Minimum Excess Risk in Bayesian Learning

Aolin Xu and Maxim Raginsky, *Senior Member, IEEE*

***Abstract*— We analyze the best achievable performance of Bayesian learning under generative models by defining and upper-bounding the minimum excess risk (MER): the gap between the minimum expected loss attainable by learning from data and the minimum expected loss that could be achieved if the model realization were known. The definition of MER provides a principled way to define different notions of uncertainties in Bayesian learning, including the aleatoric uncertainty and the minimum epistemic uncertainty. Two methods for deriving upper bounds for the MER are presented. The first method, generally suitable for Bayesian learning with a parametric generative model, upper-bounds the MER by the conditional mutual information between the model parameters and the quantity being predicted given the observed data. It allows us to quantify the rate at which the MER decays to zero as more data becomes available. Under realizable models, this method also relates the MER to the richness of the generative function class, notably the VC dimension in binary classification. The second method, particularly suitable for Bayesian learning with a parametric predictive model, relates the MER to the minimum estimation error of the model parameters from data via various continuity arguments. We also extend the definition and analysis of MER to the setting with multiple model families and the setting with nonparametric models. Along the discussions we draw some comparisons between the MER in Bayesian learning and the excess risk in frequentist learning.**

***Index Terms*— Bayesian learning, generative models, Bayes risk, excess risk, uncertainty, data processing inequality.**

I. Introduction

BAYESIAN learning under generative models has been gaining considerable attention in recent years as an alternative to the frequentist learning. In the Bayesian setting, the observed data $Z^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is modeled as conditionally i.i.d. samples generated from a probabilistic model given the model realization $P_{Z|W}$, while the model is treated either as a random element of some parametric model family drawn according to a prior distribution of the model parameters W, or as a nonparametric random process [1]. The task of Bayesian learning is to predict a quantity of interest Y based on the observed data (X, Z^n), while the quality of the predictor ψ can be assessed by the expected loss with respect to some loss function $\mathbf{E}[\ell(Y, \psi(X, Z^n))]$. Computationally, Bayesian learning often relies on posterior sampling or approximation techniques [2]–[4] and hence has much higher complexity than its frequentist counterpart; nevertheless, the Bayesian viewpoint has many attractive features, e.g., reducing overfitting [5], quantifying uncertainty in making predictions [6]–[8], enabling model compression [9], etc. In contrast with the growing attention to the computational side of Bayesian learning, its performance analysis is relatively scarce compared to the volume of literature on the theoretical analysis of frequentist learning. In this paper, we set aside the computational issues in Bayesian learning and focus on analyzing its best achievable performance under the generative model with respect to general loss functions.

Manuscript received 29 December 2020; revised 26 November 2021; accepted 31 January 2022. Date of publication 23 May 2022; date of current version 22 November 2022. The work of Maxim Raginsky was supported in part by the NSF CAREER Award under Grant CCF-1254041; in part by the Illinois Institute for Data Science and Dynamical Systems (iDS2), an NSF HDR TRIPODS Institute, under Award CCF-1934986; in part by DARPA under the LwLL (Learning with Less Labels) Program; and in part by ARO MURI (Adaptive exploitation of non-commutative multimodal information structure) under Grant W911NF-15-1-0479.

Aolin Xu is with the University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: xuaolin@gmail.com).

Maxim Raginsky is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: maxim@illinois.edu).

Communicated by A. Krishnamurthy, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2022.3176056

A. Overview of the Presentation

In Section II, we define the *minimum excess risk* (MER) in Bayesian learning as the gap between the Bayes risk $R_\ell(Y|X, Z^n)$, defined as the minimum expected loss attainable by learning from the data, and its fundamental limit $R_\ell(Y|X, W)$, defined as the minimum expected loss that would be achieved if the model parameters were known. The MER is an algorithm-independent quantity that captures the uncertainty arising from the lack of knowledge of the underlying model parameters, commonly known as *epistemic uncertainty*. Its value and rate of convergence to zero reflect the difficulty of the learning problem. The decomposition of the Bayes risk into its fundamental limit and MER also provides rigorous definitions of the aleatoric uncertainty and the minimum epistemic uncertainty in Bayesian learning, the quantification of which has become an important research topic in recent years [10]–[13]. To the authors' knowledge, the general definition of MER is new, and has not been systematically studied before.

We then present two approaches to deriving upper bounds on the MER. First, we show in Section III that, under a generic parametric generative model and for a wide range of loss functions, the MER can be upper-bounded in terms of the conditional mutual information between the model parameters and the quantity being predicted given the observed data, $I(W; Y|X, Z^n)$. This leads to asymptotic upper bounds on the MER that scale as $O(d/n)$ or $O(\sqrt{d/n})$ depending on the loss function, where d is the dimension of the parameter space and n is the data size. It also reveals an MER-information relationship in Bayesian learning, echoing the generalization-information relationship in frequentist

learning [14]. Under realizable models, it is shown that for any bounded loss function, the MER for binary classification scales as $O(d/n)$, where d is the VC dimension of the generative function class. Next, we show in Section IV how MER can be upper-bounded via various continuity or smoothness arguments. One method under this approach relies on the smoothness of the decision rule in the model parameters, while the other relies on the smoothness of the minimum expected loss as a function of the predictive model. The resulting upper bounds show the dependence of the MER on the minimum achievable estimation error of the model parameters from the data, e.g., on the minimum mean square error of the estimated model parameters $R_2(W|X, Z^n)$. This explicitly shows how the difficulty of model parameter estimation translates into the difficulty of prediction due to the model uncertainty. Order-optimal MER upper bounds for linear regression are obtained from this approach.

The analysis of the MER in the single model family setting can be extended to the setting with multiple model families. The definition of MER can also be extended to the setting with nonparametric generative models, such as Gaussian processes, and the analysis based on conditional mutual information carries over to this setting. These extensions are briefly discussed in Section V. We close by summarizing the results and making some comparisons between the MER in Bayesian learning and the excess risk in frequentist learning in Section VI.

B. Relation to Existing Works

1) Accumulated Excess Risk for Log Loss: The closest connection between this work and prior literature is the MER for the logarithmic (log) loss defined in this paper and the *accumulated* excess risks for the log loss defined in Bayesian universal source coding [15], Bayesian sequential prediction [16], Bayesian density estimation [17], and Bayesian supervised learning [18], all of which turn out to be the mutual information between the model parameters and the observed data, and are achieved by the posterior predictive distribution as a soft predictor. The only work where more general loss functions is considered is the study of sequential prediction in [16], where an upper bound on the accumulated excess risk for bounded loss functions is derived. Our definition of MER goes beyond the log loss to general loss functions, which can be unbounded, and the MER in general is achieved not necessarily by the posterior predictive distribution, but by some hard predictor according to the loss function. In Section III we show that the MER for the log loss is nevertheless an important quantity, as it can be used to upper-bound the MER for many other loss functions. Most of the above works, with the exception of [18], considered only unsupervised learning, while our results hold for both supervised and unsupervised learning. In addition, the MER defined in this work is the *instantaneous* excess risk, instead of the accumulated risk studied in above works, thus is amenable to more refined analyses. Another closely related work that considered both supervised learning and instantaneous risk is [19], where the Bayes risk of binary classification with the zero-one loss is derived by relating it to the accumulated log loss, and is further related to the VC dimension of the generative function class. As only realizable models are considered in [19], the Bayes risk there is equal to the MER. In Section III-E, we also study the MER under realizable models, but our results go beyond binary classification and the zero-one loss.

2) Convergence of Posterior Distribution: A classical frequentist analysis of Bayesian inference is the convergence of the *posterior parameter distribution* to the true model parameters, assuming the data is sampled from some *fixed* model with the true parameters [20]–[22]. This analysis has recently been extended to deep neural network models [23]. The convergence of the *posterior predictive distribution* has also been studied under the same assumption [24]. The main difference between these works and ours is the assumption on the data distribution. In our work, the underlying data distribution considered in the performance analysis stays the same as the generative model based on which the optimal predictor, or the learning algorithm, is derived. In other words, the model parameters are assumed to be randomly drawn from the prior, and the data samples are drawn from the model given the model parameters. In addition, rather than the convergence of the posterior of the model parameters, we are interested in the accuracy of the predicted quantity of interest. In Section IV we reveal how this accuracy explicitly depends on the accuracy of the model parameter estimation, by studying the expected deviation of the posterior predictive distribution from the random true model.

3) PAC-Bayes: Another loosely related line of work in statistical learning is the PAC-Bayes framework in the frequentist setting [25]–[27] and its extension as the Bayes mixture model [28]. The main difference between the Bayesian setting considered here and the PAC-Bayes framework is again the underlying data distribution. For the former, the data distribution is restricted to a parametric or nonparametric family of generative models with the data samples being *conditionally* i.i.d. given the model realization, and there is virtually no restriction on candidate predictors. For the latter, the data samples are drawn *unconditionally* i.i.d. according to a completely unknown distribution, and the hypothetical Bayes-like update takes place in a *hypothesis space* consisting of admissible predictors only. The excess risk studied in this paper is thus not directly related to the generalization error or excess risk in the PAC-Bayes method. Nevertheless, the MER-information relationship in Theorem 6 is an interesting analogue of the generalization-information relationship in the frequentist setting [14, Theorem 1] that leads to an information-theoretic derivation of the PAC-Bayes algorithm.

C. A Note on Notation

Throughout the paper, random variables are denoted by uppercase letters and their realizations are in the corresponding lowercase letters. To keep the notation uncluttered, we may use $K_{U|v}$, $P_{U|v}$, $\mathbf{E}[U|v]$ and $\mathsf{var}[U|v]$ respectively to denote the probability transition kernel $K_{U|V=v}$, the conditional distribution $P_{U|V=v}$, the conditional expectation $\mathbf{E}[U|V = v]$ and the conditional variance $\mathsf{var}[U|V = v]$. When the conditioning variables are written in uppercase letters, these quantities are

random, and expectations can be taken with respect to the conditioning variables. Throughout the paper, $D(\cdot,\cdot)$ denotes a generic statistical distance, while the KL divergence is denoted by $D_{\mathrm{KL}}(\cdot\|\cdot)$. All probability spaces considered in this paper are Borel spaces, and all functions are measurable functions. We use natural logarithms throughout the paper.

