{q(\theta'|\mathbf{z}^N)}m_{\theta'}(x)|^2. \quad (6)$$

We can see that Eqs. (5) and (6) can be computed without the knowledge of $\nu(Y|x)$ and θ^*, unlike the excess risks in Eqs. (2) and (4).

Although Eqs. (5) and (6) are widely used in practice, their theoretical understanding remains limited as discussed in Section 1. Since these quantities are used as EU metrics, it is important to analyze the convergence properties. Furthermore, their relationships with the excess risks defined in Eqs. (2) and (4) need to be clarified, since the excess risks formalize the naive intuition about the lack of data.

However, the posterior distribution q appearing in Eqs. (5) and (6) is not necessarily restricted to Bayesian posterior distributions, making it difficult to analyze them using the standard Bayesian inference theory. Additionally, since the expectation is not taken with respect to $\nu(Y|x)$, applying PAC-Bayesian theory to these metrics is challenging. Our following analyses address these issues, providing to a novel understanding of Eqs. (5) and (6).

3 EPISTEMIC UNCERTAINTY AND EXCESS RISK

In this section, we first introduce a new excess risk and its relationship with the widely used EU metrics in Section 3.1. Then, we establish their connection to the exact excess risks and derive the convergence theory for these metrics in Section 3.2. All proofs are shown in Appendixes D to F.

3.1 Bayesian Excess Risk and Epistemic Uncertainty

First, we define the new excess risk and its novel interpretation as the widely used EU metrics. To do so, we introduce the following joint distribution:

$$p^q(\theta, \mathbf{Z}^N, Z) := \nu(\mathbf{Z}^N) q(\theta|\mathbf{Z}^N)\nu(X)p(Y|X,\theta). \quad (7)$$

In contrast to the joint distribution $\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(Z)$ used in the PAC-Bayesian analysis, e.g., in Eqs. (2) and (4), the test data point of Eq. (7) follows the posterior predictive distribution. Under this setting, we define the new excess risk as follows;

Definition 1. *Given loss ℓ and joint distribution Eq. (7), we define* **Bayesian excess risk (BER)** *as*

$$\begin{aligned}\mathrm{BER}^l(Y|X,\mathbf{Z}^N) &:= \mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y, \mathbb{E}_{q(\theta'|\mathbf{Z}^N)} f_{\theta'}(X)) \\ &- \inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}} \mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X)), \quad (8)\end{aligned}$$

where the decision rule $\phi : \Theta \times \mathcal{X} \to \mathcal{A}$ takes the parameter θ and the test input x and the infimum is taken over all decision rules such that the above expectation is defined.

Note that $\mathrm{BER}^\ell(Y|X,\mathbf{Z}^N)$ is always larger than 0; see Appendix D.3 for the proof. We also formally define $\mathrm{BER}^\ell(Y|X = x, \mathbf{Z}^N = \mathbf{z}^N)$, the conditional version of BER in Appendix D.2. As shown in Theorem 1 below, the widely used EU metrics in Eqs. (5) and (6) are equivalent to this conditional BER. Before providing that result, we first discuss the intuition of BER. The BER is similar to the PER in Eq. (4). The key difference lies in how the test data is generated. In BER, the test data is generated by our model $p(Y|x,\theta)$ under the posterior distribution $q(\theta|\mathbf{z}^N)$, rather than from $\nu(Y|X = x)$. This simply changes the distribution of the expectation in the first term.

The second term, $\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X))$, represents the **Bayes risk** under the joint distribution $p^q(\theta,\mathbf{Z}^N,Z)$. Intuitively, the decision rule ϕ takes the parameter θ. Since the test data is generated from $p(y|x,\theta)$, the decision rule ϕ has access to the "true" parameter of the data distribution under this setting, resulting in the minimum achievable risk. This type of Bayes risk often arises in the analysis of the Bayesian decision theory (Xu & Raginsky, 2022). The following lemma characterizes the Bayes risk under the joint distribution $p^q(\theta,\mathbf{Z}^N,Z)$:

Lemma 1. *Conditioned on $(x, \mathbf{z}^N)$, we have*

$$\begin{aligned}&\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}} \mathbb{E}_{p^q(\theta,\mathbf{z}^N,(x,Y))}\ell(Y,\phi(\theta,x)) \\ &= \begin{cases}\mathbb{E}_{q(\theta|\mathbf{z}^N)}H[p(Y|x,\theta)], & \text{for the log loss} \\ \mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}|Y-m_\theta(x)|^2. & \text{for the squared loss}\end{cases}\end{aligned}$$

In conclusion, BER implies the loss due to insufficient data under the joint distribution $p^q(\theta,\mathbf{Z}^N,Z)$. The following theorem further elaborates on the relationship between BER and widely used EU metrics:

Theorem 1. *Conditioned on $(x,\mathbf{z}^N)$, we express BER for the log loss as $\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)$ and for the squared loss as $\mathrm{BER}^{(2)}(Y|x,\mathbf{z}^N)$. Then, we have*

$$\begin{aligned}\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) &= I_q(\theta;Y|x,\mathbf{z}^N), \\ \mathrm{BER}^{(2)}(Y|x,\mathbf{z}^N) &= \mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)].\end{aligned}$$

Thus, we can see that the widely used EU metrics in Eqs. (5) and (6) can be interpreted as the excess risk under the joint distribution $p^q(\theta,\mathbf{Z}^N,Z)$ and their behavior can be analyzed by studying BER.

Remark 1. BER optimistically evaluates the excess risk, which assumes that our model is correct *because the test data follows the posterior predictive distribution. From a different perspective, the widely used EU metrics are the excess risk assuming that the approximation of $\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(Z) \approx p^q(\theta,\mathbf{Z}^N,Z)$.*

When a model is well-specified, $\nu(y|x) = p(y|x,\theta^*)$ holds for $\theta^* \in \Theta$, the predictive distribution converges to $p(y|x,\theta^*)$ and the approximation of $\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(Z) \approx p^q(\theta,\mathbf{Z}^N,Z)$ becomes more accurate as N increases (Alquier, 2021). We discuss the quality of this approximation in Section 3.2.

We also remark that the BER resembles **Fisher information** (FI) in the sense that FI also measures the geometry of the loss function under the expectation by $p(y|x,\theta)$ not by ν. This connection is interesting since the FI matrix behaves similarly to the Hessian matrix (Thomas et al., 2020) under certain conditions, which is related to the uncertainty under Laplace approximation (Bishop, 2006).

3.2 Relationships between EU Metrics and Excess Risk

On the basis of Section 3.1, we develop a novel relationship between the BER and the excess risks.

3.2.1 In the Case of the Squared Loss

First, we show the results for the squared loss assuming that $\mathcal{Y} = \mathbb{R}$. See Appendix D.6 for $\mathcal{Y} = \mathbb{R}^d$.

Theorem 2. *Conditioned on $(x, \mathbf{z}^N)$, assume that a regression function is well-specified, that is, $\mathbb{E}_{\nu(Y|x)}[Y|x] = m_{\theta^*}(x)$ holds. Then, we have*

$$\mathrm{PER}^{(2)}(Y|x, \mathbf{z}^N) + \mathrm{BER}^{(2)}(Y|x, \mathbf{z}^N) \\ = \mathrm{ER}^{(2)}(Y|x, \mathbf{z}^N, \theta^*) \leq R_q^{(2)}(Y|x, \mathbf{z}^N). \tag{9}$$

Remark 2. *When we use a flexible model, such as a deep neural network, the assumption $\mathbb{E}_{\nu(Y|x)}[Y|x] = m_{\theta^*}(x)$ holds even when $\nu(Y|x) \neq p(Y|x, \theta^*)$, which means we misspecify the noise function in regression tasks. Many Bayesian neural networks can also satisfy this assumption because of their universal approximation abilities (Foong et al., 2020). See Appendix D.5 for details.*

Taking the expectation over $(x, \mathbf{z}^N)$ and assuming that the PAC-Bayesian bound Eq. (2) holds, from Eq. (9) we have

$$\mathrm{PER}^{(2)}(Y|X, \mathbf{Z}^N) + \mathrm{BER}^{(2)}(Y|X, \mathbf{Z}^N) \\ = \mathrm{ER}^{(2)}(Y|X, \mathbf{Z}^N, \theta^*) = \mathcal{O}\left(\sqrt{\log N/N}\right). \tag{10}$$

See Appendix C for the order discussion of Eq. (2). From Eq. (9), $\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)](= \mathrm{BER}^{(2)}(Y|x, \mathbf{z}^N))$ is a lower bound of the exact excess risk since PER is always larger than 0. This validates the naive intuition that the BER **optimistically** evaluates the exact excess risk assuming that our model is correct, which had been discussed in Remark 1. Another important implication is that, from Eq. (10), $\mathrm{BER}^{(2)}(Y|X, \mathbf{Z}^N)$ converges to 0 with the same order as the generalization error bound. Thus, Eq. (10) shows that $\mathbb{E}_{\nu(X)\nu(\mathbf{Z}^N)}\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)](= \mathrm{BER}^{(2)}(Y|X, \mathbf{Z}^N))$ decreases as we increase N, which is the promising property as the EU metric.

3.2.2 In the Case of the Log Loss

We start from the general result that never requires the well-specification of the model;

Theorem 3. *Assume that a model $-\log p(y|x, \theta)$ is σ^2 sub-Gaussian for any pairs of (x, θ). Conditioned on $(x, \mathbf{z}^N)$, we have*

$$\mathrm{BER}^{\log}(Y|x, \mathbf{z}^N) \leq 2\sqrt{\sigma^2 \mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathrm{KL}(\nu(Y|x)|p(Y|x, \theta))}$$

$$= 2\sqrt{\sigma^2(\mathrm{KL}(\nu(Y|x)|p(Y|x, \theta^*)) + \mathrm{ER}^{\log}(Y|x, \mathbf{z}^N, \theta^*))}. \tag{11}$$

Thus, we can see that BER is the lower bound of the square root of the KL divergence between the data distribution $\nu(y|x)$ and our model $p(y|x, \theta)$. Furthermore, it is decomposed into the excess risk and the model misspecification error $\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta^*))$. This formulates the intuition that the BER is the lower bound of the risk since it optimistically evaluates the incurred risks as discussed in Remark 1.

If we assume that $\nu(y|x) = p(y|x, \theta^*)$ holds, then we can relate the PER, BER, and ER as follows. This is a stronger assumption than that in Theorem 2, since it requires complete knowledge about noise.

Corollary 1. *Assume that $\nu(y|x) = p(y|x, \theta^*)$ holds and under the same setting as Theorem 3, we have*

$$\mathrm{PER}^{\log}(Y|x, \mathbf{z}^N) + \mathrm{BER}^{\log}(Y|x, \mathbf{z}^N) \\ \leq 2\sqrt{\sigma^2 \mathrm{ER}^{\log}(Y|x, \mathbf{z}^N, \theta^*)}. \tag{12}$$

Note that this result is stronger than the result obtained by dropping the model misspecification error term in Theorem 3. We point out that this result is similar to Theorem 2 of the squared loss. From Corollary 1, the square of BER is the lower bound of the exact excess risk for the log loss. Thus, we can confirm that even for the log loss, the widely used EU metric $I_q(\theta; Y|x, \mathbf{z}^N)(= \mathrm{BER}^{\log}(Y|x, \mathbf{z}^N))$ underestimates the exact excess risk.

We remark the assumption about the sub-Gaussian property. Since this assumption is for the model $p(y|x, \theta)$, it can be more easily verified. For classification, we can easily enforce the sub-Gaussianity assumption for our model by setting the output class probabilities to be always at least $\exp(-L)$, where ℓ is a positive constant. See Appendix D.11 for details.

When focusing on the convergence of the BER, from Eq. (12), taking the expectation over $(x, \mathbf{z}^N)$, we have

$$\mathrm{PER}^{\log}(Y|X, \mathbf{Z}^N) + \mathrm{BER}^{\log}(Y|X, \mathbf{Z}^N) \\ \leq 2\sqrt{\sigma^2 \mathrm{ER}^{\log}(Y|X, \mathbf{Z}^N, \theta^*)}.$$

Then if the PAC-Bayesian bound Eq. (2) holds, the right-hand side can be bounded using Eq. (2). Thus, similarly to the squared loss, $I_q(\theta; Y|X, \mathbf{Z}^N)$ decreases as we increase N, which is the desirable property as the EU metric. We can also derive the convergence of the entropy of the predictive distribution, which shows $H[p^q(Y|X, \mathbf{Z}^N)] = H[p(Y|X, \theta^*)] + \mathcal{O}(\sqrt{\ln N/N^{1/2}})$ (see Appendix D.10 for a formal statement). To the best of our knowledge, this is the first time that the convergence of EU metrics and the entropy under approximation are presented.

In summary, we obtained an important relationship between the widely used EU metrics and the exact excess risk. When a model is sufficiently flexible to satisfy the assumptions in Theorems 2, e.g., Bayesian deep learning models (Foong et al., 2020; Lu & Lu, 2020), BER serves as a lower bound for the exact excess risk for a squared loss. For a log loss, the square of BER, scaled by the sub-Gaussian constant, serves as a lower bound for the exact excess risk. This validates the interpretation that the widely used EU metrics optimistically evaluate the excess risk as discussed in Remark 1.

3.3 Relationships between EU Metrics and Objective Functions in VI

Here, we show the relationships between BER and the objective functions in VI using the theories developed in Section 3.2. It has been numerically shown that standard VI, which minimizes the KL divergence between the exact and approximate posterior distributions, often underestimates EU (Hernandez-Lobato et al., 2016; Li & Gal, 2017; Minka et al., 2005). Note that in standard VI, the loss takes the form $\mathbb{E}_{q(\theta|\mathbf{z}^N)}\ell(y, f_\theta(x))$ as in Eq. (1). To address this issue, alternative objective functions have been proposed. For example, the entropic loss $\mathrm{Ent}_\alpha^\ell(y,x) := -\frac{1}{\alpha}\ln \mathbb{E}_{q(\theta|\mathbf{z}^N)}e^{-\alpha\ell(y,f_\theta(x))}, \alpha > 0$ was proposed on the basis of the α-divergence minimization. It has been used in α-divergence dropout (α-DO) (Li & Gal, 2017), and the second-order PAC-Bayesian methods (2^{nd}-PAC) (Masegosa, 2020; Futami et al., 2021, 2022).

It has been reported that the entropic risk can capture EU better than the standard VI thanks to the mode-covering property of α-divergence, compared with the mode-seeking property of the KL divergence used in standard VI (Hernandez-Lobato et al., 2016). However, such α-divergence is implemented in the parameter space; thus, it is unclear whether the parameter space property leads to a better EU property in the prediction space under the regime of over-parametrized models such as Bayesian deep models.

Here, we provide an alternative explanation that the entropic risk has a regularization effect on EU through BER different from that of standard VI. The key observation is that the objective function of the standard VI corresponds to the log loss $R_q^{\log}$, which appears in $\mathrm{ER}^{\log}$. On the other hand, the entropic risk corresponds to $\mathrm{Ent}_{\alpha=1}^\ell(y,x) = \mathrm{PR}_q^{\log}$, appearing in $\mathrm{PER}^{\log}$. According to Eq. (11), minimizing the objective of standard VI implies the minimization of ER, which leads to the minimization of the sum of PER and BER. On the other hand, minimizing the entropic risk simply results in minimizing PER. This illustrates the fundamental difference in the regularization effect of BER between the standard VI and the entropic risk.