II. Model and Definitions

A. Bayesian Learning Under Generative Model

The basic task in supervised learning is to construct an accurate predictor of Y, a quantify of interest, given an observation X, where the knowledge of the joint distribution of X and Y is vague but can be inferred from a historical dataset $((X_1,Y_1),\ldots,(X_n,Y_n))$. In the *model-based learning* framework, a.k.a. learning under a *generative model*, the joint distribution of X and Y is assumed to be an element of a known model family. The model family can be either parametric or nonparametric. We focus on the parametric case in this work, and defer a brief discussion on the nonparametric case to Section V-B. In the case of parametric modeling, the model family is a collection of parametrized distributions $\mathcal{M}=\{P_{X,Y|w}, w\in \mathsf{W}\}$, where w represents the vector of unknown model parameters belonging to some space W. Under the *Bayesian formulation*, the vector of model parameters W is itself treated as a random quantity with a prior distribution P_W, while the data samples are conditionally i.i.d. given W. Formally, the model parameters W, the data samples $Z^n:=(Z_1,\ldots,Z_n)$ with $Z_i:=(X_i,Y_i)$, $i=1,\ldots,n$, and the pair $Z=(X,Y)$ consisting of the fresh observation X and the quantity Y to be predicted are assumed to be generated according to the joint distribution

$$P_{W,Z^n,Z}=P_W\Big(\prod_{i=1}^{n}P_{Z_i|W}\Big)P_{Z|W}, \tag{1}$$

where $P_{Z_i|W}=P_{Z|W}$ for each i. As an example of the above model, the *predictive modeling* framework, a.k.a. probabilistic discriminative model [1], further assumes that $P_{Z|W}$ factors as $P_{Z|W}=P_{X|W}K_{Y|X,W}$, with some probability transition kernel $K_{Y|X,W}$ directly describing the true predictive distribution of the quantity of interest given the observation and model parameters. It is often further assumed that X is independent of W under the predictive modeling framework. Note that the above models and the following definitions encompass the unsupervised learning problem as well, where one just ignores the observations (X^n,X) so that $Z_i=Y_i$ and $Z=Y$.

Under the generative model (1), the *Bayesian learning* problem can be phrased as a Bayes decision problem of predicting Y based on X and the labeled observations Z^n. Given an action space A and a loss function $\ell:\mathsf{Y}\times\mathsf{A}\to\mathbb{R}$, a *decision rule* $\psi:\mathsf{X}\times\mathsf{Z}^n\to\mathsf{A}$ that maps observations to an action is sought to make the expected loss $\mathbf{E}[\ell(Y,\psi(X,Z^n))]$ small. A decision rule that minimizes the expected loss among all decision rules is called a *Bayes decision rule*. The corresponding minimum expected loss is defined as the *Bayes risk* in Bayesian learning:

Definition 1: In Bayesian learning, the Bayes risk with respect to a loss function ℓ is defined as

$$R_\ell(Y|X,Z^n):=\inf_{\psi:\mathsf{X}\times\mathsf{Z}^n\to\mathsf{A}}\mathbf{E}[\ell(Y,\psi(X,Z^n))], \tag{2}$$

where the infimum is taken over all decision rules such that the above expectation is defined.

B. A Data Processing Inequality for Bayes Risk

To better understand the definition of $R_\ell(Y|X,Z^n)$, we give a brief review of the general definition of the Bayes risk and prove a useful property of it. Given a random element Y of Y, the quantity

$$R_\ell(Y):=\inf_{a\in\mathsf{A}}\mathbf{E}[\ell(Y,a)] \tag{3}$$

is known as the *Bayes envelope* [16] or the *generalized entropy* [29] of Y. Given a random element V of some space V jointly distributed with Y, the general definition of the *Bayes risk*

$$R_\ell(Y|V):=\inf_{\psi:\mathsf{V}\to\mathsf{A}}\mathbf{E}[\ell(Y,\psi(V))] \tag{4}$$

is the minimum expected loss of predicting Y based on V. It can be expressed as the expectation of the *conditional Bayes envelope* $R_\ell(Y|V=v):=\inf_{a\in\mathsf{A}}\mathbf{E}[\ell(Y,a)|V=v]$ with respect to V, as $R_\ell(Y|V)=\int_{\mathsf{V}}P_V(\mathrm{d}v)R_\ell(Y|V=v)$. The Bayes risk $R_\ell(Y|V)$ can thus be viewed as a *generalized conditional entropy* of Y given V [30], [31]. The following lemma states that the Bayes risk satisfies a *data processing inequality*.

Lemma 1: Suppose the random variables U, V and Y form a Markov chain $U-V-Y$; in other words, Y and U are conditionally independent given V. Then, for any loss function ℓ, the Bayes risk of predicting Y from U is at least as large as the Bayes risk of predicting Y from V, i.e.,

$$R_\ell(Y|U)\geq R_\ell(Y|V). \tag{5}$$

Proof: Let ψ be a Bayes decision rule for predicting Y from U. Upon observing V, a random variable U' can be sampled from $P_{U|V}$, conditionally independent of (U,Y) given V. Then $\psi(U')$ serves as a randomized prediction of Y from V. As all probability spaces under consideration are Borel spaces, the sampling of U' conditional on V can be realized by a function $f:\mathsf{V}\times[0,1]\to\mathsf{U}$ of V and an independent random variable T uniformly distributed on $[0,1]$, such that $P_{f(V,T)|V}=P_{U|V}$ [32, Lemma 3.22]. We have

$$\begin{aligned}
R_\ell(Y|V)&\leq\inf_{t\in[0,1]}\mathbf{E}[\ell(Y,\psi(f(V,t)))] &(6)\\
&\leq\mathbf{E}[\ell(Y,\psi(f(V,T)))] &(7)\\
&=\mathbf{E}[\ell(Y,\psi(U'))] &(8)\\
&=\mathbf{E}[\ell(Y,\psi(U))] &(9)\\
&=R_\ell(Y|U) &(10)
\end{aligned}$$

where (6) is due to the definition of $R_\ell(Y|V)$ and the fact that $\psi(f(\cdot,t))$ is a map from V to Y for each $t\in[0,1]$; (7) follows from the independence between T and (V,Y); (9) follows from the fact that $P_{U'|V}=P_{U|V,Y}$ due to the Markov

chain $U - V - Y$, hence $P_{U',V,Y} = P_{U,V,Y}$; and (10) follows from the definition of ψ. □

In view of the definition of μ-entropy in (14) and (15) in Section III-A, the classic data processing inequality for mutual information stating that $I(U;Y) \leq I(V;Y)$ in a Markov chain $U - V - Y$ [33] can be derived from Lemma 1 applied to the log loss. The data processing inequality for MMSE proved in [34] can also be derived from Lemma 1 applied to the quadratic loss. More importantly, Lemma 1 extends the *value of information principle* in Bayes decision making [30], which states that $R_\ell(Y) \geq R_\ell(Y|V)$, as it can be viewed as a special case of Lemma 1 when U is independent of (V, Y). Lemma 1 also extends the *principle of total evidence* [35], a.k.a. the *value of knowledge theorem* [36], which states that $R_\ell(Y|V_1) \geq R_\ell(Y|V_1, V_2)$ for arbitrary random variables V_1 and V_2 jointly distributed with Y. While the original argument in [35] overlooked the randomness of V_1, this principle can be rigorously justified by Lemma 1 as $V_1 - (V_1, V_2) - Y$ always form a Markov chain. It is also apparent from Lemma 1 or its proof that randomizing the decision rule does not help to decrease the expected loss in Bayes decision making, as $(T, V) - V - Y$ form a Markov chain for any independent random variable T to be used in the randomized decision rule.

C. Definition of Minimum Excess Risk

An immediate consequence of Lemma 1 in Bayesian learning is that the Bayes risk $R_\ell(Y|X, Z^n)$ decreases as the data size n increases, as $(X, Z^n) - (X, Z^{n+1}) - Y$ form a Markov chain. A special case of this result for linear regression with quadratic loss appears in [37]. While $R_\ell(Y|X, Z^n)$ decreases in n, it will not necessarily vanish as $n \to \infty$. We define the *fundamental limit of the Bayes risk* as the minimum expected loss when the model parameters W are known, which is attained by some "omniscient" decision rule $\Psi : \mathsf{X} \times \mathsf{W} \to \mathsf{A}$ that can directly access the model parameters.

Definition 2: In Bayesian learning, the fundamental limit of the Bayes risk with respect to a loss function ℓ is defined as

$$R_\ell(Y|X, W) = \inf_{\Psi:\mathsf{X}\times\mathsf{W}\to\mathsf{A}} \mathbf{E}[\ell(Y, \Psi(X, W))]. \tag{11}$$

For any feasible decision rule $\psi : \mathsf{X} \times \mathsf{Z}^n \to \mathsf{A}$, we can define its *excess risk* as the gap between its expected loss $\mathbf{E}[\ell(Y, \psi(X, Z^n))]$ and $R_\ell(Y|X, W)$. In this work, our interest is in the gap between the Bayes risk $R_\ell(Y|X, Z^n)$ and its fundamental limit $R_\ell(Y|X, W)$, which is the minimum achievable excess risk among all feasible decision rules:

Definition 3: The minimum excess risk (MER) with respect to a loss function ℓ is defined as

$$\mathrm{MER}_\ell := R_\ell(Y|X, Z^n) - R_\ell(Y|X, W). \tag{12}$$

The MER defined above is an algorithm-independent quantity. It quantifies the regret of the best decision rule that has access to data, but not to model parameters, relative to the best "omniscient" decision rule. It thus reflects the difficulty of the learning problem, which comes from the lack of knownedge of W. This is better illustrated by decomposing the Bayes risk as

$$R_\ell(Y|X, Z^n) = R_\ell(Y|X, W) + \mathrm{MER}_\ell. \tag{13}$$

If we view the Bayes risk as a measure of the minimum prediction uncertainty, this decomposition allows us to give formal definitions of the "aleatoric" uncertainty and the minimum "epistemic" uncertainty [10]. The first term, the fundamental limit of the Bayes risk, can be viewed as the aleatoric part of the minimum prediction uncertainty, which exists even when the model parameters are known. The second term, the MER, can be viewed as the epistemic part of the minimum prediction uncertainty, which is due to the lack of knowledge of W. In [11], a decomposition of uncertainty is proposed for the log loss and the quadratic loss, where the epistemic uncertainty is defined as $R_\ell(Y|X) - R_\ell(Y|X, W)$ when expressed by our notation; however, this definition does not take the observed data into consideration, thus does not reflect the intuitive expectation that the epistemic uncertainty should decrease as the data size increases [12]. On the contrary, the advantage of defining the minimum epistemic uncertainty as the MER is that the uncertainty becomes smaller as more data is observed, as asserted by the following result.

Theorem 1: For any loss function, MER_ℓ decreases in the data size n, and $\mathrm{MER}_\ell \geq 0$ for all n.

Proof: The claim that MER_ℓ decreases in n is due to the previously justified fact that $R_\ell(Y|X, Z^n)$ decreases in n as a consequence of Lemma 1. The claim that $\mathrm{MER}_\ell \geq 0$ is due to the Markov chain $(X, Z^n) - (X, W) - Y$ and Lemma 1. □

Intuitively, we expect that $\mathrm{MER}_\ell \downarrow 0$ as $n \to \infty$. However, except for the log loss, there are few results in the literature regarding this convergence in the general case, or regarding how the MER depends on the estimation error of the model parameters. In the following two sections, we use different methods to derive upper bounds on MER for general loss functions. We show that, in many cases, the MER can be upper-bounded either in terms of the conditional mutual information $I(W; Y|X, Z^n)$, or in terms of the minimum achievable estimation error of W from (X, Z^n). These results reflect how the MER depends on the joint distribution in (1), in particular on $P_{Z|W}$ and P_W, as well as on the loss function and the data size.

III. Upper Bounds via Conditional Mutual Information

The first method for upper-bounding the MER is to relate it to $I(W; Y|X, Z^n)$, which can be further bounded by $\frac{1}{n} I(W; Y^n|X^n)$ or $\frac{1}{n} I(W; Z^n)$. In many cases, it can be shown that $I(W; Z^n)$ is sublinear in n [38]–[40], which implies that the MER converges to zero as $n \to \infty$.

A. Logarithmic Loss

We first consider the setting where one makes "soft" predictions, such that the action space is the collection of all probability densities q with respect to a common σ-finite positive measure μ on Y. The log loss $\ell(y, q) := -\log q(y)$ penalizes those densities that assign small probabilities to the

outcome y. Based on the definitions in (3) and (4), it can be shown that

$$R_{\log}(Y) = H_\mu(Y) := -\int_{\mathsf{Y}} p_Y(y) \log p_Y(y) \mu(\mathrm{d}y) \tag{14}$$

and

$$R_{\log}(Y|V) = H_\mu(Y|V) := -\int_{\mathsf{V}} P_V(\mathrm{d}v) \int_{\mathsf{Y}} p_{Y|v}(y) \log p_{Y|v}(y) \mu(\mathrm{d}y), \tag{15}$$

which can be viewed as the μ-*entropy* of Y and the *conditional* μ-*entropy* of Y given V, and the optimal actions are the unconditional density p_Y and the conditional density $p_{Y|v}$ with respect to μ, respectively. For instance, if Y is discrete and μ is the counting measure, then $R_{\log}(Y) = H(Y)$ and $R_{\log}(Y|V) = H(Y|V)$ are the Shannon and the conditional Shannon entropy; while if $\mathsf{Y} = \mathbb{R}^p$ and μ is the Lebesgue measure, then $R_{\log}(Y) = h(Y)$ and $R_{\log}(Y|V) = h(Y|V)$ are the differential and the conditional differential entropy. (See [33] for further background on information theory.) With these definitions, the MER for the log loss is the difference between two μ-entropy terms:

$$\mathrm{MER}_{\log} = H_\mu(Y|X, Z^n) - H_\mu(Y|X, W). \tag{16}$$

A key observation is that $\mathrm{MER}_{\log}$ can be expressed in terms of the conditional mutual information:

Lemma 2: For the log loss,

$$\mathrm{MER}_{\log} = I(W; Y|X, Z^n). \tag{17}$$

Proof: The claim follows from the fact that $I(W; Y|X, Z^n) = H_\mu(Y|X, Z^n) - H_\mu(Y|X, W, Z^n)$ and that $H_\mu(Y|X, W, Z^n) = H_\mu(Y|X, W)$. The second fact is due to the Markov chain $(X, W, Z^n) - (X, W) - Y$ encoded in (1) and the definition of the conditional μ-entropy. □

Equation (17) states that $\mathrm{MER}_{\log}$ is the average reduction of the uncertainty about Y that comes from the knowledge of W, given that (X, Z^n) is already known. With this representation, using the conditional independence structure in (1) and the data processing inequality in Lemma 1 applied to the μ-entropy, we have:

Theorem 2: The MER with respect to the log loss can be upper-bounded as

$$I(W; Y|X, Z^n) \le \frac{1}{n} I(W; Y^n|X^n). \tag{18}$$

Proof: For $i = 1, \ldots, n-1$, we have

$$\begin{aligned}
&I(W; Y_i|X^n, Y^{i-1}) \\
&= H_\mu(Y_i|X^n, Y^{i-1}) - H_\mu(Y_i|W, X^n, Y^{i-1}) && (19)\\
&= H_\mu(Y_{i+1}|X^n, Y^{i-1}) - H_\mu(Y_{i+1}|W, X^n, Y^{i-1}) && (20)\\
&\ge H_\mu(Y_{i+1}|X^n, Y^i) - H_\mu(Y_{i+1}|W, X^n, Y^i) && (21)\\
&= I(W; Y_{i+1}|X^n, Y^i) && (22)
\end{aligned}$$

where (19) is due to the definitions of the conditional mutual information and the conditional μ-entropy in (15); (20) follows from the fact that $(W, X^n, Y^{i-1}, Y_i) \overset{\mathrm{d.}}{=} (W, X^n, Y^{i-1}, Y_{i+1})$[1]; and (21) follows from the fact that $H_\mu(Y_{i+1}|X^n, Y^{i-1}) \ge H_\mu(Y_{i+1}|X^n, Y^i)$ due to Lemma 1, and the fact that $H_\mu(Y_{i+1}|W, X^n, Y^{i-1}) = H_\mu(Y_{i+1}|W, X^n, Y^i) = H_\mu(Y_{i+1}|W, X_{i+1})$ as Y_{i+1} is conditionally independent of everything else given (W, X_{i+1}).

Then, from the chain rule of mutual information,

$$\begin{aligned}
&I(W; Y^n|X^n) \\
&= \sum_{i=1}^n I(W; Y_i|X^n, Y^{i-1}) && (23)\\
&\ge nI(W; Y_n|X^n, Y^{n-1}) && (24)\\
&= nI(W; Y|X, Z^{n-1}) && (25)\\
&= n\big(H_\mu(Y|X, Z^{n-1}) - H_\mu(Y|W, X, Z^{n-1})\big) && (26)\\
&\ge n\big(H_\mu(Y|X, Z^n) - H_\mu(Y|W, X, Z^n)\big) && (27)\\
&= nI(W; Y|X, Z^n) && (28)
\end{aligned}$$

where (24) is obtained by repeated application of (22); (25) is due to the fact that $(W, Z^{n-1}, Z_n) \overset{\mathrm{d.}}{=} (W, Z^{n-1}, Z)$; and (27) follows from Lemma 1 and the fact that Y is conditionally independent of everything else given (W, X). The claim follows from (28). □

Theorem 2 can be weakened to the following corollary by the fact that $I(W; Y^n|X^n) = I(W; Z^n) - I(W; X^n)$. There is no slack when X is independent of W.

Corollary 1: The MER with respect to the log loss can be upper-bounded as

$$I(W; Y|X, Z^n) \le \frac{1}{n} I(W; Z^n). \tag{29}$$

Upon maximizing over P_W on both side of (29), Corollary 1 is reminiscent of the *redundancy-capacity theorem* in universal source coding in the Bayesian setting [15], [16], where the quantity of interest is the minimum *overall* redundancy $\min_Q \mathbf{E}_{P_{W,Z^n}}[-\log Q(Z^n) + \log P_{Z^n|W}(Z^n|W)]$, which can be shown to be $I(W; Z^n)$. Therefore, from the source coding point of view, $\mathrm{MER}_{\log}$ in (17) may be interpreted as the minimum *instantaneous* redundancy of encoding a fresh sample when n data samples are observed, which is shown to be smaller than the normalized minimum overall redundancy by Corollary 1. More generally, the mutual information $I(W; Z^n)$ is also known to be the minimum *accumulated* excess risk for the log loss in Bayesian sequential prediction [16], Bayesian density estimation [17], and Bayesian supervised learning [18]. The non-asymptotic relationships between the instantaneous $\mathrm{MER}_{\log}$ and the accumulated excess risks shown in Theorem 2 and Corollary 1 hold for general model $P_{Z|W}$ and prior P_W, and allow us to quantify the rate of convergence of $\mathrm{MER}_{\log}$ by upper-bounding $I(W; Y^n|X^n)$ or $I(W; Z^n)$.

From the results of [38]–[40], if W is a d-dimensional compact subset of $\mathbb{R}^d$ and the model $P_{Z|w}$ is sufficiently smooth in w (see Appendix A.1 for rigorous statements of

[1]For random variables U and V, $U \overset{\mathrm{d.}}{=} V$ means that U and V have the same distribution.

these conditions), then

$$I(W;Z^n) = \frac{d}{2}\log\frac{n}{2\pi e} + h(W) + \frac{1}{2}\mathbf{E}\big[\log\det J_{Z|W}\big] + o(1) \quad \text{as } n\to\infty, \tag{30}$$

where $h(W)$ is the differential entropy of W, and, as a functional of $P_{Z|w}$, $J_{Z|w}$ is the Fisher information matrix about w contained in Z with respect to $P_{Z|w}$, and the expectation is taken with respect to P_W. Due to the logarithmic dependence on n in (30) and the chain rule of mutual information, it can be shown that the instantaneous mutual information under the same conditions satisfies $I(W;Z|Z^n) = O(d/n)$ as $n\to\infty$. This gives us a refined asymptotic upper bound on $\mathrm{MER}_{\log}$ whenever (30) holds than directly applying (30) to Corollary 1:

Theorem 3: Under the regularity conditions listed in Section A under which (30) holds, we have

$$I(W;Y|X,Z^n) = O\Big(\frac{d}{2n}\Big) \quad \text{as } n\to\infty. \tag{31}$$

Proof: The proof relies on [41, Lemma 6] which is stated as Lemma A3 in Appendix A.3. Suppose $(a_1,a_2,\ldots)$ and $(b_1,b_2,\ldots)$ are two sequences of real numbers such that $a_n = \sum_{i=1}^n b_i$ for all n. Lemma A3 states that, if $\lim_{n\to\infty} a_n/\log n$ and $\lim_{n\to\infty} nb_n$ exist, then they are equal. With this result and the chain rule of mutual information, we know that whenever (30) holds,

$$\lim_{n\to\infty}(n+1)I(W;Z|Z^n) = \lim_{n\to\infty}\frac{I(W;Z^n)}{\log n} = \frac{d}{2}. \tag{32}$$

The claim follows from the fact that $I(W;Y|X,Z^n) \le I(W;Z|Z^n)$. □

As we show next, the representation of $\mathrm{MER}_{\log}$ via the conditional mutual information in (17) and the resulting upper bounds derived in this subsection can be used to obtain upper bounds on the MER for other loss functions as well.