This relations can be further clarified as follows; From Eq. (12), by Cauchy-Schwartz inequality, we have

$$\mathrm{PR}_q^{\log}(Y|x,\mathbf{z}^N) \le R_q^{\log}(Y|x,\mathbf{z}^N) - \mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) - R^{\log}(Y|X,\theta^*) + \sigma^2. \tag{13}$$

Since the objective of the standard VI is $R_q^{\log}$, we can see that the entropic risk (corresponds to the left-hand side in the above) has a smaller regularization effect on the EU of the posterior predictive distribution through $-\mathrm{BER}^{\log}(Y|X,\mathbf{z}^N)$. This explains why VI on the basis of the entropic risk, such as α-DO and 2^{nd}-PAC, gives a larger EU than the standard VI.

From these relationships and considering the numerical success of α-DO and 2^{nd}-PAC in practice, balancing BER and loss function appropriately might lead to a solution that improves the evaluation of EU. To validate this conjecture, we consider the following objective function, which directly controls the loss minimization and EU evaluation on the basis of the decomposition in Eq. (13); given the data point (x,y), the objective is

$$R_q^{\log}(y|x,\mathbf{z}^N) - (1-\lambda)\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N),$$

where $0 < \lambda \le 1$ controls the regularization on BER. However, directly using this objective function requires the Monte Carlo approximation from the posterior distribution, which results in a large bias and variance. Therefore, when focusing on the Gaussian likelihood $N(y|m_\theta(x), v^2)$, we use the following upper bound of this objective function

$$\begin{aligned}&R_q^{\log}(y|x,\mathbf{z}^N) - (1-\lambda)\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)\\&\le \frac{(y-\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x))^2}{2v^2} + \frac{\ln 2\pi v^2}{2} + \lambda\frac{\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)]}{2v^2}.\end{aligned} \tag{14}$$

See Appendix F.1 for the derivation. We refer to this upper bound as a **regularized Bayesian excess risk (rBER)** and to simplify the notation, we express it as a $\ell^{\mathrm{rBER}}(y,x,\lambda,q)$. We can implement this upper bound using Monte Carlo approximation from the posterior distribution, and the obtained estimator can be unbiased and does not exhibit large variance. We can derive the following PAC-Bayesian bound for this risk function.

Theorem 4. *For any prior distribution $p(\theta)$ over Θ independent of $\mathbf{Z}^N$ and for any posterior distribution $q(\theta|\mathbf{Z}^N)$, and for any $\xi \in (0,1)$ and $c > 0$, with probability at least $1-\xi$ over the realization of training data $\mathbf{Z}^N$, we have*

$$\begin{aligned}&\mathbb{E}_{\nu(Z)}\ell^{\mathrm{rBER}}(Y,X,\lambda,q(\theta|\mathbf{Z}^N))\\&\le \frac{1}{N}\sum_{i=1}^N \ell^{\mathrm{rBER}}(Y_i,X_i,\lambda,q(\theta|\mathbf{Z}^N))\\&+\frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta)) + \frac{1}{2}\ln\frac{1}{\xi} + \frac{1}{2}\Omega_{p,\nu}(c,N)}{cN},\end{aligned} \tag{15}$$

where

$$\begin{aligned}&\Omega_{p,\nu}(c,N) := \ln \mathbb{E}_{p(\theta)p(\theta')}\mathbb{E}_{\nu(\mathbf{Z}^N)}\times\\&\exp[cN(\mathbb{E}_{\nu(Z)}L(Z,\theta,\theta') - \frac{1}{N}\sum_{n=1}^N L(Z_n,\theta,\theta'))],\\&L(z,\theta,\theta') := \frac{(y-m_\theta(x))^2}{2v^2}\\&\qquad + (\lambda-1)\frac{m_\theta^2(x) - m_\theta(x)m_{\theta'}(x)}{2v^2}.\end{aligned}$$

See Appendix F.2 for the proof. Then, our final objective function is the upper bound of this bound:

$$\mathrm{rBER}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - \mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x_i))^2}{2v^2} + \frac{\ln 2\pi v^2}{2}$$
$$+\lambda\frac{\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x_i)]}{2v^2} + \frac{\mathrm{KL}(q(\theta|\mathbf{z}^N)|p(\theta))}{N}. \qquad (16)$$

We call Eq. (16) the regularized Bayesian excess risk VI (rBER). This is an extension of standard VI. In Section 5.2, we numerically evaluate this objective function.

We should set λ to be smaller than 1 since $\lambda = 1$ corresponds to the standard VI and the standard VI often underestimates EU. See Appendix F for additional discussion on the relation to the alpha divergence minimization.

4 DISCUSSION WITH EXISTING WORK

This study puts its basis on recent analyses of the EU using Bayesian decision theory (Xu & Raginsky, 2022). The key assumption of their analysis is the correctness of the model $p(y|x,\theta)$ and evaluate its average performance over the prior distribution. While this approach effectively defines the EU, it relies on strong assumptions, making it difficult to use for real-world applications. To address this limitation, we introduced Bayesian Excess Risk (BER) in Section 3.1.

Existing analyses of uncertainty have primarily focused on calibration, clarifying when a model overestimates or underestimates uncertainty (Nixon et al., 2019; Bai et al., 2021; Naeini et al., 2015; Guo et al., 2017). In this work, we did not focus on calibration properties because considering EU alone is insufficient for addressing calibration. To properly analyze the calibration, both AU and EU must be studied together, which is a limitation of our approach.

In addition to calibration, the analysis of Gaussian processes has gained attention due to the analytic expression of their posterior predictive distributions (Fiedler et al., 2021; Lederer et al., 2019). Some studies have focused on the geometric distance between test and training data points to evaluate EU (Liu et al., 2020; Tian et al., 2021), while others have related the randomness of posterior distributions to predictions using the delta method (Nilsen et al., 2022). In contrast, the information-theoretic approach (Xu & Raginsky, 2022) focused on the loss function of the problem, defining excess risk as EU. Loss-function-based analysis has also been proposed for deterministic learning algorithms (Jain et al., 2021). Our theory, which extends the information-theoretic approach (Xu & Raginsky, 2022) to approximate Bayesian inference, clarifies the convergence properties of the variance and entropy of posterior predictive distributions.

Although the excess risk bounds in Eqs. (2) and (4) have been analyzed using PAC-Bayesian theory (Alquier, 2021; Sheth & Khardon, 2017), their connection to epistemic uncertainty (EU) has not been explored. The relationship between PAC-Bayesian theory and Bayesian inference has been studied primarily in the context of marginal likelihood (Germain et al., 2016; Rothfuss et al., 2021). Our work establishes new connections that link the uncertainty in Bayesian inference with PAC-Bayesian theory. The information-theoretic approach (Xu & Raginsky, 2022) demonstrated that EU can be expressed through conditional mutual information, and this relationship has been extended to meta-learning in some studies (Jose et al., 2021; Futami & Iwata, 2024). However, these studies assume the availability of correct models and exact posterior distributions. Our proposed analysis relaxes these assumptions, making it more applicable to practical settings.

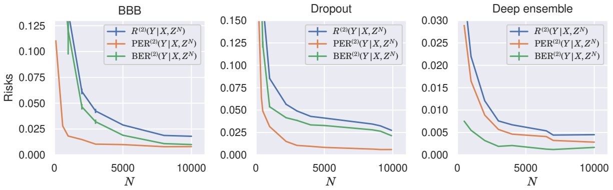

Figure 1: Results of toy data experiments: N represents the number of training data points, and the vertical line shows the value of each excess risk. The enlarged figures are shown in Appendix G.

5 NUMERICAL EXPERIMENTS

Here, we numerically evaluated theoretical findings and conjecture in Section 3. The detailed experimental settings and additional results are shown in Appendix G.

5.1 Numerical Evaluation of Theorem 2

We numerically evaluated $\mathrm{PER}^{(2)}(Y|X,\mathbf{Z}^N)$, $\mathrm{BER}^{(2)}(Y|X,\mathbf{Z}^N)$, and $R_q^{(2)}(Y|X,\mathbf{Z}^N)$ on toy datasets, where the true model is $y = 0.5x^3 + \epsilon$, $\epsilon \sim N(0,1)$, and $x \sim N(0,1)$. We used a 4-layer Bayesian neural network (BNN) for $m_\theta(x)$ with ReLU activation. The number of hidden units in each layer is 50. We assumed that this model satisfies $\mathbb{E}_\nu[Y|x] = f_{\theta^*}(x)$. We approximated its posterior distribution by Bayes by backpropagation (BBB) (Hernández-Lobato & Adams, 2015), dropout (Kendall & Gal, 2017), and deep ensemble (Lakshminarayanan et al., 2017), which are popular VI methods. The results are shown in Fig. 1. We can see that $\mathrm{PER}^{(2)}(Y|X,\mathbf{Z}^N)$ and $\mathrm{BER}^{(2)}(Y|X,\mathbf{Z}^N)$ are upper-bounded by $R_q^{(2)}(Y|X,\mathbf{Z}^N)$ and converge to zero as the number of samples increases. This is consistent with Eq. (10) in Theorem 2. We calculated the Spearman Rank Correlation (SRC) among $\mathrm{PER}^{(2)}(Y|X,\mathbf{Z}^N)$, $\mathrm{BER}^{(2)}(Y|X,\mathbf{Z}^N)$, and $R_q^{(2)}(Y|X,\mathbf{Z}^N)$ and the SRC scores were at least 0.97 suggesting high correlation among them.

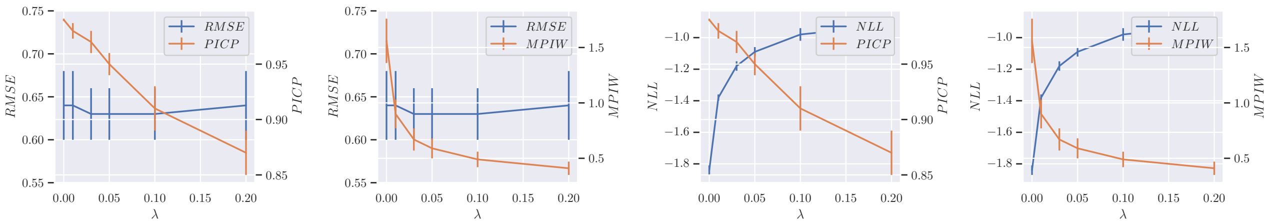

Figure 2: Performance characteristics of rBER under different λ values in regression task of Wine data in UCI dataset: In the two left figures, we plot RMSE, MPIW(0.95), and PICP. In the two right figures, we plot NLL, MPIW(0.95), and PICP. The enlarged figures are shown in Appendix G.

Table 1: Benchmark results for test RMSE, PICP, and MPIW.

Dataset	Avg. Test RMSE				Avg. Test PICP and MPIW in parenthesis			
	f-SVGD	VAR	rBER(0)	rBER(0.05)	f-SVGD	VAR	rBER(0)	rBER(0.05)
Concrete	4.33±0.8	4.30±0.7	4.47±0.6	4.48±0.7	0.82±0.03 (0.13±0.00)	0.87±0.04 (0.16±0.01)	0.99±0.02 (0.50±0.04)	**0.95±0.02** (0.25±0.02)
Boston	2.54±0.50	2.53±0.50	2.53±0.50	2.53±0.51	0.63±0.07 (0.10±0.02)	0.76±0.05 (0.14±0.01)	0.97±0.01 (0.33±0.04)	**0.92±0.04**(0.22±0.02)
Wine	0.61±0.04	0.61±0.04	0.64±0.04	0.63±0.02	0.79±0.03 (0.32±0.05)	0.85±0.02 (0.39±0.06)	0.99±0.00 (1.61±0.00)	**0.95±0.03** (0.32±0.15)
Power	3.78±0.14	3.75±0.13	3.66±0.15	3.69±0.12	0.43±0.01 (0.07±0.00)	0.82±0.01 (0.15±0.00)	0.99±0.01 (0.81±0.01)	**0.96±0.01** (0.37±0.01)
Yacht	0.64±0.28	0.60±0.28	0.75±0.41	0.78±0.48	0.92±0.04 (0.02±0.01)	0.93±0.04 (0.04±0.01)	**0.96±0.03** (0.10±0.01)	**0.94±0.04** (0.08±0.01)
Protein	3.98±0.54	3.92±0.05	3.83±0.10	3.85±0.05	0.53±0.01 (0.24±0.01)	0.83±0.00 (0.58±0.01)	1.0 ±0.00 (5.04±0.01)	**0.96±0.01** (0.86±0.00)

Table 2: Cumulative regret relative to uniform sampling.

Dataset	MAP	$\mathrm{PAC}^2_{\mathrm{E}}$	f-SVGD	VAR	rBER(0)	rBER(0.01)	rBER(0.05)
Mushroom	0.129±0.098	0.037±0.012	0.043±0.009	0.029±0.010	0.075±0.005	**0.024±0.009**	**0.021±0.004**
Financial	0.791±0.219	0.189±0.025	0.154±0.017	0.155±0.024	0.351±0.030	**0.075±0.024**	**0.075±0.031**
Statlog	0.675 ±0.287	0.032±0.003	0.010±0.000	0.006±0.000	0.145±0.223	0.005±0.001	**0.005±0.000**
CoverType	0.610±0.051	0.396±0.006	0.372±0.007	**0.291±0.004**	0.610±0.051	0.351±0.003	**0.290±0.002**

5.2 Real Data Experiments of Our Objective Function

We numerically evaluated the prediction and EU estimation performance of our rBER using Eq. (16) in regression and contextual bandit tasks. The goal of these experiments is to validate our conjecture that balancing loss minimization and BER regularization leads to improved solutions for EU evaluation.

Owing to the success of the entropic risk in particle VI (PVI) (Masegosa, 2020; Futami et al., 2021), which approximates the posterior distribution by an ensemble of models, we applied our rBER to the PVI setting. The posterior distribution is expressed as $q(\theta) := \frac{1}{M}\sum_{i=1}^{M}\delta_{\theta_i}(\theta)$, where $\delta_{\theta_i}(\theta)$ is the Dirac distribution that has a mass at θ_i and M is the size of the ensemble (see Appendix G for details about PVI). We can implement rBER by using this empirical distribution. For example, the variance term of rBER was calculated from this empirical distribution of the ensemble. We refer to rBER as rBER(0) when $\lambda = 0$ in Eq. (16).

First, we evaluated the impact of λ on the final prediction and EU evaluation performance of rBER using the UCI dataset (Dheeru & Karra Taniskidou, 2017) for regression tasks. We used a single-layer network with ReLU activation and approximated the posterior distribution with 20 ensembles. As prediction performance metrics, we evaluated RMSE and NLL. For the uncertainty evaluation performance, we evaluated the prediction interval coverage probability (PICP) (Tagasovska & Lopez-Paz, 2019), which shows the number of test observations within the estimated prediction interval; we set it to 0.95. PICP is optimal when it is close to 0.95. We also evaluated the mean prediction interval width (MPIW), which shows the average width of a prediction interval. The smaller MPIW is, the better the uncertainty estimate is when PICP is close to 0.95. From the formulation of rBER in Eq. (16), we expect that a smaller λ means a higher prediction performance and a larger EU evaluation. The results of 20 repetitions are shown in Fig. 2. As we expected, we obtained smaller PICPs and MPIWs as we increased λ. Note that rBER(1.) corresponds to the standard VI and shows an excessively small uncertainty. Thus, the objective function of standard VI overly regularizes EU. We also found that RMSE does not change markedly under different λs, perhaps because we used neural network models, which are sufficiently expressive to achieve a small RMSE even under a large λ.