B. Quadratic Loss

While the log loss is naturally used for assessing "soft" predictions, it is also a common practice to make "hard" predictions, e.g., the actions can be elements in Y. When $\mathsf{Y}=\mathsf{A}=\mathbb{R}$, a commonly used loss function is the quadratic loss $\ell(y,a)=(y-a)^2$. For any V that statistically depends on Y, the conditional Bayes envelope with respect to the quadratic loss is $R_2(Y|V=v) = \mathsf{var}[Y|v]$, the optimal action is the conditional mean $\mathbf{E}[Y|v]$, and the corresponding Bayes risk

$$R_2(Y|V) = \mathbf{E}[\mathsf{var}[Y|V]] \tag{33}$$

is the minimum mean square error (MMSE) of estimating Y from V. In this case, the MER in Bayesian learning turns out to be

$$\mathrm{MER}_2 = \mathbf{E}\big[\mathsf{var}[Y|X,Z^n]\big] - \mathbf{E}\big[\mathsf{var}[Y|X,W]\big]. \tag{34}$$

More generally, when $\mathsf{Y}=\mathsf{A}=\mathbb{R}^p$ and $\ell(y,a)=\|y-a\|^2$ with $\|\cdot\|$ denoting the l_2 norm, the MER in this case is

$$\mathrm{MER}_2 = \mathbf{E}\big[\|Y-\mathbf{E}[Y|X,Z^n]\|^2\big] - \mathbf{E}\big[\|Y-\mathbf{E}[Y|X,W]\|^2\big] \tag{35}$$

$$= \mathbf{E}\big[\|\mathbf{E}[Y|X,Z^n]-\mathbf{E}[Y|X,W]\|^2\big], \tag{36}$$

where the second equality follows from the fact that $\mathbf{E}[Y|X,W] = \mathbf{E}[Y|X,W,Z^n]$ and the orthogonality principle in MMSE estimation [42].

Under the assumption that $\|Y\|\le b$, using a result that connects MMSE difference to conditional mutual information [34, Theorem 10], we can upper-bound MER_2 in terms of $I(W;Y|X,Z^n)$:

Theorem 4: If $\mathsf{Y}=\{y\in\mathbb{R}^p : \|y\|\le b\}$ for some $b>0$, then for the quadratic loss,

$$\mathrm{MER}_2 \le 2b^2 I(W;Y|X,Z^n). \tag{37}$$

Proof: [34, Theorem 10] states that if $\|Y\|\le b$, then for any (U,V) jointly distributed with Y,

$$R_2(Y|U) - R_2(Y|U,V) \le 2\,b^2 I(V;Y|U). \tag{38}$$

Using this result and the fact that $R_2(Y|X,W) = R_2(Y|X,W,Z^n)$, we obtain (37). □

With Theorem 4, all the upper bounds on $\mathrm{MER}_{\log}$ derived in Section III-A can be used to further upper-bound MER_2. In particular, whenever (30) holds, we have $\mathrm{MER}_2 = O(b^2 d/n)$ as $n\to\infty$.

C. Zero-One Loss

Another loss function we consider for hard predictions is the zero-one loss $\ell(y,a)=\mathbf{1}\{y\ne a\}$ with $\mathsf{Y}=\mathsf{A}$. For any V that statistically depends on Y, the conditional Bayes envelope with respect to the zero-one loss is $R_{01}(Y|V=v) = 1-\max_{y\in\mathsf{Y}} P_{Y|v}(y)$, the optimal action is the conditional mode $\arg\max_{y\in\mathsf{Y}} P_{Y|v}(y)$, and the corresponding Bayes risk is

$$R_{01}(Y|V) = 1 - \mathbf{E}[\max_{y\in\mathsf{Y}} P_{Y|V}(y)], \tag{39}$$

with expectation taken with respect to V. The MER for the zero-one loss is

$$\mathrm{MER}_{01} = \mathbf{E}[\max_{y\in\mathsf{Y}} P_{Y|X,W}(y)] - \mathbf{E}[\max_{y\in\mathsf{Y}} P_{Y|X,Z^n}(y)], \tag{40}$$

where the expectations are taken with respect to the conditioning variables. In this case, as the loss function takes values in $[0,1]$, Theorem 6 stated in the next subsection gives an upper bound for MER_{01} in terms of $I(W;Y|X,Z^n)$:

Corollary 2: For the zero-one loss,

$$\mathrm{MER}_{01} \le \sqrt{\frac{1}{2}I(W;Y|X,Z^n)}. \tag{41}$$

From Theorem 3, we know that whenever (30) holds, $\mathrm{MER}_{01} = O(\sqrt{d/n})$ as $n\to\infty$.

For the special case of binary classification, where $\mathsf{Y}=\{0,1\}$, the Bayes risk $R_{01}(Y|X,Z^n)$ is studied in [19] and is upper-bounded in terms of $H(Y^n|X^n)$. When the model is realizable, that is, when $Y=g(X,W)$ with some generative function $g:\mathsf{X}\times\mathsf{W}\to\{0,1\}$, it is also observed in [19] that $H(Y^n|X^n)$ can be further upper-bounded in terms of the *VC dimension* of the generative function class $\mathcal{G}=\{g(\cdot,w):\mathsf{X}\to\{0,1\}, w\in\mathsf{W}\}$, defined as

$$V(\mathcal{G}) := \sup\Big\{n\in\mathbb{N} : \sup_{x^n\in\mathsf{X}^n}\big|\{(g(x_1,w),\ldots,g(x_n,w)), w\in\mathsf{W}\}\big| = 2^n\Big\}. \tag{42}$$

The Sauer-Shelah lemma [43], [44] states that, if $V(\mathcal{G}) = d$, then for all $x^n \in \mathsf{X}^n$,

$$\big|\{(g(x_1,w),\ldots,g(x_n,w)), w \in \mathsf{W}\}\big| \le \sum_{k=1}^{d} \binom{n}{k} \le en^d. \tag{43}$$

As $\mathrm{MER}_{01} \le R_{01}(Y|X,Z^n)$, the results in [19] lead to the following MER upper bounds.

Theorem 5: If $\mathsf{Y} = \{0,1\}$, then

$$\mathrm{MER}_{01} \le \frac{1}{2\log 2} H(Y|X,Z^n) \le \frac{1}{2n\log 2} H(Y^n|X^n). \tag{44}$$

Moreover, if $Y = g(X,W)$ with some function $g : \mathsf{X}\times\mathsf{W} \to \mathsf{Y}$, and the function class $\mathcal{G} = \{g(\cdot,w) : \mathsf{X} \to \mathsf{Y}, w \in \mathsf{W}\}$ has VC dimension d, then

$$\mathrm{MER}_{01} \le O\Big(\frac{d}{2n\log 2}\Big) \quad \text{as } n \to \infty. \tag{45}$$

These upper bounds also hold for $\frac{1}{2\log 2}\mathrm{MER}_{\log}$ in the same settings.

Proof: The proof of (44) is essentially drawn from [19]. Using our notation,

$$\begin{aligned}
\mathrm{MER}_{01} &\le R_{01}(Y|X,Z^n) && (46)\\
&= \mathbf{E}\Big[\min_{y\in\{0,1\}} \mathbb{P}[Y=y|X,Z^n]\Big] && (47)\\
&\le \mathbf{E}\Big[\frac{1}{2\log 2} h_2\big(\mathbb{P}[Y=1|X,Z^n]\big)\Big] && (48)\\
&= \frac{1}{2\log 2} H(Y|X,Z^n) && (49)\\
&\le \frac{1}{2n\log 2} H(Y^n|X^n), && (50)
\end{aligned}$$

where (46) follows from the fact that $R_{01}(Y|X,W) \ge 0$; (47) follows from (39) and the assumption that $\mathsf{Y} = \{0,1\}$; (48) follows from the fact that $\min\{p, 1-p\} \le \frac{1}{2\log 2} h_2(p)$ for $p \in [0,1]$, where $h_2(\cdot)$ is the binary entropy function; and (50) can be proved by the chain rule of Shannon entropy and the fact that $H(Y_i|X^n, Y^{i-1})$ decreases as i increases, similar to the proof of Theorem 2.

The proof of (45) relies on the observation made in [19] that $H(Y^n|X^n) \le d\log n + 1$ under a realizable model whenever the VC dimension of $\mathcal{G}$ is d, which is due to the Sauer-Shelah lemma (43). Additionally, from $H(Y^n|X^n) = \sum_{i=1}^n H(Y_i|X^n, Y^{i-1})$ and Lemma A3, we have

$$\lim_{n\to\infty} (n+1) H(Y|X,Z^n) = \lim_{n\to\infty} \frac{H(Y^n|X^n)}{\log n} \le d \tag{51}$$

whenever these limits exist, which proves (45).

The upper bounds also hold for $\frac{1}{2\log 2}\mathrm{MER}_{\log}$ because $\mathrm{MER}_{\log} \le H(Y|X,Z^n)$, as $H(Y|X,W) \ge 0$ when Y is discrete. □

In Section III-E, we discuss the MER under realizable models in more general settings, where the results go beyond binary classification and zero-one loss.

D. General Loss Functions

Now we derive a general upper bound for the MER with respect to a wide range of loss functions. For an arbitrary loss function $\ell : \mathsf{Y}\times\mathsf{A} \to \mathbb{R}$, let $\Psi^* : \mathsf{X}\times\mathsf{W} \to \mathsf{Y}$ be the optimal omniscient decision rule such that $\mathbf{E}[\ell(Y,\Psi^*(X,W))] = R_\ell(Y|X,W)$. Given (X,Z^n), let W' be a sample from the posterior distribution $P_{W|X,Z^n}$ conditionally independent of everything else given (X,Z^n). Then the MER can be upper-bounded by

$$\mathrm{MER}_\ell \le \mathbf{E}[\ell(Y,\Psi^*(X,W'))] - \mathbf{E}[\ell(Y,\Psi^*(X,W))]. \tag{52}$$

Here, $\Psi^*(X,W')$ is a plug-in decision rule, where we first estimate W by W' from (X,Z^n), and then plug W' in Ψ^* to predict Y given X. The right side of (52) is the excess risk of this plug-in decision rule. Under regularity conditions on the moment generating function of $\ell(Y,\Psi^*(X,W'))$ under the conditional distribution $P_{Y,W'|X,Z^n}$, we have the following upper bound on MER_ℓ in terms of $I(\Psi^*(X,W);Y|X,Z^n)$.

Theorem 6: Assume there is a function $\varphi(\lambda)$ defined on $[0,b)$ for some $b \in (0,\infty]$, such that

$$\log \mathbf{E}_{x,z^n}\Big[\exp\Big\{-\lambda\Big(\ell(Y,\Psi^*(x,W'))- \mathbf{E}_{x,z^n}\big[\ell(Y,\Psi^*(x,W'))\big]\Big)\Big\}\Big] \le \varphi(\lambda) \tag{53}$$

for all $0 \le \lambda < b$ and all (x,z^n), where $\mathbf{E}_{x,z^n}[\cdot]$ denotes the conditional expectation with respect to (Y,W') given $(X,Z^n) = (x,z^n)$. Then

$$\mathrm{MER}_\ell \le \varphi^{*-1}\left(I(\Psi^*(X,W);Y|X,Z^n)\right), \tag{54}$$

where $\varphi^*(\gamma) := \sup_{0\le\lambda<b}\{\lambda\gamma - \varphi(\lambda)\}$, $\gamma \in \mathbb{R}$, is the Legendre dual of φ, and $\varphi^{*-1}(u) := \sup\{\gamma \in \mathbb{R} : \varphi^*(\gamma) \le u\}$, $u \in \mathbb{R}$, is the generalized inverse of φ^*. In addition, if $\varphi(\lambda)$ is strictly convex over $(0,b)$ and $\varphi(0) = \varphi'(0) = 0$, then $\lim_{x\downarrow 0}\varphi^{*-1}(x) = 0$.

Proof: We have the following chain of inequalities:

$$\begin{aligned}
&\mathrm{MER}_\ell\\
&\le \mathbf{E}[\ell(Y,\Psi^*(X,W'))] - \mathbf{E}[\ell(Y,\Psi^*(X,W))] && (55)\\
&= \mathbf{E}\big[\mathbf{E}[\ell(Y,\Psi^*(X,W')) - \ell(Y,\Psi^*(X,W))|X,Z^n]\big]\\
&\le \mathbf{E}\big[\varphi^{*-1}\big(D_{\mathrm{KL}}(P_{Y,\Psi^*(X,W)|X,Z^n}\| P_{Y,\Psi^*(X,W')|X,Z^n})\big)\big] && (56)\\
&= \varphi^{*-1}\big(\mathbf{E}\big[D_{\mathrm{KL}}(P_{Y,\Psi^*(X,W)|X,Z^n}\| P_{Y,\Psi^*(X,W')|X,Z^n})\big]\big) && (57)\\
&= \varphi^{*-1}\big(I(\Psi^*(X,W);Y|X,Z^n)\big) && (58)
\end{aligned}$$

where (56) follows from the assumption (53) in the statement of the theorem and Lemma A2 stated in Appendix A.2 applied to $P = P_{Y,\Psi^*(x,W)|x,z^n}$ and $Q = P_{Y,\Psi^*(x,W')|x,z^n}$, and the expectation is taken with respect to (X,Z^n); (57) follows from the concavity of φ^{*-1}, which is due to the convexity of φ^*, and Jensen's inequality; (58) follows from the fact that W' is conditionally i.i.d. of W and conditionally independent of Y given (X,Z^n). The last claim of the theorem comes from the fact that under the assumptions on φ, its Legendre dual φ^* is increasing on $[0,\infty)$ and continuous at 0 and so is the inverse φ^{*-1}. □

An example for the condition in (53) to hold is when the random variable $\ell(Y, \Psi^*(x, W'))$ is σ^2-*subgaussian*[2] conditionally on $(X, Z^n) = (x, z^n)$. In this case, (53) holds with $b = \infty$ and $\varphi(\lambda) = \sigma^2\lambda^2/2$, and we have the following corollary.

Corollary 3: If $\ell(Y, \Psi^*(x, W'))$ is σ^2-*subgaussian* conditionally on $(X, Z^n) = (x, z^n)$ for all (x, z^n), then

$$\mathrm{MER}_\ell \le \sqrt{2\sigma^2 I(\Psi^*(X, W); Y|X, Z^n)}. \tag{59}$$

Using the fact that if $\ell(\cdot, \cdot) \in [a, b]$ then ℓ is $(b-a)^2/4$-subgaussian under any distribution of the arguments, Corollary 3 can provide upper bound for the MER under any bounded loss functions. More generally, Theorem 6 can be applied in the situation where the loss function is unbounded and non-subgaussian. In Appendix B, we present such a case where an MER upper bound for the quadratic loss in linear regression is derived based on Theorem 6.

From the data processing inequality of mutual information,

$$I(\Psi^*(X, W); Y|X, Z^n) \le I(W; Y|X, Z^n). \tag{60}$$

Since φ^{*-1} defined in Theorem 6 is an increasing function on $[0, \infty)$, the upper bounds in (54) and (59) can be weakened by replacing $I(\Psi^*(X, W); Y|X, Z^n)$ with $I(W; Y|X, Z^n)$ or any of its upper bounds derived in Section III-A. In particular, when (30) and Theorem 3 hold in addition with the assumption in Theorem 6, we have $\mathrm{MER}_\ell = O(\varphi^{*-1}(d/2n))$ as $n \to \infty$.

Theorem 6 also provides a connection between the MER and the mutual information between the *observed data* and the *learned model parameters*. If X is independent of W, then $P_{W|X,Z^n} = P_{W|Z^n}$, and (W', Z^n) have the same joint distribution as (W, Z^n). In this case, when the condition in Corollary 3 is satisfied, upper-bounding $I(W; Y|X, Z^n)$ in (59) by $\frac{1}{n} I(W; Z^n)$ according to Corollary 1 leads to the following result.

Corollary 4: If X is independent of W in addition to the condition in Corollary 3, then

$$\mathrm{MER}_\ell \le \sqrt{\frac{2\sigma^2}{n} I(Z^n; W')}, \tag{61}$$

where $I(Z^n; W')$ is the mutual information between the data and the learned model parameters sampled from the posterior distribution $P_{W|Z^n}$.

Corollary 4 is an analogue of the generalization-information relationship in the frequentist learning [14, Theorem 1], where it is shown that the generalization error in frequentist learning can be upper-bounded in terms of the mutual information between the *observed data* and the *learned hypothesis*. From (54), we also know that when the more general condition in Theorem 6 is satisfied, we have $\mathrm{MER}_\ell \le \varphi^{*-1}(\frac{1}{n} I(Z^n; W'))$, which is analogous to upper bounds on the generalization error in [45], [46].

E. Realizable Models and Connection to VC Dimension

In Section III-C we have presented the MER for the zero-one loss under the realizable model of binary classification. Here, we present a few results on the MER for general loss functions under general realizable models. These results provide tighter asymptotic MER bounds under realizable models than directly using the results obtained in the previous subsection. Following the observations made in [19], these results also show how the key quantities in classical frequentist learning theory, notably the Vapnik-Chervonenkis (VC) dimension, can be naturally brought into the MER analysis in Bayesian learning through the information-theoretic framework proposed in this work.

A realizable model is a model where the quantity of interest Y is determined by the observation X and the model parameters W through a *generative function* $g : \mathsf{X} \times \mathsf{W} \to \mathsf{Y}$. Under a realizable model, $g(X, W')$ can serve as a plug-in decision rule, where W' is a sample from the posterior distribution $P_{W|X,Z^n}$, conditionally independent of everything else given (X, Z^n). It is observed in a follow-up work [47] (which has appeared after the initial version of this paper was posted) that the generalization error bounds developed in [48] for the realizable setting of frequentist learning can be adapted to MER bounds for realizable models in Bayesian learning. In particular, [47, Lemma 3] shows that for a loss function $\ell \in [0, b]$, if $R_\ell(Y|X, W) = 0$, then $\mathrm{MER}_\ell \le 3b I(W; Y|X, Z^n)$. Following this approach, the next result provides an upper bound for the MER under realizable models, with a better prefactor and a tighter conditional mutual information term.

Theorem 7: For a loss function $\ell \in [0, b]$, if there exists a function $g : \mathsf{X} \times \mathsf{W} \to \mathsf{Y}$ such that $\ell(Y, g(X, W)) = 0$ almost surely with respect to the joint distribution $P_{W,X,Y}$, then

$$\mathrm{MER}_\ell \le \frac{b}{\log 2} I(g(X, W); Y|X, Z^n). \tag{62}$$

Proof: We have the following chain of inequalities:

$$\begin{aligned}
\mathrm{MER}_\ell &\le \mathbf{E}[\ell(Y, g(X, W'))] && (63)\\
&= \mathbf{E}\big[\mathbf{E}[\ell(Y, g(X, W'))|X, Z^n]\big] && (64)\\
&\le \int \frac{b}{\log 2} I(g(x, W); Y|X = x, Z^n = z^n)\\
&\qquad P_{X,Z^n}(\mathrm{d}x, \mathrm{d}z^n) && (65)\\
&= \frac{b}{\log 2} I(g(X, W); Y|X, Z^n), && (66)
\end{aligned}$$

where (63) follows from the assumption that $\ell(Y, g(X, W)) = 0$, the minimum loss, which implies that $R_\ell(Y|X, W) = 0$; (65) follows by applying Lemma 3 stated below to the joint distribution $P_{g(x,W),Y|x,z^n}$ for all (x, z^n) under P_{X,Z^n}. □

The following lemma used in the proof of Theorem 7 is adapted from [48, Theorem 5.7], where it is developed for bounding the generalization error in the realizable setting of frequentist learning.

Lemma 3: Let V and Y be jointly distributed random variables on $\mathsf{V} \times \mathsf{Y}$. Let V' be an independent copy of V, that is, $P_{V'} = P_V$ and V' is independent of (V, Y). For a function $\ell : \mathsf{Y} \times \mathsf{V} \to [0, b]$, if $\ell(Y, V) = 0$ almost surely with respect to $P_{V,Y}$, then

$$\mathbf{E}[\ell(Y, V')] \le \frac{b}{\log 2} I(V; Y). \tag{67}$$

Proof: We follow the symmetrization idea used in the proof of [48, Theorem 5.7]. Let $\tilde{V} = (V_0, V_1)$ with V_0 and

[2] A random variable U is σ^2-subgaussian if $\mathbf{E}[e^{\lambda(U - \mathbf{E}U)}] \le e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$.