We conducted similar experiments on other datasets in the UCI dataset and compared our method with existing PVI methods, f-SVGD (Wang et al., 2019), $\mathrm{PAC}^2_{\mathrm{E}}$ (Masegosa, 2020), and VAR (Futami et al., 2021). The results of 20 repetitions are shown in Table 1. Owing to space limitations, the results of the $\mathrm{PAC}^2_{\mathrm{E}}$, other λs, and the negative log-likelihood are shown in Appendix G. The existing PVIs showed small PICPs and MPIWs, indicating that the existing methods underestimate uncertainty. rBER(0) showed large PICPs and MPIWs since the Bayesian excess risk is not regularized. rBER(0.05) showed a moderate MPIW with a higher PICP and almost identical prediction performance in RMSE. Thus, rBER successfully controlled the prediction and uncertainty evaluation performance.

From the regression experiments, we numerically found that rBER achieved high fitting performance and improved EU evaluation performance by setting a small λ. Owing to this success, we applied rBER to contextual bandit problems (Riquelme et al., 2018). In contextual bandit tasks, we need to balance the trade-off between exploitation and exploration to achieve a small cumulative regret. For that purpose, our algorithms must appropriately control the prediction and uncertainty evaluation performance. Thus, we expect that our rBER is suitable for these tasks. For our experiments, we used the Thompson sampling algorithm with BNN, which has two hidden layers. We used 20 ensembles for approximating the posterior distribution. The results of 10 repetitions are shown in Table 2. Since our approach outperformed other methods, it showed better prediction and uncertainty control than the existing methods. This implies that the objective function of rBER, which directly optimizes the sum of PER and BER, is better than the entropic loss, which implicitly optimizes BER, for achieving high prediction and uncertainty evaluation performance.

6 CONCLUSION AND LIMITATION

In this study, we analyzed widely used EU metrics in VI and provided novel relationships between these metrics and excess risks. We examined the objective function of VI based on these relationships. Finally, we proposed a new VI and applied it to PVI. A clear limitation of our work is that the developed theory applies only to squared loss and log loss, while other popular EU metrics, such as expected calibration error in classification and Fisher information in mean-field VI, remain unaddressed. In future work, we aim to explore the relationship between BER and these EU metrics, providing a rigorous theory to guarantee their behavior, which is crucial for practical applications.

Acknowledgements

FF was supported by JSPS KAKENHI Grant Number JP23K16948. FF was supported by JST, PRESTO Grant Number JPMJPR22C8, Japan.

References

Alquier, P. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

Alquier, P. and Ridgway, J. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475 – 1497, 2020a. doi: 10.1214/19-AOS1855. URL https://doi.org/10.1214/19-AOS1855.

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

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

Bai, Y., Mei, S., Wang, H., and Xiong, C. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In *International Conference on Machine Learning*, pp. 566–576. PMLR, 2021.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Brown, L. D. and Purves, R. Measurable selections of extrema. *The annals of statistics*, pp. 902–912, 1973.

Clarke, B. and Barron, A. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990. doi: 10.1109/18.54897.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Fiedler, C., Scherer, C. W., and Trimpe, S. Practical and rigorous uncertainty bounds for gaussian process regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7439–7447, 2021.

Foong, A., Burt, D., Li, Y., and Turner, R. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:15897–15908, 2020.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, 32, 2019.

Futami, F. and Iwata, T. Information-theoretic analysis of Bayesian test data sensitivity. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1099–1107. PMLR, 02–04 May 2024.

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

Futami, F., Iwata, T., Ueda, N., Sato, I., and Sugiyama, M. Predictive variational bayesian inference as risk-seeking optimization. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5051–5083. PMLR, 28–30 Mar 2022.

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

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Haussler, D. and Opper, M. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernandez-Lobato, D., and Turner, R. Black-box alpha divergence minimization. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1511–1520, New York, New York, USA, 20–22 Jun 2016. PMLR.

Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pp. 1861–1869. PMLR, 2015.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Jain, M., Lahlou, S., Nekoei, H., Butoi, V., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*, 2021.

Janz, D., Hron, J., Mazur, P., Hofmann, K., Hernández-Lobato, J. M., and Tschiatschek, S. Successor uncertainties: exploration and uncertainty in temporal difference learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Jose, S. T., Park, S., and Simeone, O. Information-theoretic analysis of epistemic uncertainty in bayesian meta-learning. *arXiv preprint arXiv:2106.00252*, 2021.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lederer, A., Umlauft, J., and Hirche, S. Posterior variance analysis of gaussian processes with application to average learning curves, 2019.

Li, Y. and Gal, Y. Dropout inference in bayesian neural networks with alpha-divergences. In *International conference on machine learning*, pp. 2052–2061. PMLR, 2017.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pp. 2378–2386, 2016.

Lu, Y. and Lu, J. A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems*, 33: 3094–3105, 2020.

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

Masegosa, A., Lorenzen, S., Igel, C., and Seldin, Y. Second order pac-bayesian bounds for the weighted majority vote. *Advances in Neural Information Processing Systems*, 33, 2020.

Minka, T. et al. Divergence measures and message passing. 2005.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Nguyen, A. T., Tran, T., Gal, Y., Torr, P., and Baydin, A. G. Kl guided domain adaptation. In *International Conference on Learning Representations*, 2021.

Nilsen, G. K., Munthe-Kaas, A. Z., Skaug, H. J., and Brun, M. Epistemic uncertainty quantification in deep learning classification by the delta method. *Neural Networks*, 145: 164–176, 2022.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR workshops*, 2019.

Ortega, L. A., Cabañas, R., and Masegosa, A. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11720–11743. PMLR, 2022.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9116–9126. PMLR, 18–24 Jul 2021.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Sheth, R. and Khardon, R. Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Sheth, R. and Khardon, R. Pseudo-bayesian learning via direct loss minimization with applications to sparse gaussian process models. In Zhang, C., Ruiz, F., Bui, T., Dieng, A. B., and Liang, D. (eds.), *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pp. 1–18. PMLR, 08 Dec 2020.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Sun, S., Zhang, G., Shi, J., and Grosse, R. FUNCTIONAL VARIATIONAL BAYESIAN NEURAL NETWORKS. In *International Conference on Learning Representations*, 2019.

Tagasovska, N. and Lopez-Paz, D. Single-model uncertainties for deep learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Thomas, V., Pedregosa, F., Merriënboer, B., Manzagol, P.-A., Bengio, Y., and Le Roux, N. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3503–3513. PMLR, 2020.

Tian, J., Yung, D., Hsu, Y.-C., and Kira, Z. A geometric perspective towards neural calibration via sensitivity decomposition. *Advances in Neural Information Processing Systems*, 34, 2021.

Wang, Z., Ren, T., Zhu, J., and Zhang, B. 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, 2019. URL https://openreview.net/forum?id=BkgtDsCcKQ.

Watanabe, S. *Algebraic geometry and statistical learning theory*. Cambridge university press, 2009.

Watanabe, S. *Mathematical theory of Bayesian statistics*. Chapman and Hall/CRC, 2018.

Wei, Y. and Khardon, R. Variational inference on the final-layer output of neural networks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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

Xu, A. and Raginsky, M. Minimum excess risk in bayesian learning. *IEEE Transactions on Information Theory*, 68 (12):7935–7955, 2022. doi: 10.1109/TIT.2022.3176056.

Ye, N. and Zhu, Z. Bayesian adversarial learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Sec.2 and 3.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, see Sec.5 and the Supplementary material.]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, see the Supplementary material.]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes, see Sec.2 and 3.]

   (b) Complete proofs of all theoretical results. [Yes, see the Supplementary material.]

   (c) Clear explanations of any assumptions. [Yes, see Sec.2 and 3.]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, see the Supplementary material.]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, see Sec. 5 and the Supplementary material.]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, see the Supplementary material.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes, see the Supplementary material.]

   (b) The license information of the assets, if applicable. [Yes, see the Supplementary material.]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Notation

Distributions

$\nu(z)$	A data generating distribution
$p(y\|x,\theta)$	A model
$p(\theta)$	A prior distribution
$q(\theta\|\mathbf{z}^N)$	A posterior distribution
$p^q(y\|x,\mathbf{z}^N)$	The predictive distribution obtained by the expectation over $q(\theta\|\mathbf{z}^N)$
$p^q(\theta,\mathbf{z}^N,z)$	The approximate joint distribution defined as $\nu(\mathbf{z}^N)q(\theta\|\mathbf{z}^N)\nu(x)p(y\|x,\theta)$

Risk functions

$R^\ell_q(Y\|X,\mathbf{Z}^N)$	A test error defined as $\mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{q(\theta\|\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\ell(Y,f_\theta(X))$
$r^\ell_q(\mathbf{Z}^N)$	A training error defined as $\mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{q(\theta\|\mathbf{Z}^N)}\sum_{n=1}^N\ell(Y_n,f_\theta(X_n))/N$
$\mathrm{ER}^\ell(Y\|X,\mathbf{Z}^N,\theta^*)$	The excess risk defined as $R^\ell_q(Y\|X,\mathbf{Z}^N)-R^\ell(Y\|X,\theta^*)$.
$\mathrm{PR}^\ell_q(Y\|X,\mathbf{Z}^N)$	A prediction risk defined as $\mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\ell(Y,\mathbb{E}_{q(\theta\|\mathbf{Z}^N)}f_\theta(X))$
$\mathrm{PER}^\ell(Y\|X,\mathbf{Z}^N)$	The prediction excess risk defined as $\mathrm{PR}^\ell_q(Y\|X,\mathbf{Z}^N)-R^\ell(Y\|X,\theta^*)$
$\mathrm{BPR}^\ell_q(Y\|X,\mathbf{Z}^N)$	The Bayesian prediction risk defined as $\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\mathbb{E}_{q(\theta'\|\mathbf{Z}^N)}f_{\theta'}(X))$
$\mathrm{BER}^\ell(Y\|X,\mathbf{Z}^N)$	The Bayesian excess risk as $\mathrm{BPR}^\ell(Y\|X,\mathbf{Z}^N)-\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X))$.

B Summary of settings

Here we summarize the concepts and definitions of joint distributions and risks used in this work.

B.1 The setting used in the PAC-Bayesian theory (Section 2.1)

- Joint distribution of data and parameter θ (All the data is i.i.d.):

$$\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(Z).$$

- Prediction excess risk:

$$\begin{aligned}\mathrm{PER}^\ell(Y|X,\mathbf{Z}^N)&:=\mathrm{PR}^\ell_q(Y|X,\mathbf{Z}^N)-R^\ell(Y|X,\theta^*),\\ \mathrm{PR}^\ell_q(Y|X,\mathbf{Z}^N)&:=\mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\ell(Y,\mathbb{E}_{q(\theta|\mathbf{Z}^N)}f_\theta(X)).\end{aligned}$$

B.2 Our setting defined in Section 3.1

- Joint distribution (The training data is i.i.d. The test data follows the model):

$$p^q(\theta, \mathbf{Z}^N, Z) := \nu(\mathbf{Z}^N) q(\theta|\mathbf{Z}^N)\nu(X)p(Y|X,\theta)$$

- Bayesian excess risk:

$$\mathrm{BER}^\ell(Y|X,\mathbf{Z}^N) := \mathrm{BPR}^\ell_q(Y|X,\mathbf{Z}^N) - \inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}} \mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X)),$$
$$\mathrm{BPR}^\ell_q(Y|X,\mathbf{Z}^N) := \mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\mathbb{E}_{q(\theta'|\mathbf{Z}^N)}f_{\theta'}(X)).$$

B.3 Notations for the general loss function and its relation to the squared and log loss

We define the loss function $l : \mathcal{Y}\times\mathcal{A}\to\mathbb{R}$ to measure the performance of supervised learning, where $\mathcal{A}$ is the action space. We define a hypothesis $f_\theta(x) : \mathcal{X}\to\mathcal{A}$ and given (y,x) and f, we represent the loss as $\ell(y, f_\theta(x))$.

We use a probability model $p(y|x,\theta)$ with its mean expressed as $m_\theta(x)\in\mathbb{R}$.

- Squared loss: $\ell(y,a) = |y-a|^2$, where $\mathcal{Y}=\mathcal{A}=\mathbb{R}$. $f_\theta(x) = m_\theta(x)$ and $\ell(y, m_\theta(x)) = |y-m_\theta(x)|^2$.
- Log loss: $\ell(y,p) = -\ln p(y)$, where p is the probability distribution of Y and $\mathcal{A}$ is the set of all distributions on Y. $f_\theta(x) = p(\cdot|x,\theta)$ and $\ell(y, p(\cdot|x,\theta)) = -\ln p(y|x,\theta)$

C Preliminaries of the PAC-Bayesian theory

We briefly introduce the PAC-Bayesian theory. The typical PAC-Bayesian bound provides us the high-probability guarantee about the gap between the test error $\tilde{R}(\theta) := \mathbb{E}_{\nu(Z)}\ell(Y, f_\theta(X))$ and the training dataset $\tilde{r}(\theta) := \frac{1}{N}\sum_{n=1}^N \ell(Y_n, f_\theta(X_n))$ (Here we do not take the expectation over $\mathbf{Z}^N$);

Theorem 5. *(Alquier et al., 2016) Given a data generating distribution ν, for any prior distribution $p(\theta)$ over Θ independent of $\mathbf{Z}^N$ and for any $\xi\in(0,1)$ and $c>0$, with probability at least $1-\xi$ over the choice of training data $\nu(\mathbf{Z}^N)$, for all probability distributions $q(\theta|\mathbf{z}^N)$ over Θ, we have*

$$\mathbb{E}_q\tilde{R}(\theta) \le \mathbb{E}_q\tilde{r}(\theta) + \frac{\mathrm{KL}(q|p) + \ln\xi^{-1} + \Omega_{p,\nu}(c,N)}{cN},$$

where $\Omega_{p,\nu}(c,N) := \ln\mathbb{E}_{p(\theta)}\mathbb{E}_{\nu(\mathbf{Z}^N)}\exp[cN(\tilde{R}(\theta)-\tilde{r}(\theta))]$.

This constant $\Psi_{p,\nu}$ depends on the property of the loss function and the data generating distribution and prior. For example, when $\ell(Y, f_\theta(x)) - \mathbb{E}_\nu\ell(Y, f_\theta(X))$ satisfies the σ^2 sub-Gaussian property, and by setting $c = 1/\sqrt{N}$, we have

$$\mathbb{E}_q\tilde{R}(\theta) \le \mathbb{E}_q\tilde{r}(\theta) + \frac{\mathrm{KL}(q|p) + \ln\xi^{-1} + \sigma^2/2}{\sqrt{N}}.$$

Note that we say that the real-valued random variable X satisfies the σ^2 sub-Gaussian property if for any $t\in\mathbb{R}$, the following relation holds

$$\log\mathbb{E}e^{tX} \le \frac{t^2\sigma^2}{2}.$$

On the other hand, we introduced the bound in expectation in the main paper (Alquier, 2021). For example when we assume the σ^2 sub-Gaussian property and $c=\lambda$, we have

$$R^\ell_q(Y|X,\mathbf{Z}^N) \le r^\ell_q(\mathbf{Z}^N) + \frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))}{\lambda} + \frac{\lambda\sigma^2}{2N}, \tag{17}$$

see Alquier (2021) and Alquier et al. (2016) for the proof and other settings of PAC-Bayesian bounds.

Next, we introduce the PAC-Bayesian bound Eq.(2), which considers the excess risk. As introduced in Alquier (2021), if the loss l satisfies the L-Lipschitz property, then we can convert the above PAC-Bayes bound Eq. (17) to the excess risk bound; from Theorem 4.1 in Alquier (2021), we have the oracle inequality

$$R^{\ell}_{\hat{q}}(Y|X,\mathbf{Z}^N) \leq \inf_{q\in\mathcal{P}(\Theta)} R^{\ell}_{q}(Y|X,\mathbf{Z}^N) + \frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))}{\lambda} + \frac{\lambda\sigma^2}{2N},$$

where $\hat{q}$ is the Gibbs distribution defined similar to Eq. (1);

$$\hat{q}(\theta|\mathbf{z}^N) = \operatorname*{argmin}_{q(\theta|\mathbf{z}^N)\in\mathcal{P}(\Theta)} r^{\ell}_{q}(\mathbf{z}^N) + \frac{\mathrm{KL}(q(\theta|\mathbf{z}^N)|p(\theta))}{\lambda}.$$

where $\mathcal{P}(\Theta)$ is a set of all probability distribution over θ.

Then we expand $R^{\ell}_{q}(Y|X,\mathbf{Z}^N)$ around the optimal parameter θ^* using the Lipschitz property;

$$R^{\ell}_{q}(Y|X,\mathbf{Z}^N) \leq R^{\ell}(Y|X,\theta^*) + L\mathbb{E}_{\mathbf{Z}^N}\mathbb{E}_q\|\theta-\theta^*\| = R^{\ell}(Y|X,\theta^*) + L\mathbb{E}_q\|\theta-\theta^*\|,$$

where the last expectation is taken over the prior distribution. The example 2.2 shown in Alquier (2021) is that setting the prior as $N(\theta|0,\gamma^2 I_d)$ and $\gamma=\sqrt{\sigma^2}/\sqrt{N}$, we obtain

$$R^{\ell}_{\hat{q}}(Y|X,\mathbf{Z}^N) \leq R^{\ell}(Y|X,\theta^*) + L\sqrt{\frac{d\sigma^2}{N}} + \frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))}{\lambda} + \frac{\lambda\sigma^2}{2N}. \tag{18}$$

By assuming that both the prior and posterior are Gaussian distributions, there exists a positive constant C such that

$$R^{\ell}_{\hat{q}}(Y|X,\mathbf{Z}^N) \leq R^{\ell}(Y|X,\theta^*) + \sqrt{\frac{C\log N}{N}}.$$

where C only depends on the problem, such as the Lipschitz constant of the loss, mean and variance of the prior and posterior distribution.

For the squared loss, the above Lipschitz bound cannot be used. In such a case, Alquier & Ridgway (2020b) introduced the modified Lipschitz continuity; we assume that for any $\theta,\ \theta'\in\Theta$, there exists a measurable function $M(Z)$ such that

$$\ell(Y,f_\theta(X)) - \ell(Y,f_{\theta'}(X)) \leq M(Z)\|\theta-\theta'\|,$$

and assume $\mathbb{E}_Z M(Z) < L < \infty$. Then we have

$$R^{\ell}_{q}(Y|X,\mathbf{Z}^N) \leq R^{\ell}(Y|X,\theta^*) + \mathbb{E}_{\mathbf{Z}^N}\mathbb{E}_{q(\theta|\mathbf{Z}^N)}\mathbb{E}_{Z'}M(Z')\|\theta-\theta^*\| = R^{\ell}(Y|X,\theta^*) + L\mathbb{E}_p\|\theta-\theta^*\|.$$

Then we can obtain a similar bound as Eq. (18).

Another widely used excess risk bound is the assumption under the Bernstein condition (Alquier & Ridgway, 2020a); if there exists a constant $K>0$ such that for any $\theta\in\Theta$,

$$\mathbb{E}_Z(\ell(Y,f_\theta(X)) - R^{\ell}(Y|X,\theta^*))^2 \leq K(R^{\ell}(Y|X,\theta) - R^{\ell}(Y|X,\theta^*)).$$

This is known as the condition that achieves faster convergence. This assumption is satisfied, for example, under the classification without noise, the loss with Lipschitz continuity and strong convexity (Alquier, 2021).

Then under the Bernstein condition with some constant $K>0$ and loss is bounded above by some constant C, from Theorem 4.3 in Alquier (2021), we have

$$\begin{aligned} &R^{\ell}_{q}(Y|X,\mathbf{Z}^N) - R^{\ell}(Y|X,\theta^*) \\ &\leq \inf_q \left(R^{\ell}_{q}(Y|X,\mathbf{Z}^N) - R^{\ell}(Y|X,\theta^*) + \frac{\max(2K,C)\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))}{N} \right). \end{aligned}$$

See also Alquier et al. (2016) for additional discussion about the general form of the excess risk bound under the Bernstein condition. We finally remark that the excess risk of the log loss and the squared loss and its relation to the PER has extensively been studied in Sheth & Khardon (2017).

D Proofs of theorems in Section 3

Here we present the proofs of Section 3. In this section, we use the definition $\mathrm{BPR}_q^\ell(Y|X,\mathbf{Z}^N) := \mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\mathbb{E}_{q(\theta'|\mathbf{Z}^N)}f_{\theta'}(X))$ to simplify the notation.

D.1 Remarks about swapping the expectation

Some of our proof requires the swapping expectations about $\nu(Y|x)$ and $q(\theta|\mathbf{z}^N)$. This is expressed as the following assumption:

Assumption 1.

$$\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{\nu(Y|x)}\ell(Y,f_\theta(x)) = \mathbb{E}_{\nu(Y|x)}\mathbb{E}_{q(\theta|\mathbf{z}^N)}\ell(Y,f_\theta(x)).$$

This means that conditions of Fubini's theorem must hold. This is a mild assumption as discussed in Sheth & Khardon (2017). For completeness, we cite their discussions here. When considering the log loss, the integrand can be negative, so we need to be careful about swapping the expectation. For discrete Y, such as classification tasks, $p(y|x,\theta)\leq 1$ implies the log loss is always positive, thus the condition is satisfied. For continuous Y, let us consider the Gaussian likelihood $p(y|x,\theta) = N(y|m_\theta(x),v^2)$. Then we have

$$\log p(y|x,\theta) \leq B := -\log\sqrt{2\pi} - \log v^2,$$

thus $-\log p(y|x,\theta) + B \geq 0$ holds. So by redefining the log loss by adding B, we can enforce the condition of Fubini's theorem.

Finally, existing work (Sheth & Khardon, 2017) suggested modifying the log loss as

$$-\log\left(\frac{1-c}{\max_{\theta,x,y} p(y|x,\theta)}p(y|x,\theta) + c\right),$$

by using a small constant c. Then we can enforce the condition of Fubini's theorem in general.

D.2 Conditional expectation of excess risks

In this section, we define the conditional version of the excess risks. Note that PER and BER are defined as

$$\begin{aligned}\mathrm{PER}^\ell(Y|X,\mathbf{Z}^N) &:= \mathrm{PR}_q^\ell(Y|X,\mathbf{Z}^N) - R^\ell(Y|X,\theta^*),\\ \mathrm{BER}^\ell(Y|X,\mathbf{Z}^N) &:= \mathrm{BPR}_q^\ell(Y|X,\mathbf{Z}^N) - \inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X)).\end{aligned}$$

We define the conditional excess risk as

$$\mathrm{PER}^\ell(Y|x,\mathbf{z}^N) := \mathrm{PR}_q^\ell(Y|x,\mathbf{z}^N) - R^\ell(Y|x,\theta^*), \tag{19}$$

$$\mathrm{BER}^\ell(Y|x,\mathbf{z}^N) := \mathrm{BPR}_q^\ell(Y|x,\mathbf{z}^N) - \inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}\ell(Y,\phi(\theta,x)). \tag{20}$$

It has been proved in Lemma 3.4 in Steinwart & Christmann (2008) that if the action space is $\mathbb{R}$ following relation holds

$$\mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\mathrm{PER}^\ell(Y|x,\mathbf{z}^N) = \mathrm{PER}^\ell(Y|X,\mathbf{Z}^N), \tag{21}$$

$$\mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\mathrm{BER}^\ell(Y|x,\mathbf{z}^N) = \mathrm{BER}^\ell(Y|X,\mathbf{Z}^N). \tag{22}$$

Moreover, from Theorem 3 in Brown & Purves (1973), the above relation holds for the log loss. Thus, we can naturally connect the conditional and unconditional definitions of PER and BER for squared loss and log loss.

Explicit calculation

In the following, we explicitly calculate how relations Eqs. (21) and (22) hold for the squared and log loss.

The first terms in (19) and (20) can easily be expressed as

$$\begin{aligned}\mathrm{PR}^{\ell}_{q}(Y|X,\mathbf{Z}^N) &= \mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{\nu(Z)}\ell(Y,\mathbb{E}_{q(\theta|\mathbf{Z}^N)}f_\theta(X)) \\ &= \mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\mathbb{E}_{\nu(Y|X=x)}\ell(Y,\mathbb{E}_{q(\theta|\mathbf{z}^N)}f_\theta(x)) \\ &= \mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\mathrm{PR}^{\ell}_{q}(Y|x,\mathbf{z}^N),\end{aligned}$$

and

$$\begin{aligned}\mathrm{BPR}^{\ell}_{q}(Y|X,\mathbf{Z}^N) &= \mathbb{E}_{\nu(\mathbf{Z}^N)}\mathbb{E}_{\nu(X)q(\theta|\mathbf{Z}^N)p(Y|X,\theta)}\ell(Y,\mathbb{E}_{q(\theta'|\mathbf{Z}^N)}f_{\theta'}(X)) \\ &= \mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\mathbb{E}_{q(\theta|\mathbf{Z}^N=\mathbf{z}^N)p(Y|X=x,\theta)}\ell(Y,\mathbb{E}_{q(\theta'|\mathbf{z}^N)}f_{\theta'}(x)) \\ &= \mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\mathrm{BPR}^{\ell}_{q}(Y|x,\mathbf{z}^N).\end{aligned}$$

The second term in (19) can easily be expressed as

$$\mathrm{R}^{\ell}(Y|X,\theta^*) = \mathbb{E}_{\nu(Z)}\ell(Y,f_{\theta^*}(X)) = \mathbb{E}_{\nu(X=x)}\mathbb{E}_{\nu(Y|X=x)}\ell(Y,f_{\theta^*}(x)) = \mathbb{E}_{\nu(X=x)}\mathrm{R}^{\ell}(Y|x,\theta^*),$$

Next, we calculate the Bayes risks in the second term of Eq. (20). First, for the squared loss, we have

$$\begin{aligned}\mathbb{E}_{\nu(Z)}[\ell(Y,\tilde{\phi}(X))] &= \mathbb{E}_{\nu(Z)}(Y-\tilde{\phi}(X))^2 \\ &= \mathbb{E}_{\nu(Z)}(Y-\mathbb{E}_{\nu(Y'|X)}[Y'|X])^2 + \mathbb{E}_{\nu(X)}(\mathbb{E}_{\nu(Y'|X)}[Y'|X]-\tilde{\phi}(X))^2,\end{aligned}$$

where $\mathbb{E}_{\nu(Y'|X)}[Y'|X]$ is the conditional expectation. Thus, infimum is achieved by setting $\mathbb{E}_{\nu(Y'|X)}[Y'|X]=\tilde{\phi}(X)$ and we have

$$\begin{aligned}\inf_{\tilde{\phi}:\mathcal{X}\to\mathcal{A}}\mathbb{E}_{\nu(Z)}[\ell(Y,\tilde{\phi}(X))] &= \mathbb{E}_{\nu(Z)}(Y-\mathbb{E}_{\nu(Y'|X)}[Y'|X])^2 \\ &= \mathbb{E}_{\nu(X=x)}\mathbb{E}_{\nu(Y|X=x)}(Y-\mathbb{E}_{\nu(Y'|X)}[Y'|X])^2 \\ &= \mathbb{E}_{\nu(X=x)}\inf_{\tilde{\phi}:\mathcal{X}\to\mathcal{A}}\mathbb{E}_{\nu(Y|x)}[\ell(Y,\tilde{\phi}(x))].\end{aligned}$$

We can show the same statement for BER as follows:

$$\begin{aligned}\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X)) &= \mathbb{E}_{\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X)}(Y-\mathbb{E}_{p(Y'|X,\theta)}[Y'|X])^2 \\ &= \mathbb{E}_{\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X=x)}(Y-\mathbb{E}_{p(Y'|X=x,\theta)}[Y'|x])^2 \\ &= \mathbb{E}_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}\ell(Y,\phi(\theta,x)),\end{aligned}$$

where we used the tower property of the conditional expectation

Next, we discuss the conditional Bayes risk for log loss. We use the relation, $-\mathbb{E}_p\ln p = \inf_{q>>p}\mathbb{E}_p[-\ln q]$, where $q>>p$ implies that p is absolutely continuous with respect to q. Then, by definition, it is clear that

$$\mathrm{BPR}^{\log}(Y|x,\mathbf{z}^N) = -\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\ln\mathbb{E}_{q(\theta'|\mathbf{z}^N)}p(Y|x,\theta') = H[p^q(Y|x,\mathbf{z}^N)],$$

where $p^q(Y|x,\mathbf{z}^N)=\mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta)$ is the conditional predictive distribution. Also, by definition, the Bayes risk for the log loss is given as

$$\begin{aligned}\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X)) &= \inf_{p'}\mathbb{E}_{\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X)p(Y|X,\theta)}[-\ln p'(Y|X,\theta)] \\ &= -\mathbb{E}_{\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X)p(Y|X,\theta)}\,[\ln p(Y|X,\theta)] \\ &= E_{\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X=x)p(Y|X=x,\theta)}H[p(Y|X=x,\theta)] \\ &= E_{\nu(\mathbf{Z}^N=\mathbf{z}^N)\nu(X=x)}\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}E_{q(\theta|\mathbf{Z}^N)p(Y|x,\theta)}\ell(Y,\phi(\theta,x)).\end{aligned}$$

Thus, it is clear that the relations shown in Eqs. (21) and (22) hold for the squared and log losses.

D.3 Proof of $\mathrm{BER}^{\ell}(Y|X,\mathbf{Z}^N)\geq 0$

We remark that for the squared loss and log loss, the first term of BER can be written as

$$\mathbb{E}_{p^{\mathrm{f}}(\theta,\mathbf{Z}^N,Z)}\ell(Y,\mathbb{E}_{q(\theta'|\mathbf{Z}^N)}f_{\theta'}(X))=\inf_{\psi:\mathcal{Z}^N\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\mathbf{Z}^N,Z,\theta)}[\ell(Y,\psi(X,\mathbf{Z}^N))].$$

This expression is similar to the definition of the first term in MER, which is introduced as the special type of excess risk in Xu & Raginsky (2022). Then, applying the same technique in Lemma 1 (Xu & Raginsky, 2022), we can show that $\mathrm{BPR}_q^{\ell}(Y|X,\mathbf{Z}^N)$ satisfies the data processing inequality. Then given a Markov chain, for example, $(X,\mathbf{Z}^N)-(X,Z^{N+1})-Y$, then we have $\mathrm{BPR}_q^{\ell}(Y|X,\mathbf{Z}^N)\geq\mathrm{BPR}_q^{\ell}(Y|X,Z^{N+1})$. Consider a Markov chain $(X,\mathbf{Z}^N)-(X,\theta)-Y$. Then by the data processing inequality, we have $\mathrm{BPR}_q^{\ell}(Y|X,\mathbf{Z}^N)\geq\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{p^q(\theta,\mathbf{Z}^N,Z)}\ell(Y,\phi(\theta,X))$ since the Bayes error uses the parameter of $p^q(\mathbf{z}^N,\theta,z)$ directly. This concludes the proof.