V_1 being i.i.d. samples from P_V. Let S and S' be i.i.d. uniform Bernoulli random variables independent of $\tilde{V}$, with $\overline{S} = 1-S$, and $\overline{S'} = 1-S'$. With these random variables at hand, we can construct $V = \tilde{V}_S$, $V' = \tilde{V}_{\overline{S}}$, and Y to be jointly distributed with V and conditionally independent of everything else given V. Then, following the technique used in the proof of [48, Theorem 5.7], for any $u > 0$ and $t > 0$,

$$\begin{aligned}
&\mathbf{E}[\ell(Y,V')]\\
=&\mathbf{E}[\ell(Y,V')] - \mathbf{E}[u\ell(Y,V)] && (68)\\
=&\mathbf{E}[\ell(Y,\tilde{V}_{\overline{S}})] - \mathbf{E}[u\ell(Y,\tilde{V}_S)] && (69)\\
=&\mathbf{E}\big[\mathbf{E}[\ell(Y,\tilde{V}_{\overline{S}}) - u\ell(Y,\tilde{V}_S)|\tilde{V}]\big] && (70)\\
\le&\frac{1}{t}\Big(I(S;Y|\tilde{V}) +\\
&\mathbf{E}\Big[\log \mathbf{E}\big[\exp\big\{t\big(\ell(Y,\tilde{V}_{\overline{S'}}) - u\ell(Y,\tilde{V}_{S'})\big)\big\}|\tilde{V}\big]\Big]\Big) && (71)\\
=&\frac{1}{t}\Big(I(S;Y|\tilde{V}) + \mathbf{E}\Big[\log \mathbf{E}\big[\mathbf{E}\big[\exp\big\{t\big(\ell(Y,\tilde{V}_{\overline{S'}})\\
&\quad - u\ell(Y,\tilde{V}_{S'})\big)\big\}|Y,S,\tilde{V}\big]|\tilde{V}\big]\Big]\Big) && (72)\\
=&\frac{1}{t}\Big(I(S;Y|\tilde{V}) +\\
&\quad \mathbf{E}\Big[\log \mathbf{E}\big[\tfrac{1}{2}e^{t\ell(Y,\tilde{V}_{\overline{S}})} + \tfrac{1}{2}e^{-ut\ell(Y,\tilde{V}_{\overline{S}})}|\tilde{V}\big]\Big]\Big) && (73)
\end{aligned}$$

where (68) is due to the assumption that $\ell(Y,V) = 0$ almost surely; (71) is due to the Donsker-Varadhan theorem, which implies that

$$\begin{aligned}
D_{\mathrm{KL}}(P_{S,Y|\tilde{V}}\|P_{S',Y|\tilde{V}}) \ge\ &\mathbf{E}\big[t\big(\ell(Y,\tilde{V}_{\overline{S}}) - u\ell(Y,\tilde{V}_S)\big)\big|\tilde{V}\big]\\
&- \log \mathbf{E}\big[\exp\big\{t\big(\ell(Y,\tilde{V}_{\overline{S'}}) - u\ell(Y,\tilde{V}_{S'})\big)\big\}\big|\tilde{V}\big];
\end{aligned}$$

and (73) follows from the fact that S' is equally likely to be S or $\overline{S}$ conditional on S, writing out the inner-most expectation in this way, and by setting $\ell(Y,\tilde{V}_S)$ to 0.

Setting $t = \frac{\log 2}{b}$ and sending $u \to \infty$, we see that the inner expectation in (73) is upper-bounded by 1, which leads to

$$\mathbf{E}[\ell(Y,V')] \le \frac{b}{\log 2} I(S;Y|\tilde{V}). \tag{74}$$

The claim follows from the observation made in [49] that $I(S;Y|\tilde{V}) \le I(\tilde{V},S;Y) = I(V;Y)$, where the equality holds because Y is conditionally independent of $(\tilde{V},S)$ given $V = \tilde{V}_S$. □

Theorem 7 can be weakened by the data processing inequality of mutual information,

$$I(g(X,W);Y|X,Z^n) \le I(W;Y|X,Z^n), \tag{75}$$

which can be further bounded by $\frac{1}{n}I(W;Y^n|X^n)$ or $\frac{1}{n}I(W;Z^n)$ due to Theorem 2 or Corollary 1. When Y is discrete, $I(W;Y^n|X^n)$ can be further bounded by $H(Y^n|X^n)$. It implies that under a realizable model with discrete Y, not necessarily binary, the MER with respect to a bounded loss function can be upper-bounded nonasymptotically on the order of $\frac{1}{n}H(Y^n|X^n)$.

With a realizable model, a natural question to ask is how the MER depends on the richness of the *generative function class* $\mathcal{G} = \{g(\cdot,w) : \mathsf{X} \to \mathsf{Y}, w \in \mathsf{W}\}$. When $\mathsf{Y} = \{0,1\}$, one featuring combinatorial quantity that measures the richness of $\mathcal{G}$ is its VC dimension, defined in (42). The connection between $H(Y^n|X^n)$ and $V(\mathcal{G})$ is observed in [19], in the setting of binary classification with the zero-one loss. We make use of it here to obtain a corollary of Theorem 7, which extends the results in Theorem 5 as it applies to more general loss functions.

Corollary 5: Under a realizable model with $\mathsf{Y} = \{0,1\}$, if the function class $\mathcal{G} = \{g(\cdot,w) : \mathsf{X} \to \mathsf{Y}, w \in \mathsf{W}\}$ has VC dimension d, then for any loss function $\ell \in [0,b]$,

$$\mathrm{MER}_\ell \le O\Big(\frac{b}{\log 2} \cdot \frac{d}{n}\Big) \quad \text{as } n \to \infty. \tag{76}$$

Proof: By (75) and Theorem 2, and the assumptions that the model is realizable and Y is discrete, the upper bound in Theorem 7 can be weakened to

$$\begin{aligned}
\mathrm{MER}_\ell &\le \frac{b}{\log 2} H(Y|X,Z^n) && (77)\\
&\le \frac{b}{n\log 2} H(Y^n|X^n). && (78)
\end{aligned}$$

The Sauer-Shelah lemma as stated in (43) implies that $H(Y^n|X^n) \le d\log n + 1$. In addition, from the chain rule of Shannon entropy and Lemma A3, we have

$$\lim_{n\to\infty}(n+1)H(Y|X,Z^n) = \lim_{n\to\infty}\frac{H(Y^n|X^n)}{\log n} \le d,$$

similar to the proof of (45) in Theorem 5. This proves the claim in view of (78). □

The VC dimension plays a key role in the frequentist learning theory, in bounding the excess risk in terms of the richness of the *hypothesis class*, which amounts to the set of decision rules. In Bayesian learning, while there is no restriction on the decision rules, Corollary 5 shows that the VC dimension of the generative function class plays a similar role in upper-bounding the MER.

IV. Upper Bounds via Functional and Distributional Continuities

In the previous section, the upper bounds are derived by relating the MER to $I(W;Y|X,Z^n)$. In this section, we explore alternative methods for bounding the MER, either in terms of the smoothness of the optimal omniscient decision rule in model parameters, or in terms of the smoothness of the minimum expected loss in the predictive model. These smoothness, or continuity properties enable us to bound the MER via the accuracy of estimated parameters from the data. The following lemma is instrumental for this approach.

Lemma 4 ([50], [51]): Let (U,ρ) be a metric space. If U and U' are two random elements of U that are conditionally i.i.d. given another random element V of some space V, i.e., $P_{U,U'|V=v} = P_{U|V=v}P_{U'|V=v}$ and $P_{U|V=v} = P_{U'|V=v}$ for all $v \in \mathsf{V}$, then $\mathbf{E}[\rho(U',U)] \le 2R_\rho(U|V)$. Moreover, if $\mathsf{U} = \mathbb{R}^d$, then $\mathbf{E}[\|U'-U\|^2] = 2R_2(U|V)$.

As an aside, Lemma 4 provides us with a means for evaluating the performance of making randomized prediction by sampling from the posterior predictive distribution $P_{Y|X,Z^n}$, via upper-bounding the corresponding MER. Let Y' be sampled from $P_{Y|X,Z^n}$, which can be realized by first

sampling W' from $P_{W|X,Z^n}$ then Y' from $P_{Y|X,W'}$. Then, for any metric ℓ on Y, we have

$$\mathbf{E}[\ell(Y,Y')] \le 2R_\ell(Y|X,W) + 2\mathrm{MER}_\ell. \tag{79}$$

A. Via Continuity of Optimal Omniscient Decision Rule

1) General Upper Bound: We start from (52) which states that $\mathrm{MER}_\ell \le \mathbf{E}[\ell(Y,\Psi^*(X,W'))] - \mathbf{E}[\ell(Y,\Psi^*(X,W))]$, where Ψ^* is the optimal omniscient decision that achieves $R_\ell(Y|X,W)$ when W is known, and W' is a sample from the posterior distribution $P_{W|X,Z^n}$ conditionally independent of everything else given (X,Z^n). The MER can be upper-bounded in terms of the smoothness of the function $\ell(y,\Psi^*(x,w))$ in w and the accuracy of approximating W by W'.

Theorem 8: If $\mathsf{W}=\mathbb{R}^d$ and W is independent of X, then

$$\mathrm{MER}_\ell \le \mathbf{E}\Big[\sup_{y\in\mathsf{Y}}\sup_{w\in\mathsf{W}} \|\nabla_w \ell(y,\Psi^*(X,w))\|\Big]\cdot \sqrt{2R_2(W|Z^n)}, \tag{80}$$

where Ψ^* is the optimal omniscient decision rule for the loss function ℓ.