D.4 Proof of Theorem 1

Here we consider the conditional quantities of them. The formal definitions of the conditional fundamental limit of learning and total risk are given in Appendix D.2.

For the log loss, we use the property of entropy. For any probability distributions p and q, the entropy satisfies $H[p]:=-\mathbb{E}_p\ln p=\inf_{q>>p}\mathbb{E}_p[-\ln q]$, where $q>>p$ implies that p is absolutely continuous with respect to q . Then, by definition, it is clear that

$$\mathrm{BPR}^{\log}(Y|x,\mathbf{z}^N)=-\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\ln\mathbb{E}_{q(\theta'|\mathbf{z}^N)}p(Y|x,\theta')=H[p^q(Y|x,\mathbf{z}^N)],$$

where $p^q(Y|x,\mathbf{z}^N)=\mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta)$ is the conditional predictive distribution. Also, by definition, the Bayes risk for the log loss is given as

$$\begin{aligned}\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}\ell(Y,\phi(\theta,x))&=\inf_{p'}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}[-\ln p'(Y|x,\theta)]\\&=-\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\left[\ln p(Y|x,\theta)\right]\\&=E_{q(\theta|\mathbf{z}^N)}H[p(Y|x,\theta)].\end{aligned}$$

Thus,

$$\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)=\mathrm{BPR}^{\log}(Y|x,\mathbf{z}^N)-\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}\ell(Y,\phi(\theta,x))=I_q(\theta;Y|x,\mathbf{z}^N).$$

Next, for the squared loss, recall that for any random variable Y with distribution p, we have

$$\inf_a\mathbb{E}_p|Y-a|^2=\mathbb{E}_p|Y-\mathbb{E}_{p(Y')}Y'|^2=\mathrm{Var}(Y).$$

Using this relation, we have

$$\begin{aligned}&\mathrm{BER}_q^{(2)}(Y|x,\mathbf{z}^N)\\&=\mathrm{BPR}^{(2)}(Y|x,\mathbf{z}^N)-\inf_{\phi:\Theta\times\mathcal{X}\to\mathcal{A}}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}\ell(Y,\phi(\theta,x))\\&=\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}(Y-\mathbb{E}_{q(\theta'|\mathbf{z}^N)}m_{\theta'}(x))^2-\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}(Y-m_\theta(x))^2\\&=-2\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x)\mathbb{E}_{q(\theta'|\mathbf{z}^N)}m_{\theta'}(x)+(\mathbb{E}_{q(\theta'|\mathbf{z}^N)}m_{\theta'}(x))^2+2\mathbb{E}_{q(\theta|\mathbf{z}^N)}(m_\theta(x))^2-\mathbb{E}_{q(\theta|\mathbf{z}^N)}(m_\theta(x))^2\\&=\mathbb{E}_{q(\theta|\mathbf{z}^N)}(m_\theta(x))^2-(\mathbb{E}_{q(\theta'|\mathbf{z}^N)}m_{\theta'}(x))^2\\&=\mathbb{E}_{q(\theta|\mathbf{z}^N)}(m_\theta(x)-\mathbb{E}_{q(\theta'|\mathbf{z}^N)}m_{\theta'}(x))^2\\&=\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)].\end{aligned}$$

This concludes the proof.

D.5 Proof of Theorem 2

Proof. To prove this theorem, we first introduce the following lemma about the Jensen gap

Lemma 2. *For any (x, y) and any posterior distribution conditioned on $\mathbf{Z}^N = \mathbf{z}^N$, we have*

$$\mathbb{E}_{q(\theta|\mathbf{z}^N)}(y - m_\theta(x))^2 = (y - \mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x))^2 + \mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)].$$

Proof. By definition,

$$\begin{aligned}\mathbb{E}_{q(\theta|\mathbf{z}^N)}(y - m_\theta(x))^2 &= y^2 - 2y\mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x) + \mathbb{E}_{q(\theta|\mathbf{z}^N)}[m_\theta(x)^2] \\ &= (y - \mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x))^2 + \mathbb{E}_{q(\theta|\mathbf{z}^N)}[m_\theta(x)^2] - [\mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x)]^2 \\ &= (y - \mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x))^2 + \mathbb{E}_{q(\theta|\mathbf{z}^N)}[m_\theta(x) - \mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x)]^2 \\ &= (y - \mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x))^2 + \mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)].\end{aligned}$$

This concludes the proof. □

First, we can prove Eq.(9) by the direct calculation. By definition

$$\begin{aligned}\mathrm{ER}^{(2)}(Y|x, \mathbf{z}^N, \theta^*) &\coloneqq R_q^{(2)}(Y|x, \mathbf{z}^N) - R^{(2)}(Y|x, \theta^*) \\ &= \mathbb{E}_{\nu(Y|x)q(\theta|\mathbf{z}^N)}(Y - m_\theta(x))^2 - \mathbb{E}_{\nu(Y|x)}(Y - \mathbb{E}_{\nu(Y|x)}[Y|x])^2 \\ &= -2m_{\theta*}(x)\mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x) + \mathbb{E}_{q(\theta|\mathbf{z}^N)}[m_\theta(x)^2] + m_{\theta*}(x)^2 \\ &= (m_{\theta*}(x) - \mathbb{E}_{q(\theta|\mathbf{z}^N)} m_\theta(x))^2 + \mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)] \\ &= \mathrm{PER}^{(2)}(Y|x, \mathbf{z}^N) + \mathrm{BER}_q^{(2)}(Y|x, \mathbf{z}^N),\end{aligned}$$

where we used the relation $\mathbb{E}_{\nu(Y|x)}[Y|x] = m_{\theta*}(x)$. By definition,

$$\mathrm{ER}^{(2)}(Y|x, \mathbf{z}^N, \theta^*) = R_q^{(2)}(Y|x, \mathbf{z}^N) - R^{(2)}(Y|x, \theta^*) \geq 0,$$

and for the squared loss for any action a, we have $\ell(y, a) \geq 0$. Combined these, we have

$$\mathrm{ER}^{(2)}(Y|x, \mathbf{z}^N, \theta^*) \leq R_q^{(2)}(Y|x, \mathbf{z}^N).$$

This concludes the proof of Eq.(9). □

The unconditional relation is derived by using relations in Eqs. (21) and (22). Finally, we get Eq.(10) by the PAC-Bayesian bound Eq.(2).

Finally, we discuss the assumption about $\mathbb{E}_{\nu(Y|x)}[Y|x] = m_{\theta*}(x)$. This means that we only need to correctly specify the “average” of the conditional distribution of Y given x, and we do not need to care about the other properties of the distribution of Y. In the regression task, $\mathbb{E}_{\nu(Y|x)}[Y|x] = m_{\theta*}(x)$ implies that the regression function is well specified and $\nu(Y|x) \neq p(y|x, \theta^*)$ means that the noise function is misspeficied.

For example, assume that $\nu(y|x)$ is the multi-modal distribution and $p(y|x, \theta) = N(y|m_\theta(x), \sigma^2)$. Then Theorem 2 still holds if average $E_{\nu(Y|x)}[Y]$ is correctly specified by $m_\theta(x)$. However, the distribution of $p(y|x, \theta)$ and $\nu(y|x)$ can be clearly different.

D.5.1 Discussion about the model misspecification of the regression function

We further discuss when $\mathbb{E}_{\nu(Y|x)}[Y|x] \neq m_{\theta*}(x)$. We simply express $\mathbb{E}_{\nu(Y|x)}[Y|x]$ as $\eta(x)$. By definition

$$\begin{aligned}
\mathrm{ER}^{(2)}(Y|x,\mathbf{z}^N,\theta^*) &\coloneqq R_q^{(2)}(Y|x,\mathbf{z}^N) - R^{(2)}(Y|x,\theta^*)\\
&= \mathbb{E}_{\nu(Y|x)q(\theta|\mathbf{z}^N)}(Y-m_\theta(x))^2 - \mathbb{E}_{\nu(Y|x)}(Y-m_{\theta*}(x))^2\\
&= -2\eta(x)\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x) + \mathbb{E}_{q(\theta|\mathbf{z}^N)}[m_\theta(x)^2] + 2\eta(x)m_{\theta*}(x) - m_{\theta*}(x)^2\\
&= -2m_{\theta*}(x)\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x) + \mathbb{E}_{q(\theta|\mathbf{z}^N)}[m_\theta(x)^2] + m_{\theta*}(x)^2\\
&- 2(\eta(x)-m_{\theta*}(x))\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x) + 2(\eta(x)-m_{\theta*}(x))m_{\theta*}(x)\\
&= \mathrm{PER}^{(2)}(Y|x,\mathbf{z}^N) + \mathrm{BER}_q^{(2)}(Y|x,\mathbf{z}^N) + 2(\eta(x)-m_{\theta*}(x))(m_{\theta*}(x)-\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x))\\
&\leq \mathrm{PER}^{(2)}(Y|x,\mathbf{z}^N) + \mathrm{BER}_q^{(2)}(Y|x,\mathbf{z}^N) + 2\|\eta(x)-m_{\theta*}(x)\|_2\|m_{\theta*}(x)-\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x)\|_2.
\end{aligned}$$

The final term represents the misspecification error of the regression function.

D.6 Discussion about $\mathcal{Y} = \mathbb{R}^d$

For $\mathcal{Y} = \mathbb{R}^d$, Lemma 2 and Theorem 1 hold since we can proceed the proof in the same way for $\mathcal{Y} = \mathbb{R}$. Thus, we can proceed the proof of Theorem 2 for $\mathcal{Y} = \mathbb{R}^d$ in the same way as $\mathcal{Y} = \mathbb{R}$. Thus Theorem 2 holds in $\mathcal{Y} = \mathbb{R}^d$.

As for Theorem 3, we consider $p(Y|x,\theta) = N(y|m_\theta(x), \mathrm{diag}(v^2))$, where $v^2 \in \mathbb{R}^d$, where $\mathrm{diag}(v^2)$ is the diagonal matrix with each entry is v_i^2. Then Theorem 3 still holds.

D.7 Proof of Theorem 3

We use the following change-of-measure inequality, also known as the transportation lemma (Boucheron et al., 2013; Xu & Raginsky, 2022; Xu, 2020).

Lemma 3. *Let W be a real-valued integrable random variable with probability distribution p. Let h be a convex and continuously differentiable function on a interval $(0,b]$ and assume $h(0) = h'(0) = 0$. Define for every $x \geq 0$, $h^*(x) = \sup_{0\leq\rho<b}\{\rho x - h(\rho)\}$ and let for every $y \geq 0$, $h^{*-1}(y) \coloneqq \sup\{x \in \mathbb{R} : h^*(x) \leq y\}$. Then if*

$$\ln \mathbb{E}_{p(W)} e^{\rho(W - \mathbb{E}_{p(W)}W)} \leq h(\rho),$$

is satisfied, for any probability distribution q, which is absolutely continuous with respect to p such that $\mathrm{KL}(q|p) \leq \infty$, we have

$$\mathbb{E}_{q(W)}W - \mathbb{E}_{p(W)}W \leq h^{*-1}(\mathrm{KL}(q|p)).$$

When $h(\rho) = \rho^2\sigma^2/2$ and $b = \infty$, this assumption is σ^2 sub-Gaussian property and $h^{*-1}(x) = \sqrt{2x}$.

Then we have the following relations

$$\begin{aligned}
&\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)\\
&= \mathbb{E}_{\nu(Y|x)}\mathbb{E}_{q(\theta|\mathbf{z}^N)}[-\ln p(Y|x,\theta) + \ln p(Y|x,\theta^*)]\\
&= \mathbb{E}_{\nu(Y|x)}\mathbb{E}_{q(\theta|\mathbf{z}^N)}[-\ln p(Y|x,\theta) + \ln \mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta) - \ln \mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta) + \ln p(Y|x,\theta^*)]\\
&= \mathbb{E}_{\nu(Y|x)}\mathbb{E}_{q(\theta|\mathbf{z}^N)}[-\ln p(Y|x,\theta) + \ln \mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta)] + \mathrm{PER}^{\log}(Y|x,\mathbf{z}^N)\\
&\leq \mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}[-\ln p(Y|x,\theta) + \ln \mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta)] + \mathbb{E}_{q(\theta|\mathbf{z}^N)}2\sqrt{\sigma^2\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta))} + \mathrm{PER}^{\log}(Y|x,\mathbf{z}^N)\\
&= -\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) + \mathrm{PER}^{\log}(Y|x,\mathbf{z}^N) + \mathbb{E}_{q(\theta|\mathbf{z}^N)}2\sqrt{\sigma^2\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta))},
\end{aligned}$$

where we used the Lemma 3 for $q = \nu(Y|x)$ and $p = p(Y|x,\theta)$ and the assumption that $\ln p(y|x,\theta)$ are σ^2 sub-Gaussian.

Thus, we have

$$\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) + \mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*) \leq \mathrm{PER}^{\log}(Y|x,\mathbf{z}^N) + 2\sqrt{\sigma^2\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta))}.$$

Note that the following relation holds by the Jensen inequality;

$$\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)\geq \mathrm{PER}^{\log}(Y|x,\mathbf{z}^N).$$

Thus, we have

$$\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N))\leq 2\sqrt{\sigma^2\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta))}.$$

Note that the following relation holds;

$$\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta)=\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta^*)+\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*).$$

By taking the expectation, we have

$$\begin{aligned}\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)&\leq \mathbb{E}_{q(\theta|\mathbf{z}^N)}2\sqrt{\sigma^2\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta))}\\&=\mathbb{E}_{q(\theta|\mathbf{z}^N)}2\sqrt{\sigma^2(\mathrm{KL}(\nu(Y|x)|p(Y|x,\theta^*)+\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*))}\end{aligned}$$

D.8 Proof of Corollary 1 (Well-specified setting)

$\nu(y|x)=p(y|x,\theta^*)$

For simplicity, we define the excess risk of log loss as $\ell(y,x,\theta,\theta^*)=-\ln p(y|x,\theta)-(-\ln p(y|x,\theta^*))$.