Proof: We have

$$\begin{aligned}
&\mathrm{MER}_\ell\\
&\le \mathbf{E}[\ell(Y,\Psi^*(X,W')) - \ell(Y,\Psi^*(X,W))]\\
&\le \mathbf{E}\Big[\sup_{w\in\mathsf{W}} \|\nabla_w \ell(Y,\Psi^*(X,w))\|\cdot\|W'-W\|\Big]\\
&\le \mathbf{E}\Big[\sup_{y\in\mathsf{Y}}\sup_{w\in\mathsf{W}} \|\nabla_w \ell(y,\Psi^*(X,w))\|\Big]\mathbf{E}[\|W'-W\|]\\
&\le \mathbf{E}\Big[\sup_{y\in\mathsf{Y}}\sup_{w\in\mathsf{W}} \|\nabla_w \ell(y,\Psi^*(X,w))\|\Big]\sqrt{2R_2(W|Z^n)},
\end{aligned}$$

where we used (52), Lemma A5, the assumption that X and W are independent, and the fact that $\mathbf{E}[\|W'-W\|] \le \sqrt{\mathbf{E}[\|W'-W\|^2]} = \sqrt{2R_2(W|Z^n)}$ due to Lemma 4. □

a) Example: constant Ψ^*: An extreme case where Theorem 8 can be useful is when $\Psi^* : \mathsf{X}\times\mathsf{W}\to\mathsf{Y}$ does not dependent on W under certain loss functions, in which case Theorem 8 guarantees that the MER is zero. For example, if $Y_i = g(X_i,W)V_i$ and $Y = g(X,W)V$, with some $g : \mathsf{X}\times\mathsf{W}\to\mathbb{R}$ and (V^n,V) being i.i.d. zero-mean random variables independent of (W,X^n,X), then for the quadratic loss,

$$\begin{aligned}
\Psi^*(X,W) &= \mathbf{E}[g(X,W)V|X,W] &(81)\\
&= g(X,W)\mathbf{E}[V] &(82)\\
&\equiv 0, &(83)
\end{aligned}$$

hence $\mathrm{MER}_2 = 0$ by Theorem 8. It implies that for this example

$$\begin{aligned}
R_2(Y|X,Z^n) &= R_2(Y|X,W) &(84)\\
&= \mathbf{E}\big[g(X,W)^2V^2\big] &(85)\\
&= \mathbf{E}\big[g(X,W)^2\big]\mathsf{var}[V], &(86)
\end{aligned}$$

which shows that a small MER does not necessarily mean a small Bayes risk $R_\ell(Y|X,Z^n)$.

b) Example: logistic regression: Another example where Theorem 8 can be applied to is bounding the MER for logistic regression with the log loss. Bayesian logistic regression is an instance under the predictive modeling framework, where $\mathsf{Y}=\{0,1\}$, $W\in\mathbb{R}^d$ is assumed to be independent of X, and the predictive model is specified by $K_{Y|x,w}(1) = \sigma(w^\top\phi(x))$, with $\sigma(a) := 1/(1+e^{-a})$, $a\in\mathbb{R}$, being the logistic sigmoid function, and $\phi(x)\in\mathbb{R}^d$ being the feature vector of the observation. For the log loss, the optimal omniscient decision rule $\Psi^*(x,w)$ is the Bernoulli distribution with bias $\sigma(w^\top\phi(x))$, the same as $K_{Y|x,w}$. Since $|\frac{\mathrm{d}}{\mathrm{d}a}\log\sigma(a)| = |1-\sigma(a)| \le 1$ and $|\frac{\mathrm{d}}{\mathrm{d}a}\log(1-\sigma(a))| = |-\sigma(a)| \le 1$, from Theorem 8 we have

$$\mathrm{MER}_{\log} \le \mathbf{E}[\|\phi(X)\|]\sqrt{2R_2(W|Z^n)}. \tag{87}$$

This result explicitly shows that the MER in logistic regression depends on how well we can estimate the model parameters from data, as it is dominated by $R_2(W|Z^n)$, the MMSE of estimating W from Z^n. For logistic regression, a closed-form expression for this MMSE may not exist. Nevertheless, any upper bound on it that is nonasymptotic in d and n will translate to a nonasymptotic upper bound on the MER. In Section IV-C.3 we continue the discussion of this example with different upper-bounding methods, where the dependence on $R_2(W|Z^n)$ can be improved when it is small.

2) Realizable Models With Additive Noise: The smoothness of $\Psi^*(x,w)$ in w can lead to potentially tighter MER bounds under realizable models, possibly with additive noise. Consider the generative model of the form $Y_i = g(X_i,W)+V_i$ and $Y = g(X,W)+V$, where the generative function $g:\mathsf{X}\times\mathsf{W}\to\mathbb{R}$ is some parametric nonlinearity in general, which could be approximated by a neural network, the parameter vector $W\in\mathbb{R}^d$ is independent of (X^n,X), and the additive noise (V^n,V) are i.i.d. real-valued random variables independent of (W,X^n,X). This model encompasses both linear and nonlinear Bayesian regression problems. We have the following MER bounds for the quadratic loss under this model.

Theorem 9: Under the model considered above, for the quadratic loss,

$$\begin{aligned}
\mathrm{MER}_2 &\le 2R_2(g(X,W)|X,Z^n) &(88)\\
&\le 2\mathbf{E}\Big[\sup_{w\in\mathsf{W}}\|\nabla_w g(X,w)\|^2\Big]R_2(W|Z^n). &(89)
\end{aligned}$$

Proof: Under the model considered above, for the quadratic loss,

$$\begin{aligned}
\Psi^*(X,W) &= \mathbf{E}[g(X,W)+V|X,W] &(90)\\
&= g(X,W)+\mathbf{E}[V]. &(91)
\end{aligned}$$

We have

$$\begin{aligned}
&\mathrm{MER}_2\\
&\le \mathbf{E}\big[(Y-g(X,W')-\mathbf{E}[V])^2\big] -\\
&\quad \mathbf{E}\big[(Y-g(X,W)-\mathbf{E}[V])^2\big] &(92)\\
&= \mathbf{E}\big[(g(X,W')-g(X,W))^2\big] + \mathsf{var}[V]\\
&\quad - \mathsf{var}[V] &(93)\\
&= 2R_2(g(X,W)|X,Z^n) &(94)\\
&\le \mathbf{E}\Big[\sup_{w\in\mathsf{W}}\|\nabla_w g(X,w)\|^2\cdot\|W'-W\|^2\Big] &(95)
\end{aligned}$$

$$= 2\mathbf{E}\Big[\sup_{w\in\mathsf{W}} \|\nabla_w g(X,w)\|^2\Big] R_2(W|Z^n), \tag{96}$$

where (92) follows from (52); (93) follows from the independence between (X,W) and V; (94) follows from the fact that $g(X,W')$ and $g(X,W)$ are conditionally independent given (X,Z^n), and Lemma 4; (95) follows from (93) and Lemma A5; and (96) follows from the independence between X and W, and Lemma 4. □

a) Example: linear regression: Theorem 9 can be applied to bounding the MER of the linear regression problem with the quadratic loss. Bayesian linear regression is an instance of the noisy realizable model considered above, where $g(x,w) = w^\top\phi(x)$, $\phi(x) \in \mathbb{R}^d$ is the feature vector of the observation x, and (V^n, V) are i.i.d. samples from $\mathcal{N}(0,\sigma^2)$. With the Gaussian prior of model parameters $P_W = \mathcal{N}(0,\sigma_W^2\mathbf{I}_d)$, the MMSE for estimating W from Z^n has a closed-form expression

$$R_2(W|Z^n) = \mathbf{E}[\mathrm{tr}(C_{W|Z^n})], \tag{97}$$

where

$$C_{W|Z^n} = \Big(\frac{1}{\sigma_W^2}\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{\Phi}\mathbf{\Phi}^\top\Big)^{-1} \tag{98}$$

is the conditional covariance matrix of W given Z^n, which only depends on X^n through the $d\times n$ feature matrix $\mathbf{\Phi} = [\phi(X_1),\ldots,\phi(X_n)]$. Under this model, we also have $\nabla_w g(x,w) = \phi(x)$, hence Theorem 9 implies that

$$\mathrm{MER}_2 \le 2\mathbf{E}[\|\phi(X)\|^2]\mathbf{E}[\mathrm{tr}(C_{W|Z^n})]. \tag{99}$$

Under the above model with Gaussian prior, it can be shown that the posterior predictive distribution $P_{Y|x,z^n}$ is Gaussian with variance $\sigma^2+\phi(x)^\top C_{W|z^n}\phi(x)$. From this we can obtain exact expressions for the MER and alternative upper bounds:

$$\begin{aligned}\mathrm{MER}_{\log} &= \frac{1}{2}\mathbf{E}\Big[\log\Big(1+\frac{1}{\sigma^2}\phi(X)^\top C_{W|Z^n}\phi(X)\Big)\Big]\\ &\le \frac{\mathbf{E}[\|\phi(X)\|^2]}{2\sigma^2}\mathbf{E}[\mathrm{tr}(C_{W|Z^n})],\end{aligned} \tag{100}$$

and

$$\begin{aligned}\mathrm{MER}_2 &= \mathbf{E}\big[\phi(X)^\top C_{W|Z^n}\phi(X)\big]\\ &\le \mathbf{E}[\|\phi(X)\|^2]\mathbf{E}[\mathrm{tr}(C_{W|Z^n})].\end{aligned} \tag{101}$$

The upper bounds in (100) and (101) are justified by noting that

$$\phi(X)^\top C_{W|Z^n}\phi(X) = \|C_{W|Z^n}^{1/2}\phi(X)\|^2 \tag{102}$$

$$\le \sigma_1^2\big(C_{W|Z^n}^{1/2}\big)\|\phi(X)\|^2 \tag{103}$$

$$\le \mathrm{tr}(C_{W|Z^n})\|\phi(X)\|^2, \tag{104}$$

where $\sigma_1(\cdot)$ is the largest singular value of the underlying matrix. A special choice of the d feature functions composing the feature vector is such that they are orthonormal with respect to P_X, namely $\int_{\mathsf{X}}\phi_i(x)\phi_j(x)P_X(\mathrm{d}x) = \mathbf{1}\{i=j\}$ for $i,j\in\{1,\ldots,d\}$. In this case, $\mathbf{\Phi}\mathbf{\Phi}^\top \approx n\mathbf{E}[\phi(X)\phi(X)^\top] = n\mathbf{I}_d$, hence $\mathbf{E}[\mathrm{tr}(C_{W|Z^n})] \sim O(d/n)$, implying that MER_2 scale with d and n as $O(d/n)$ according to (101). It further implies that the upper bound (99) given by Theorem 9 is order-optimal for vanishing MER. We continue the discussion of this example in Section IV-C.3.