From the assumption that $\ln p(y|x,\theta^*)$ are $\ln p(y|x,\theta)$ σ^2 sub-Gaussian, from the Lemma 3, we have,

$$\mathbb{E}_{p(Y|x,\theta^*)}\ell(Y,x,\theta,\theta^*)-\mathbb{E}_{p(Y|x,\theta)}\ell(Y,x,\theta,\theta^*)\leq 2\sqrt{\sigma^2\mathrm{KL}(p(Y|x,\theta^*)|p(Y|x,\theta)))}$$

By taking the expectation $\mathbb{E}_{q(\theta|\mathbf{z}^N)}$, we have

$$\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta^*)}\ell(Y,x,\theta,\theta^*)-\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\ell(Y,x,\theta,\theta^*)\leq 2\mathbb{E}_{q(\theta|\mathbf{z}^N)}\sqrt{\sigma^2\mathrm{KL}(p(Y|x,\theta^*)|p(Y|x,\theta)))}$$

We can rewrite the right-hand side as

$$\begin{aligned}&\mathrm{KL}(p(Y|x,\theta^*)|p(Y|x,\theta))\\&=\mathbb{E}_{p(Y|x,\theta^*)}[-\log p(Y|x,\theta)]-\mathbb{E}_{p(Y|x,\theta^*)}[-\log p(Y|x,\theta^*)]\\&=\mathbb{E}_{\nu(Y|x)}[-\log p(Y|x,\theta)]-\mathbb{E}_{\nu(Y|x)}[-\log p(Y|x,\theta^*)]=\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*).\end{aligned}\tag{23}$$

Thus, we have

$$\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta^*)}\ell(Y,x,\theta,\theta^*)-\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\ell(Y,x,\theta,\theta^*)\leq 2\mathbb{E}_{q(\theta|\mathbf{z}^N)}\sqrt{\sigma^2(\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*))}.$$

Next, we evaluate the left-hand side of Eq. (23). As for the first term, we have

$$\begin{aligned}\mathbb{E}_{p(Y|x,\theta^*)}\mathbb{E}_{q(\theta|\mathbf{z}^N)}\ell(Y,x,\theta,\theta^*)&=\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta^*)}[-\ln p(Y|x,\theta)+\ln p(Y|x,\theta^*)]\\&\geq \mathbb{E}_{\nu(Y|x)}[-\ln\mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta)+\ln p(Y|x,\theta^*)]\\&=\mathrm{PER}^{\log}(Y|x,\mathbf{z}^N)\end{aligned}\tag{24}$$

where the first inequality is the Jensen inequality

As for the second term in Eq. (23), we have

$$\begin{aligned}-\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\ell(Y,x,\theta,\theta^*)&=\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}[-\ln p(Y|x,\theta^*)+\ln p(Y|x,\theta)]\\&\geq \mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}[-\ln\mathbb{E}_{q(\theta|\mathbf{z}^N)}p(Y|x,\theta)+\ln p(Y|x,\theta)]\\&=\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N).\end{aligned}$$

where we used the relation $H[p]:=-\mathbb{E}_p\ln p=\inf_{q>>p}\mathbb{E}_p[-\ln q]$ in the inequality.

Combined these, we have

$$\mathrm{PER}^{\log}(Y|x,\mathbf{z}^N)+\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)\leq 2\mathbb{E}_{q(\theta|\mathbf{z}^N)}\sqrt{\sigma^2\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)}.$$

Based on the conditional and un-conditional BER and PER discussed in Appendix D.2, we can get the unconditional version of the theorem by taking the expectation $\mathbb{E}_{\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X=x)}$ instead of $\mathbb{E}_{q(\theta|\mathbf{z}^N)}$ and moving the expectation inside the square root by using the Jensen inequality.

D.9 Discussion about the logistic regression and softmax likelihood

Here we discuss the logistic regression. For logistic regression, we define the model as $p(Y=1|x,\theta)=\mathrm{sig}(\phi_\theta(x))$ where $\mathrm{sig}=1/(1+e^{-x})$ is the sigmoid function and $\phi_\theta(x)$ is the feature vector.

Then we can upper bound the reverse KL divergence as follows:

$$\begin{aligned}\mathrm{KL}(p(Y|x,\theta)|p(Y|x,\theta^*)) &\leq \log\left(1+\frac{2\mathrm{TV}^2(p(Y|x,\theta)|p(Y|x,\theta^*))}{\min(\mathrm{sig}(\phi_\theta(x)),1-\mathrm{sig}(\phi_\theta\phi(x)))}\right)\\ &\leq \log\left(1+4e^{\phi_\theta(x)}\mathrm{TV}^2(p(Y|x,\theta)|p(Y|x,\theta^*))\right)\\ &\leq \log\left(1+8e^{\phi_\theta(x)}\mathrm{KL}(p(Y|x,\theta^*)|p(Y|x,\theta))\right)\end{aligned}$$

where TV is the total variation distance, and we used the reverse Pinsker inequality (Sason & Verdú, 2016) to upper bound the reverse KL divergence by total variation distance, and then we upper bound the total variation distance by Pinsker inequality, which results in forward KL divergence. We also used the following relation;

$$\min(\mathrm{sig}(\phi_\theta(x)),1-\mathrm{sig}(\phi_\theta\phi(x)))\geq\frac{e^{-\phi_\theta(x)}}{2}.$$

Thus, we have

$$\mathrm{PER}^{\log}(Y|x,\mathbf{x}^N)+\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)\leq\log\left(1+8e^{\phi_\theta(x)}\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)\right)+\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*).$$

Thus, we can see that the sum of PER and BER is upper-bounded by ER.

For softmax classification problem, when there exists K classes, that is, $y=k$ with $k\in 1,\cdots,K$, we define the model as $p(Y=k|x,\theta)=e^{\phi_{\theta,k}(x)}/\sum_k e^{\phi_{\theta,k}(x)}$, where $\phi_{\theta,k}(x)$ is the feature vector. Similar to the logistic regression, we have

$$\begin{aligned}\mathrm{KL}(p(Y|x,\theta)|p(Y|x,\theta^*)) &\leq \log\left(1+\frac{2\mathrm{TV}^2(p(Y|x,\theta)|p(Y|x,\theta^*))}{\min_k e^{\phi}_{\theta,k}(x)/\sum_k e^{\phi_{\theta,k}(x)}}\right)\\ &\leq \log\left(1+\sum_k e^{\phi_{\theta,k}(x)}\max_k e^{-\phi_{\theta,k}(x)}\mathrm{TV}^2(p(Y|x,\theta)|p(Y|x,\theta^*))\right)\\ &\leq \log\left(1+2\sum_k e^{\phi_{\theta,k}(x)}\max_k e^{-\phi_{\theta,k}(x)}\mathrm{KL}(p(Y|x,\theta^*)|p(Y|x,\theta))\right)\end{aligned}$$

Thus, the input of the softmax function is bounded, and the sum of PER and BER is upper-bounded by ER, similar to the logistic regression.

However, as we discuss in Appendix D.11, we can easily enforce the condition that the log loss is bounded for the classification task. Thus, we can easily upper bound the softmax and logistic regression model.

D.10 Convergence of the entropy

Corollary 2. *Under the same assumption as Corollary 1, assume that $-\ln p(y|x,\theta)$ satisfies the σ^2 sub-Gaussian property and $\nu(y|x)=p(y|x,\theta^*)$ holds. Conditioned on $(x,\mathbf{z}^N)$, we have*

$$H[p^q(Y|x,\mathbf{z}^N)]\leq R_q^{\log}(Y|x,\mathbf{z}^N)+2\sqrt{2\sigma^2\mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)},$$

and if the excess risk bound of the PAC-Bayesian theory Eq.(2) holds, we have

$$H[p^q(Y|X,\mathbf{Z}^N)]-H[p(Y|X,\theta^*)]=\mathcal{O}(\sqrt{\ln N/N^{1/2}}).$$

Proof. If the log loss $-\ln p(Y|x,\theta)$ satisfies the σ^2 sub-Gaussian property, conditioned on $(x,\theta,\mathbf{z}^N)$, using Lemma 3 for the sub-Gaussian case, we have

$$\mathbb{E}_{p(Y|x,\theta^*)}(-\ln p(Y|x,\theta))-\mathbb{E}_{p(Y|x,\theta)}(-\ln p(Y|x,\theta))\leq\sqrt{2\sigma^2\mathrm{KL}(p(Y|x,\theta^*)|p(Y|x,\theta))}.$$

Thus, by taking the expectation about $q(\theta|\mathbf{z}^N)$, we have

$$\mathbb{E}_{q(\theta|\mathbf{z}^N)}H[p(Y|x,\theta)] \leq R_q^{\log}(Y|x,\mathbf{z}^N) + \sqrt{2\sigma^2 \mathrm{ER}^{\log}(Y|x,\mathbf{z}^N,\theta^*)}. \tag{25}$$

we used the relation $H[p] := -\mathbb{E}_p \ln p = \inf_{q>>p} \mathbb{E}_p[-\ln q]$ in the inequality and we used Eq. (23).

Next, we take the expectation over $\nu(\mathbf{Z}^N)q(\theta|\mathbf{Z}^N)\nu(X)$ in Eq.(25) instead of $q(\theta|\mathbf{z}^N)$, we have

$$\begin{aligned}
&H[p^q(Y|X,\mathbf{Z}^N)] - H[p(Y|X,\theta^*)] \\
&\leq R_q^{\log}(Y|X,\mathbf{Z}^N) - H[p(Y|X,\theta^*)] + \sqrt{2\sigma^2 \mathrm{ER}^{\log}(Y|X,\mathbf{Z}^N,\theta^*)} \\
&= \mathrm{ER}^{\log}(Y|X,\mathbf{Z}^N,\theta^*) + \sqrt{2\sigma^2 \mathrm{ER}^{\log}(Y|X,\mathbf{Z}^N,\theta^*)}.
\end{aligned}$$

Then from the excess risk bound of the PAC-Bayesian theory Eq.(2), we have $\mathrm{ER}^{\log}(Y|X,\mathbf{Z}^N,\theta^*) = \mathcal{O}(\ln N/N^{1/2})$, we get the bound. □

D.11 Bounded log loss

In Appendix D.9, we discussed the setting of logistic and softmax classification. For such classification tasks, we can enforce the bounded condition of the log loss by simple rescaling. For example, when considering softmax classification with C classes, we enforce the classifier so that each predicted class probability is at least $\exp(-L)$. Then, clearly, the log loss is always bounded by L. This can be achieved as follows (We cite the technique introduced in Nguyen et al. (2021)). If the softmax classifier outputs $p_1, \cdots, p_C$, we augment it into $(p_1 \cdot K + \delta, \cdots, p_C \cdot K + \delta)$, where $\delta := \exp(-L)$ and $K := 1 - \delta \cdot C$. The advantage of this rescaling is that this never changes the output prediction class while upper-bounding the log loss. Then, we can use the sub-Gaussian bound for the relation between BER, PER, and ER.

Other than classification, by using a constant $c \in (0,1)$, we introduce a smoothed version of the log loss as

$$-\log((1-c)p(y|x,\theta) + c).$$

Then, the log loss will be bounded.

D.12 Additional discussion for the general loss function

As introduced in Section 2.2, this study primarily focuses on variance and mutual information, which are widely used in Bayesian inference and uncertainty quantification tasks. We demonstrated in Section 3 that these measures are deeply connected to squared loss and logarithmic loss, respectively. On the other hand, for general convex loss functions, it is possible to develop a theoretical framework for excess risk under sub-Gaussian assumptions, as shown in Theorem 3. That is, we can define the Bayesian excess risk and its relation to the excess risk by utilizing the similar proof technique as in Theorem 3. On the other hand, for such general loss functions, it becomes more challenging to interpret how they can serve as measures of uncertainty, unlike squared loss and logarithmic loss.

E Discussion of entropic risk

E.1 Relation between the log loss and the squared loss

Before showing the relation between posterior variance and mutual information, we introduce an important lemma used in the analysis.

Lemma 4 (Lemma 1 in (Haussler & Opper, 1997))**.** *Let $P(w)$ be a measure on a set W and $Q(v)$ be a measure on a set V. For any real-valued function $u(w,v)$, we have*

$$-\int_V dQ(v) \ln \int_W dP(w) e^{u(w,v)} \leq -\ln \int_W dP(w) e^{\int_V dQ(v)u(w,v)}.$$

For completeness, we show the proof.

Proof. For any real-valued functions u_1 and u_2 and $0 \leq \alpha \leq 1$, we have

$$\begin{aligned}\int_W dP(w) e^{\alpha u_1(w)+(1-\alpha)u_2(w)} &= \int_W dP(w) \left(e^{u_1(w)}\right)^{\alpha} \left(e^{u_2(w)}\right)^{1-\alpha} \\ &\leq \left(\int_W dP(w) e^{u_1(w)}\right)^{\alpha} \left(\int_W dP(w) e^{u_2(w)}\right)^{1-\alpha},\end{aligned}$$

where we used Hölder's inequality. Taking the logarithmic function, this shows that $\ln dP(w)e^{u(w,v)}$ is convex in u. Thus, the result follows by using the Jensen inequality. □

Using this, we show the relation between posterior variance and the mutual information ($\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N)$).

Lemma 5. *For the Gaussian likelihood $p(y|x,\theta) = N(y|m_\theta(x), v^2)$, we have*

$$\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) \leq \frac{\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)]}{2v^2}.$$

Proof.

$$\begin{aligned}\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) &= I_\nu(\theta; Y|x,\mathbf{z}^N) \\ &= \mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}\left[-\ln \mathbb{E}_{q(\theta'|\mathbf{z}^N)} p(Y|x,\theta') + \ln p(Y|x,\theta)\right] \\ &= -\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)} \ln \mathbb{E}_{q(\theta'|\mathbf{z}^N)} e^{\ln p(Y|x,\theta')-\ln p(Y|x,\theta)}.\end{aligned}$$

Then, applying Lemma 4, we have

$$\begin{aligned}&\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) \\ &\leq -\ln \mathbb{E}_{q(\theta'|\mathbf{z}^N)} e^{\mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)} \ln p(Y|x,\theta')-\ln p(Y|x,\theta)} \\ &\leq -\ln \mathbb{E}_{q(\theta'|\mathbf{z}^N)} e^{-\frac{1}{2v^2}\mathbb{E}_{q(\theta|\mathbf{z}^N)p(Y|x,\theta)}(Y-m_{\theta'}(x))^2-(Y-m_\theta(x))^2} \\ &\leq -\ln \mathbb{E}_{q(\theta'|\mathbf{z}^N)} e^{-\frac{1}{2v^2}\mathbb{E}_{q(\theta|\mathbf{z}^N)}(v^2+f_\theta^2(x)-2m_{\theta'}(x)m_\theta(x)+m_{\theta'}^2(x)-(v^2+f_\theta^2(x)+f_\theta^2(x)-2m_\theta^2(x)))} \\ &= -\ln \mathbb{E}_{q(\theta'|\mathbf{z}^N)} e^{-\frac{1}{2v^2}(\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta^2(x)-2\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x)m_{\theta'}(x)+m_{\theta'}^2(x))} \\ &\leq \frac{\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)]}{2v^2}.\end{aligned}$$

□

E.2 Discussion about the direct loss minimization

In the main paper, we focus on the entropic risk. Here we discuss the relationship between direct loss minimization (DLM) and the EU. As written in the main paper, the DLM method minimizes the prediction risk (PR) directly, so it can be seen as the pseudo-Bayesian inference (Sheth & Khardon, 2020). The common choice of the loss function is the log loss, so we minimize the negative of the log-likelihood of the approximate posterior distribution in the DLM method. This has been used in 2nd-order PAC Bayesian inference and the entropic loss with $\alpha = 1$.

Here we pick up the squared loss in DLM. From Lemma 4, we have

$$\mathbb{E}_{q(\theta|\mathbf{z}^N)}(y - m_\theta(x))^2 = (y - \mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x))^2 + \mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)]. \tag{26}$$

The left-hand side is known as the Gibbs risk and is used in the standard regression methods, such as the pseudo-Bayesian learning or MC dropout methods. On the other hand, in the DLM, we minimize the first term of the right-hand side of Eq.(26), which corresponds to $PR^{(2)}$. Thus, from Eq.(26), we can see that minimizing the Gibbs risk in the left-hand side results in regularizing PR and BER simultaneously. It is also clear that minimizing only PR results in a larger EU compared to minimizing the Gibbs risk. This fact has been used to obtain the diverse ensembles in deep learning (Ortega et al., 2022).

We also remark that the purpose of the DLM is to derive better (randomized) hypotheses for a given loss function. For regression tasks, this leads to minimizing the squared loss to obtain regression functions on average. In contrast, our analyses focus on clarifying the relationship between epistemic uncertainty, loss functions, and prediction functions. Our results provide insights into the connections between DLM and epistemic uncertainty. Furthermore, our proposed algorithm (BER) can be interpreted as an extension of DLM with a novel regularization term, designed to better uncertainty quantification.

E.3 Relation to the function space VI

In Sun et al. (2019); Wei & Khardon (2024), they considered the variational inference in the function space, that is, given a function f, they introduced $p(y|f)$ as the likelihood and $q(f)$ as the variational posterior distribution over f. It has been conjectured that developing the VI over the function space results in better uncertainty quantification compared to the VI over the parameter space when considering the neural network models.

Our theory can potentially apply to those VI in function spaces. Specifically, for the squared loss, we expect that Theorem 2 still holds since the proof of Theorem 2 might be applied without modification when we consider the posterior distribution over the function space, rather than the parameter space. As a result, the relationship in Eq. (10), where BER serves as a lower bound for excess risk, also holds in this case.

For the logarithmic loss, we expect that Theorem 3 and Corollary 1 also hold. The key elements in these proofs are the sub-Gaussian and convexity of the logarithmic loss, and thus, we believe that proofs for those statements can be extended to the posterior distribution over the function space.

However, since we are dealing with conditional probabilities and the infimum of decision rules, extending these concepts to a potentially infinite-dimensional function space requires more care mathematically. Therefore, we leave the pursuit of rigorous results as future work.

F Detailed description of the proposed method

F.1 Derivation of Eq. (14)

Here we derive Eq. (14) using the definition of PER as follows;

$$
\begin{aligned}
R^{\log}(y|x,\mathbf{z}^N) - (1-\lambda)\mathrm{BER}^{\log}(Y|x,\mathbf{z}^N) &\leq \mathbb{E}_{q(\theta|\mathbf{z}^N)}\frac{|y-m_\theta(x)|^2}{2v^2} + \frac{\ln 2\pi v^2}{2} - (1-\lambda)\frac{\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)]}{2v^2} \\
&= \frac{(y-\mathbb{E}_{q(\theta|\mathbf{z}^N)}m_\theta(x))^2}{2v^2} + \frac{\ln 2\pi v^2}{2} + \lambda\frac{\mathrm{Var}_{\theta|\mathbf{z}^N}[m_\theta(x)]}{2v^2},
\end{aligned}
$$

where we used Lemma 5 in the first inequality, and we used Jensen inequality for the second inequality, and we used Lemma 2 in the final line.

F.2 Derivation of PAC-Bayesian bound

First, we show the PAC-Bayesian bound for our proposed method. Following the high-probability bound of Theorem 5, we have the following generalization error bound:

Theorem 6. *Given a distribution $\nu(Z)$, for any prior distribution $p(\theta)$ over Θ independent of $\mathbf{Z}^N$ and for any $\xi \in (0,1)$ and $c > 0$, with probability at least $1-\xi$ over the choice of training data $\mathbf{Z}^N$, for any posterior distribution $q(\theta|\mathbf{Z}^N)$ over Θ, we have*

$$
\begin{aligned}
&\mathbb{E}_{\nu(Z)}\left[\frac{(Y-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta(X))^2}{2v^2} + \lambda\frac{\mathrm{Var}_{\theta|\mathbf{Z}^N}[m_\theta(x)]}{2v^2}\right] + \frac{\ln 2\pi v^2}{2} \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{E}_{q(\theta|\mathbf{Z}^N)}(Y_i-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta(X_i))^2}{2v^2} + \lambda\frac{\mathrm{Var}_{\theta|\mathbf{Z}^N}[m_\theta(X_i)]}{2v^2}\right] + \frac{\ln 2\pi v^2}{2} + \frac{\mathrm{KL}(q|p)+\frac{1}{2}\ln\frac{1}{\xi}+\frac{1}{2}\Omega_{p,\nu}(c,N)}{cN},
\end{aligned}
$$

where

$$
\begin{aligned}
\Omega_{p,\nu}(c,N) &:= \ln \mathbb{E}_{p(\theta)p(\theta')}\mathbb{E}_{\nu(\mathbf{Z})^N}\exp[cN(\mathbb{E}_{\nu(Z)}L(Z,\theta,\theta') - \frac{1}{N}\sum_{n=1}^{N}L(Z_n,\theta,\theta'))], \\
L(z,\theta,\theta') &:= \frac{(y-m_\theta(x))^2}{2v^2} + (\lambda-1)\frac{m_\theta^2(x) - m_\theta(x)m_{\theta'}(x)}{2v^2}.
\end{aligned}
$$

Proof. The proof is similar to Masegosa et al. (2020). First, note that

$$\begin{aligned}&\frac{(y-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta(x))^2}{2v^2}+\lambda\frac{\mathrm{Var}_{\theta|\mathbf{Z}^N}[m_\theta(x)]}{2v^2}\\&=\mathbb{E}_{q(\theta|\mathbf{Z}^N)}\frac{(y-m_\theta(x))^2}{2v^2}+(\lambda-1)\frac{\mathrm{Var}_{\theta|\mathbf{Z}^N}[m_\theta(x)]}{2v^2}\\&=\mathbb{E}_{q(\theta|\mathbf{Z}^N)}\frac{(y-m_\theta(x))^2}{2v^2}+(\lambda-1)\frac{\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta^2(x)-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta(x)\mathbb{E}_{q(\theta'|\mathbf{Z}^N)}m_{\theta'}(x)}{2v^2}.\end{aligned}$$

Based on this, we define the tandem loss as

$$L(z,\theta,\theta'):=\frac{(y-m_\theta(x))^2}{2v^2}+(\lambda-1)\frac{m_\theta^2(x)-m_\theta(x)m_{\theta'}(x)}{2v^2}.$$

Then by considering the prior $p(\theta,\theta')=p(\theta)p(\theta')$, using Theorem 5, we have

$$\begin{aligned}\mathbb{E}_{\nu(Z)q(\theta|\mathbf{Z}^N)q(\theta'|\mathbf{Z}^N)}L(Z,\theta,\theta')\leq&\frac{1}{N}\sum_{n=1}^N\mathbb{E}_{q(\theta|\mathbf{Z}^N)q(\theta'|\mathbf{Z}^N)}L(z_n,\theta,\theta')\\&+\frac{\mathrm{KL}(q(\theta|\mathbf{Z}^N)q(\theta'|\mathbf{Z}^N)|p(\theta)p(\theta'))+\ln\xi^{-1}+\Omega_{p,\nu}(c,N)}{cN},\end{aligned}$$

where

$$\Omega_{p,\nu}(c,N):=\ln\mathbb{E}_{p(\theta)p(\theta')}\mathbb{E}_{\nu(\mathbf{Z})^N}\exp[cN(\mathbb{E}_{\nu(Z)}L(Z,\theta,\theta')-\frac{1}{N}\sum_{n=1}^N L(Z_n,\theta,\theta'))].$$

Since $\mathrm{KL}(q(\theta|\mathbf{Z}^N)q(\theta'|\mathbf{Z}^N)|p(\theta)p(\theta'))=2\mathrm{KL}(q(\theta|\mathbf{Z}^N)|p(\theta))$, by setting $c=2c'$, we get the result. □

Thus, the constant $\Omega_{p,\nu}$ depends only on the setting of the problem. We optimize the right-hand side of Eq.(15) as the objective function.

Next, we discuss the relation between rBER and standard VI. The objective function of standard VI is

$$-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}\ln N(y|m_\theta(x),v^2)=\frac{(y-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta(x))^2+\mathrm{Var}_{\theta|\mathbf{Z}^N}[m_\theta(x)]}{2v^2}+\frac{1}{2}\ln 2\pi v^2.$$

Thus, we can interpret that the log loss of the Gaussian likelihood corresponds to the prediction risk and Bayesian excess risk. Since the prediction risk corresponds to the prediction performance, standard VI implicitly controls the prediction performance and the Bayesian excess risk. Our rBER can be regarded as

$$\mathrm{rBER}(\lambda)=\frac{1}{2v^2}\frac{1}{N}\sum_{n=1}^N\left(\mathrm{PR}_q^{(2)}(y_n|x_n,\mathbf{Z}^N)+\lambda\mathrm{BER}^{(2)}(y_n|x_n,\mathbf{Z}^N)\right)+\frac{1}{2}\ln 2\pi v^2+\frac{1}{N}\mathrm{KL}(q|p).$$

Thus, BER has a flexible weight for regularizing the uncertainty. Numerically, when $\lambda=0$, this corresponds to the setting where we simply optimize $\mathrm{PR}_q^{(2)}(y|x,\mathbf{Z}^N)$. This means we only consider the fitting performance. We numerically found that $\lambda=0$ results in large uncertainty due to the lack of regularization. When $\lambda=1$, we found that the uncertainty is underestimated.

Finally, we remark on the relation between Bayesian excess risk and $\mathrm{Var}m_\theta(x)$. Since we focus on the log loss, we can consider the following type of objective function.

$$\frac{1}{N}\sum_{n=1}^N\left[\frac{\mathbb{E}_Q(y_n-\mathbb{E}_{q(\theta|\mathbf{Z}^N)}m_\theta(x_n))^2}{2v^2}+\lambda\mathrm{BER}^{\log}(y_n|x_n,\mathbf{Z}^N)\right]+\frac{\ln 2\pi v^2}{2}+\frac{1}{N}\mathrm{KL}(q|p),\tag{27}$$

where we use the Bayesian excess risk directly, instead of $\mathrm{Var}m_\theta(x)$. Note that from Appendix E, $\mathrm{BER}^{\log}(Y|x,\mathbf{Z}^N)\leq\mathrm{Var}[m_\theta(x)]/v^2$ holds for the Gaussian likelihood. Thus, Eq.(27) and our BER behave similarly. From the numerical point of view, implementing $\mathrm{Var}[m_\theta(x)]$ is easier than Eq.(27) since we calculate the variance of the prediction.

F.3 Additional discussion with alpha-divergence

About the forward and reverse KL divergences, their relationship with α-divergence has been discussed in Hernandez-Lobato et al. (2016); Li & Gal (2017); Minka et al. (2005). Specifically, the reverse KL corresponds to α-divergence with $\alpha = 0$, and the forward KL corresponds to $\alpha = 1$. Moreover, minimizing the α-divergence aligns with the Entropic risk.

Eq. (13) and the corresponding objective functions of our proposed rBER reveal that these α-divergence-based objectives include regularization about the uncertainty about the choice of λ. Specifically, in our method, setting $\lambda = 0$ excludes variance regularization, making it equivalent to $\alpha = 1$, i.e., the forward KL. On the other hand, setting $\lambda = 1$ corresponds to $\alpha = 0$, i.e., the forward KL. Intermediate values of λ relate to α-divergence with α between 0 and 1. As noted in Hernandez-Lobato et al. (2016); Li & Gal (2017); Minka et al. (2005), varying α between 0 and 1 captures different levels of uncertainty. Our method instead adjusts λ to explore a different perspective on uncertainty measures.

When $\alpha = 1$, rBER corresponds to the standard VI. Here, we discuss how rBER behaves for $\alpha \neq 1$. Recall that the proposed BER is based on Eq. (13), which corresponds to the minimization of the bound in Corollary 1.

Next, we discuss whether Corollary 1 holds for $\alpha \neq 1$. For $\alpha = 1$, the Entropic risk for $\alpha = 1$, i.e., $\mathrm{Ent}^{\ell}_{\alpha=1}(y, x) = PR$, corresponds to PER in Corollary 1. We considered whether the PER term in Corollary 1 could be replaced with the general Entropic risk, $E_{Y|x}[\mathrm{Ent}^{\ell}_{\alpha}(y, x)]$. If so, a corresponding relation for Eq. (13) under $\alpha \neq 1$ could be derived, enabling the rBER to incorporate $\alpha \neq 1$.

From the proof of Corollary 1, we can see that the PER in Corollary 1 can indeed be rewritten in terms of Entropic risk. Specifically, the PER term on the left-hand side of Corollary 1 can be replaced with

$$E_{Y|x}[\mathrm{Ent}^{\ell}_{\alpha}(Y, x) + \log p(y|x, \theta^*)],$$

where the loss ℓ is the logarithmic loss), and the inequality still holds. Thus, we can derive Eq. (13) for $\alpha \neq 1$ and from this relation, we can derive the objective function of BER for $\alpha \neq 1$ similarly to the case of $\alpha = 1$.

This can be confirmed as follows; From Eq. (24) in the proof of Corollary 1, we have

$$\begin{aligned}\mathrm{PER}^{log}(Y|x, \mathbf{z}^N) &= \mathbb{E}_{\nu(Y|x)}[-\ln \mathbb{E}_{q(\theta|\mathbf{z}^N)} p(Y|x, \theta) + \ln p(Y|x, \theta^*)] \\ &\leq \mathbb{E}_{q(\theta|\mathbf{z}^N)} \mathbb{E}_{p(Y|x,\theta)}[-\ln p(Y|x, \theta) + \ln p(Y|x, \theta^*)]\end{aligned}$$

and this inequality upper bounds the PER by the Jensen inequality. We can modify this as $\mathbb{E}_{\nu(Y|x)}\mathrm{Ent}^{l}_{\alpha}(y, x) \leq \mathbb{E}_{q(\theta|\mathbf{z}^N)}\mathbb{E}_{p(Y|x,\theta)}[-\ln p(Y|x, \theta)]$ We then proceed with the rest of the proof the same as Corollary 1.

G Numerical experiments

In this section, we describe the detailed settings of the experiments. We also present the additional experimental results. We used two NVIDIA GPUs with 32GB memory (NVIDIA GeForce RTX 3090) and CPU (Intel(R) Core(TM) i9-10850K CPU, 3.60GHz) with 32 GB memory for all the numerical experiments.

G.1 Particle variational inference

Here we briefly explain the PVI and existing methods. In PVI, we use the empirical distribution $\rho(\theta) = \frac{1}{M}\sum_{i=1}^{M} \delta_{\theta_i}(\theta)$ as the posterior distribution. Here $\delta_{\theta_i}(\theta)$ is the Dirac distribution with a mass at θ_i. We refer to these M samples as particles. PVI (Liu & Wang, 2016; Wang et al., 2019) approximates the posterior through iteratively updating the empirical distribution by interacting them with each other:

$$\theta_i^{\mathrm{new}} \leftarrow \theta_i^{\mathrm{old}} + \eta v_i(\{\theta_{i'}^{\mathrm{old}}\}_{i'=1}^{M}), \tag{28}$$

where $v(\{\theta\})$ is the update direction. v is composed of the gradient term and the repulsion term to enhance the diversity of the posterior distribution since we are often interested in the multi-modal information of the posterior distribution. For the update direction about v, see f-SVGD in Wang et al. (2019) and VAR in Futami et al. (2021) for details.

We follow the approach in VAR in Futami et al. (2021). They proposed using the gradient of the PAC-Bayesian bound for

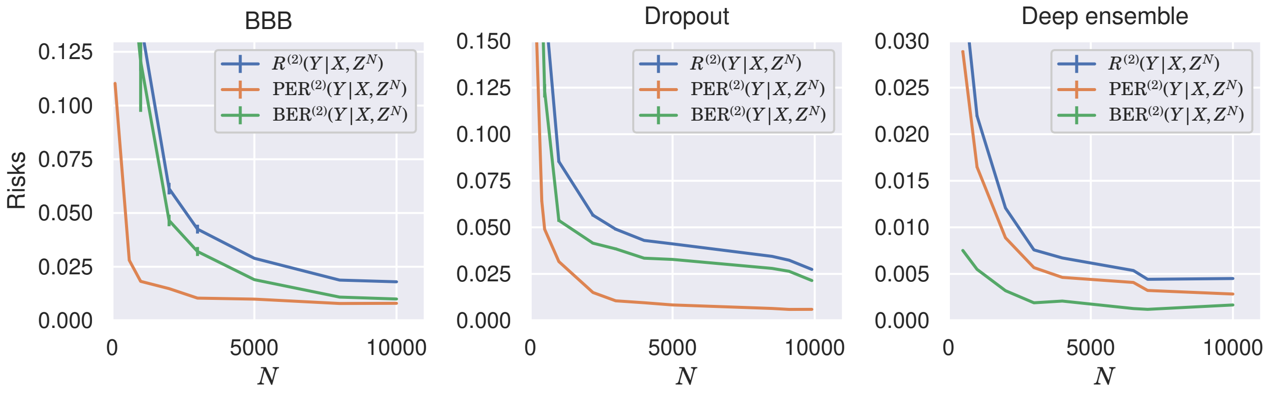

Figure 3: Result of toy data experiments: N represents a number of training data points, and the vertical line is the value of each excess risk.

the update direction v in Eq.(28). For example, VAR uses

$$v_i = \partial_i \mathcal{F}(\{\theta_i\}_{i=1}^N),$$

$$\mathcal{F}(\{\theta_i\}_{i=1}^N) := -\frac{1}{NM}\sum_{i=1}^{M}\sum_{n=1}^{N}[\ln p(y_n|x_n,\theta_i)+R(y_n,x_n)]+\frac{1}{N}\mathrm{KL}(\rho(\theta)|p(\theta)),$$

where R is the repulsion term to enhance the diversity. See Futami et al. (2021) for details. Following their setting, we consider using the following update direction

$$v_i = \partial_i \mathrm{rBER}(\lambda)$$

$$\mathrm{rBER}(\lambda) = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \mathbb{E}_{\rho(\theta)}m_\theta(x_n)|^2}{2v'^2}+\lambda\frac{\mathrm{Var}[m_\theta(x_n)]}{2v'^2}+\frac{\ln 2\pi v'^2}{2}+\frac{1}{N}\mathrm{KL}(\rho(\theta)|p(\theta)),$$

where $\rho(\theta)=\frac{1}{M}\sum_{i=1}^{M}\delta_{\theta_i}(\theta)$. We optimize v' by gradient descent.

G.2 Settings for Toy data experiments

For these experiments, we used the implementation in the previous work (Amini et al., 2020). For the toy data experiments, we used the Adam optimizer with the step size 0.0001 in the implementation of Amini et al. (2020). In the deep ensemble method, the number of ensembles is 5. We set other hyperparameters as the same as in Amini et al. (2020). Here we show the enlarged version of Fig. 1 in Fig. 3. (Fig. 3 is the results of 10 repetitions.)

G.3 BNN regression for UCI dataset

We used the same setting as the previous work (Wang et al., 2019; Futami et al., 2021). We used the Adam optimizer with a learning rate of 0.004. We used a batch size of 100 and ran 500 epochs for the dataset size to be smaller than 1000. For a larger dataset, we used a batch size of 1000 and ran 3000 epochs.

To calculate the PICP, we first calculate the 95% prediction interval. We then calculate the number of test data points included inside the prediction interval.

To calculate the MPIW, we calculated the mean of the prediction interval and normalized it by the maximum length of the test data point; $\max y_{\mathrm{test}} - \min y_{\mathrm{test}}$.

We show the additional results here. We show the result of PAC_E^2 and the negative log-likelihood.

We found that PAC_E^2 is similar to BER(0) measured in the negative log-likelihood, MPIW, and PICP. However, the prediction performance of PAC_E^2 in RMSE is significantly worse than BER(0). This is because the objective function of PAC_E^2 is

Table 3: Benchmark results on test PICP and MPIW.

Dataset	Avg. Test PICP and MPIW in parenthesis				
	f-SVGD	VAR	PAC^2_E	rBER(0)	rBER(0.05)
Concrete	0.82±0.03 (0.13±0.00)	0.87±0.04 (0.16±0.01)	0.97±0.02 (0.57±0.04)	0.99±0.02 (0.50±0.04)	**0.95±0.02** (0.25±0.02)
Boston	0.63±0.07 (0.10±0.02)	0.76±0.05 (0.14±0.01)	**0.94±0.04** (0.40±0.04)	0.97±0.01 (0.33±0.04)	0.92±0.04 (0.22±0.02)
Wine	0.79±0.03 (0.32±0.05)	0.85±0.02 (0.39±0.06)	0.98±0.01 (1.06±0.01)	0.99±0.00 (1.61±0.00)	**0.95±0.03** (0.32±0.15)
Power	0.43±0.01 (0.07±0.00)	0.82±0.01 (0.15±0.00)	0.99±0.00 (0.57±0.02)	0.99±0.01 (0.81±0.01)	0.96±0.01 (0.37±0.01)
Yacht	0.92±0.04 (0.02±0.01)	0.93±0.04 (0.04±0.01)	0.97±0.03 (0.07±0.00)	**0.96±0.03** (0.10±0.01)	**0.94±0.04** (0.08±0.01)
Protein	0.53±0.01 (0.24±0.01)	0.83±0.00 (0.58±0.01)	0.98 ±0.00 (1.44±0.06)	1.0 ±0.00 (5.04±0.01)	**0.96±0.01** (0.86±0.00)

Table 4: Benchmark results on negative test log-likelihood

Dataset	Avg. negative test log likelihood				
	f-SVGD	VAR	PAC^2_E	rBER(0)	rBER(0.05)
Concrete	-2.85±0.15	-2.81±0.06	-3.16±0.03	-3.50±0.03	-3.06±0.05
Boston	-2.34±0.31	-2.34±0.24	-2.61±0.08	-2.55±0.05	-2.38±0.16
Wine	-0.89±0.08	-0.90±0.06	-1.26±0.02	-1.84±0.03	-1.08±0.03
Power	-2.75±0.03	-2.80±0.03	-3.17±0.03	-3.95±0.04	-2.86±0.01
Yacht	-0.81±0.67	-0.87±0.38	-0.81±0.11	-1.62±0.17	-1.46±0.30
Protein	-2.70±0.00	-2.84±0.00	-3.30±0.00	-4.45±0.02	-2.94±0.00

the negative log-likelihood of the predictive distribution; thus, the performance in RMSE is not guaranteed. On the other hand, the objective function of BER(0) is based on the squared loss. Thus, it can show performance in RMSE. We show the results, which are not presented in the main paper owing to space limitations in Table 3 and 4.

Next, we evaluated how the RMSE, PICP, MPIW, and negative log-likelihood behave by changing λ in BER. We show the results in Fig.4 and 5. (Fig. 5 is the enlarged version of Fig. 2 in the main paper.)

We confirmed that the prediction performance measured in RMSE does not depend on the choice of λ. On the other hand, other measures depend on λ significantly. The ideal PICP is 0.95. Thus, we should choose λ by cross-validation. We also found that even $\lambda = 1.$ corresponds to the standard VI. It underestimates the PICP.

We have conducted additional numerical experiments for $\alpha = 0, 0.5, 1$. The results confirm that our method, which directly controls uncertainty via variance, outperforms the α-divergence variational inference in terms of uncertainty quantification.

G.4 Contextual bandit tasks

Here we explain the setting of the task. Our experiments follow the setting in Wang et al. (2019); Futami et al. (2021). Denote the context set as $\mathcal{S}$. For each time step t, an agent receives context $s_t \in \mathcal{S}$ from the environment. The agent choose action $a_t \in \{1, \dots, A\}$ based on the context s_t and get a reward $r_{a_t,t}$. We would like to minimize the pseudo-regret

$$R_T = \max_{\phi:\mathcal{S}\to\{1,\dots,A\}} \mathbb{E}\left[\sum_{t=1}^{T} r_{g(s_t),t} - \sum_{t=1}^{T} r_{a_t,t}\right],$$

where ϕ maps the context to the action. We consider a prior $\mu_{s,i,0}$ over a reward of context s and action i. Then, we update the prior to a posterior distribution using the observed reward. Following the previous work, we use Thompson sampling to select the action as

$$r_t \in \operatorname*{argmax}_{i=\{1,\dots,K\}} \hat{r}_{i,t}, \quad \hat{r}_{i,t} \sim \mu_{s,i,t}.$$

We consider a neural network regression model following the previous work (Wang et al., 2019; Futami et al., 2021), where the input is the context, and the output is the K-dimensional action. We place a prior distribution over the parameters of the network. We approximate the posterior distribution of the neural network parameters by PVI. All the hyperparameters are precisely the same as in the previous work (Wang et al., 2019).

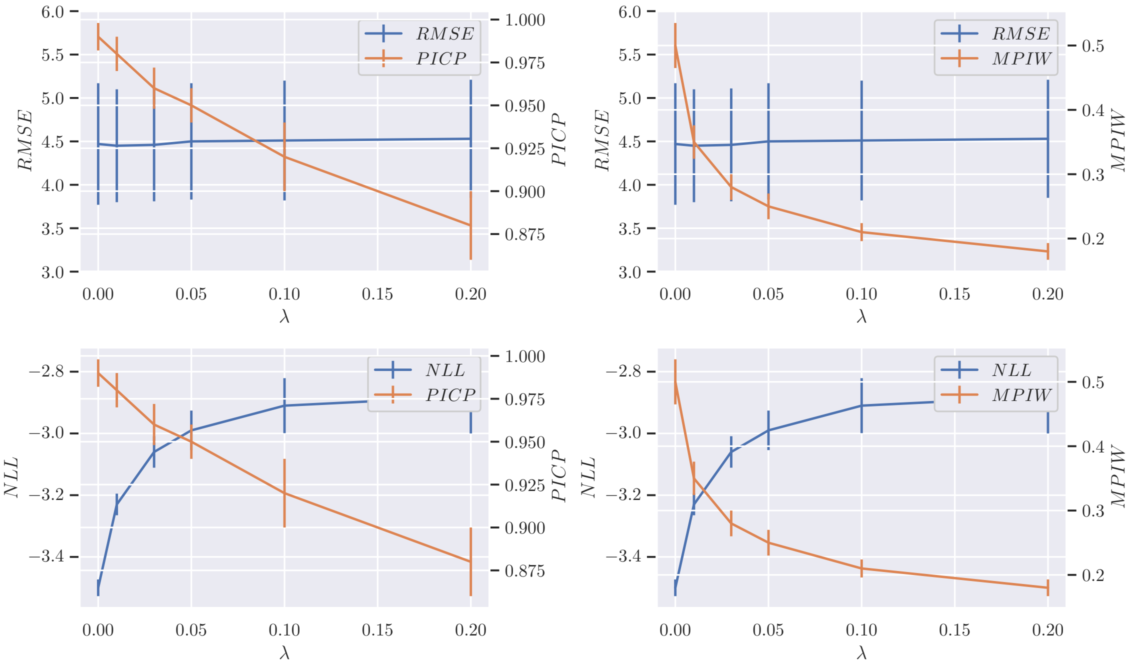

Figure 4: Concrete data in UCI dataset.

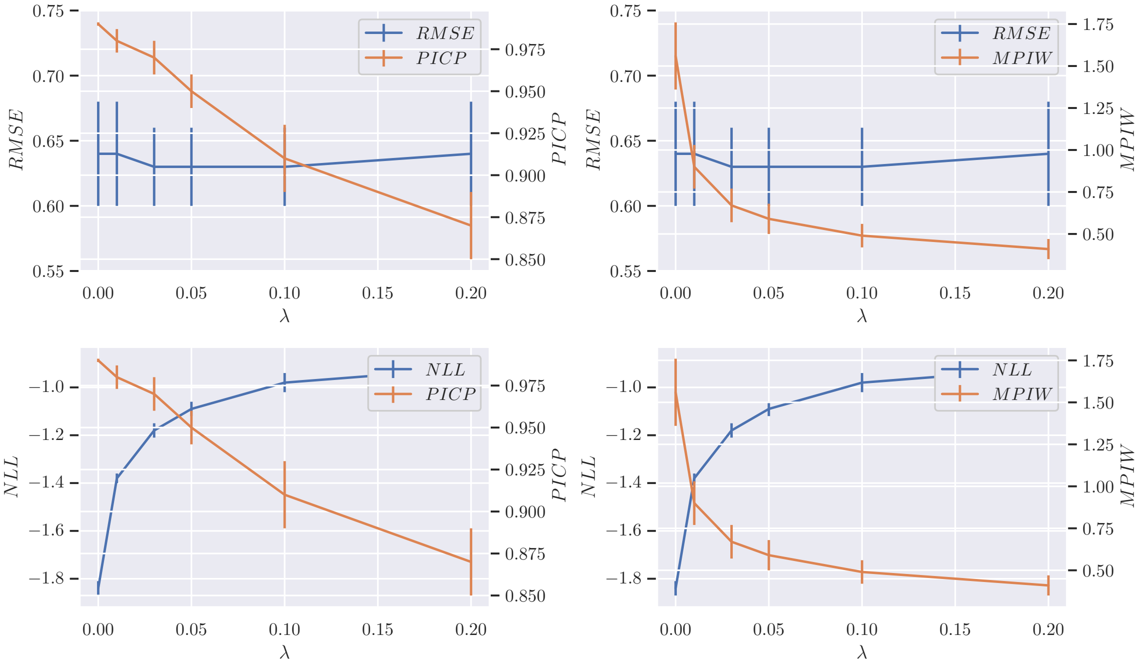

Figure 5: Wine data in UCI dataset.

Here we show additional results: Specifically, we tested the Bayes by Backpropagation (BBB) and Dropout methods on real data experiments, which were also used in the toy data experiments. For deep ensemble, the results are already included in the paper under the name MAP. Additionally, since our proposed method is based on α-divergence VI, we conducted experiments for $\alpha = 0.5, 1$. Below, we summarize those results for contextual bandit tasks.

Among the additional baseline methods, dropout showed relatively good performance. However, it still underperformed

Table 5: Cumulative regret relative to uniform sampling.

Dataset	BBB	Dropout	BB$\alpha(=1.0)$	BB$\alpha(=0.5)$
Mushroom	0.036 ± 0.002	0.056 ± 0.012	0.543 ± 0.000	0.539 ± 0.000
Financial	0.23 ± 0.013	0.17 ± 0.007	0.40 ± 0.007	0.70 ± 0.035
Statlog	0.126 ± 0.015	0.025 ± 0.006	0.194 ± 0.017	0.211 ± 0.014
CoverType	0.582 ± 0.022	0.306 ± 0.002	0.394 ± 0.005	0.605 ± 0.000

compared to the proposed method and other PVI methods, which are shown in the paper. On the other hand, α-divergence VI showed higher regret (worse performance), as the posterior uncertainty was excessively large.

The results confirm that our method, which directly controls uncertainty via variance, outperforms other baseline methods for the uncertainty quantification.