B. Deviation of Posterior Predictive Distribution From True Predictive Model

As described in Section II-A, under the predictive modeling framework, the generative model is specified as $P_{Z|W} = P_{X|W}K_{Y|X,W}$, with a parametrized probability transition kernel $K_{Y|X,W}$ describing the true predictive model of Y given X. An alternative method for upper-bounding the MER under this framework is by examining the deviation of the posterior predictive distribution $P_{Y|X,Z^n}$ from the true predictive model $K_{Y|X,W}$ in terms of a suitable *convex statistical distance* between them. Here, by a convex statistical distance we mean any statistical distance $(P,Q)\mapsto D(P,Q)$ that is convex in the first argument when the second one is held fixed, or convex in the second argument while the first one is held fixed. For example, any f-divergence, including the commonly used total variation distance, KL divergence and χ^2-divergence, is jointly convex in both arguments [52]. As another example, consider the pth power of p-Wasserstein distance between two Borel probability measures P and Q on $\mathbb{R}^m$ with finite second moments [53]:

$$\mathcal{W}_p^p(P,Q) := \inf_{\pi\in\Pi(P,Q)}\mathbf{E}_{(X,Y)\sim\pi}[\|X-Y\|^p], \tag{105}$$

where $\Pi(P,Q)$ denotes the collection of all couplings of P and Q, i.e., Borel probability measures on $\mathbb{R}^m\times\mathbb{R}^m$ with marginals P and Q. As shown in Lemma A4 in Appendix A.4, $(P,Q)\mapsto\mathcal{W}_p^p(P,Q)$ is also jointly convex. The following lemma is key for relating the deviation of $P_{Y|X,Z^n}$ from $K_{Y|X,W}$ to the estimation error of model parameters.

Lemma 5: Let W' be a sample from the posterior distribution $P_{W|X,Z^n}$, such that W and W' are conditionally i.i.d. given (X,Z^n). Then for any (w,x,z^n) and any statistical distance D that is convex in the first argument,

$$D(P_{Y|x,z^n},K_{Y|x,w}) \le \mathbf{E}[D(K_{Y|x,W'},K_{Y|x,w})|x,z^n], \tag{106}$$

and consequently,

$$\mathbf{E}[D(P_{Y|X,Z^n},K_{Y|X,W})] \le \mathbf{E}[D(K_{Y|X,W'},K_{Y|X,W})] \tag{107}$$

where the expectations are taken with respect to the joint distribution of (W,X,Z^n,W'). Similarly, for any (w,x,z^n) and any statistical distance D that is convex in the second argument,

$$D(K_{Y|x,w},P_{Y|x,z^n}) \le \mathbf{E}[D(K_{Y|x,w},K_{Y|x,W'})|x,z^n], \tag{108}$$

and consequently,

$$\mathbf{E}[D(K_{Y|X,W},P_{Y|X,Z^n})] \le \mathbf{E}[D(K_{Y|X,W},K_{Y|X,W'})]. \tag{109}$$

Proof: From the joint distribution in (1), it follows that for any (w,z^n,z),

$$P_{Y|x,z^n}(y) = \int_{\mathsf{W}} K_{Y|x,w'}(y)P_{W|x,z^n}(\mathrm{d}w') \tag{110}$$

If the statistical distance considered here is convex in the first argument, we have

$$D(P_{Y|x,z^n},K_{Y|x,w}) \le \int_{\mathsf{W}} D(K_{Y|x,w'},K_{Y|x,w})P_{W|x,z^n}(\mathrm{d}w'), \tag{111}$$

which proves (106). Taking expectations over the conditioning terms, we obtain (107). The proof of (108) and (109) follows the same argument when D is convex in the second argument. □

Whenever the convex statistical distance $D(K_{Y|x,w'}, K_{Y|x,w})$ can be upper-bounded via $\|w' - w\|$ or $\|w' - w\|^2$, Lemma 4 can be used to further upper-bound $\mathbf{E}[D(K_{Y|X,W'}, K_{Y|X,W})]$ in terms of the minimum achievable estimation error of W. In the following two subsections, we use two different methods together with Lemma 5 and Lemma 4 to convert upper bounds on the deviation of $P_{Y|X,Z^n}$ from $K_{Y|X,W}$ into upper bounds on the MER for various loss functions.

C. From Deviation of Posterior Predictive Distribution to MER Bound

1) Via Conditional Mutual Information Upper Bound: For the log loss, we can directly upper-bound $I(W;Y|X,Z^n)$ in terms of the KL divergence between $K_{Y|X,W}$ and $P_{Y|X,Z^n}$, and arrive at the following result with Lemma 5.

Theorem 10: When $P_{Z|W} = P_{X|W}K_{Y|X,W}$, let W' be a sample from the posterior distribution $P_{W|X,Z^n}$, conditionally independent of everything else given (X, Z^n). Then,

$$\mathrm{MER}_{\log} \le \mathbf{E}[D_{\mathrm{KL}}(K_{Y|X,W}\|K_{Y|X,W'})] \tag{112}$$

where the expectation is taken with respect to the joint distribution of (W, W', X).

Proof: From (17), we have

$$\begin{aligned}\mathrm{MER}_{\log} &= I(W;Y|X,Z^n) && (113)\\ &= \mathbf{E}[D_{\mathrm{KL}}(P_{Y|X,Z^n,W}\|P_{Y|X,Z^n})] && (114)\\ &= \mathbf{E}[D_{\mathrm{KL}}(P_{Y|X,W}\|P_{Y|X,Z^n})] && (115)\\ &\le \mathbf{E}[D_{\mathrm{KL}}(K_{Y|X,W}\|K_{Y|X,W'})], && (116)\end{aligned}$$

where (115) follows from the fact that Y is conditionally independent of Z^n given (X, W); and (116) is from Lemma 5 and the fact that $D_{\mathrm{KL}}(P\|Q)$ is convex in Q for a fixed P. □

In Section IV-C.3 we continue with the example of logistic regression, where Theorem 10 can be used with Lemma 4 to bound the MER in terms of the MMSE of estimating W from data.

2) Via Continuity of Generalized Entropy: The second method for relating the MER to the deviation of posterior predictive distribution is directly comparing $R_\ell(Y|X,Z^n)$ against $R_\ell(Y|X,W)$, via the distributional continuity of the generalized entropy. We examine classification and regression problems separately.

a) Classification: For classification problems where Y is finite, we consider both the soft classification with the log loss and the hard classification with the zero-one loss. The MER upper bounds rely on the continuity properties of the Shannon entropy and the maximal probability, respectively, as stated in the following lemma, with proofs provided in Appendix A.6. For more general discussions on the continuity of generalized entropy, the reader may refer to [54], [55].

Lemma 6: Let P and Q be distributions on a finite set Y such that $\min_{y\in\mathsf{Y}} Q(y) > 0$. Then

$$H(P) - H(Q) \le \big(-\log \min_{y\in\mathsf{Y}} Q(y)\big) d_{\mathrm{TV}}(P,Q), \tag{117}$$

$$\max_{y\in\mathsf{Y}} Q(y) - \max_{y\in\mathsf{Y}} P(y) \le d_{\mathrm{TV}}(P,Q), \tag{118}$$

where $d_{\mathrm{TV}}(P,Q) := \frac{1}{2}\sum_{y\in\mathsf{Y}} |P(y) - Q(y)|$ is the total variation distance between P and Q.

Compared with the well-known Shannon entropy difference bound in terms of the total variation distance $|H(P) - H(Q)| \le 2\, d_{\mathrm{TV}}(P,Q)\log(|\mathsf{Y}|/2d_{\mathrm{TV}}(P,Q))$ when $d_{\mathrm{TV}}(P,Q) \le 1/4$ [33, Theorem 17.3.3], the bound given in (117) is not as tight in $|\mathsf{Y}|$, but is tighter in $d_{\mathrm{TV}}(P,Q)$, which leads to sharper MER bounds when the data size is large. Armed with Lemma 6 and Lemma 5, we have the following MER bounds for classification problems.

Theorem 11: If Y is finite, then for the log loss,

$$\begin{aligned}\mathrm{MER}_{\log} \le \sup_{x\in\mathsf{X},\, w\in\mathsf{W}} &(-\log \kappa(x,w))\cdot\\ &\mathbf{E}[d_{\mathrm{TV}}(K_{Y|X,W'}, K_{Y|X,W})],\end{aligned} \tag{119}$$

where $\kappa(x,w) := \min_{y\in\mathsf{Y}} K_{Y|x,w}(y)$, W' is a sample from $P_{W|X,Z^n}$, conditionally independent of everything else given (X, Z^n), and the expectation is with respect to $P_{W,W',X}$. In addition, for the zero-one loss,

$$\mathrm{MER}_{01} \le \mathbf{E}[d_{\mathrm{TV}}(K_{Y|X,W'}, K_{Y|X,W})]. \tag{120}$$

Proof: When Y is finite, for the log loss,

$$\begin{aligned}&\mathrm{MER}_{\log}\\ &= H(Y|X,Z^n) - H(Y|X,W) && (121)\\ &= \int \big(H(Y|x,z^n) - H(Y|x,w)\big) P(\mathrm{d}w,\mathrm{d}x,\mathrm{d}z^n)\\ &\le \int \Big(-\log \min_{y\in\mathsf{Y}} K_{Y|x,w}(y)\Big) d_{\mathrm{TV}}(P_{Y|x,z^n}, P_{Y|x,w})\\ &\qquad P(\mathrm{d}w,\mathrm{d}x,\mathrm{d}z^n) && (122)\\ &\le \sup_{w\in\mathsf{W},x\in\mathsf{X}} \big(-\log \min\nolimits_{y\in\mathsf{Y}} K_{Y|x,w}(y)\big)\\ &\qquad \mathbf{E}\big[d_{\mathrm{TV}}(P_{Y|X,Z^n}, P_{Y|X,W})\big] && (123)\\ &\le \sup_{w\in\mathsf{W},x\in\mathsf{X}} \big(-\log \min\nolimits_{y\in\mathsf{Y}} K_{Y|x,w}(y)\big)\\ &\qquad \mathbf{E}\big[d_{\mathrm{TV}}(P_{Y|X,W'}, P_{Y|X,W})\big] && (124)\end{aligned}$$

where (122) follows from Lemma 6; and (124) follows from Lemma 5.

For the zero-one loss,

$$\begin{aligned}&\mathrm{MER}_{01}\\ &= \mathbf{E}[\max\nolimits_{y\in\mathsf{Y}} K_{Y|X,W}(y)] - \mathbf{E}[\max\nolimits_{y\in\mathsf{Y}} P_{Y|X,Z^n}(y)]\\ &= \int \big(\max\nolimits_{y\in\mathsf{Y}} K_{Y|x,w}(y) - \max\nolimits_{y\in\mathsf{Y}} P_{Y|x,z^n}(y)\big)\\ &\qquad P(\mathrm{d}w,\mathrm{d}x,\mathrm{d}z^n) && (125)\\ &\le \int d_{\mathrm{TV}}(K_{Y|x,w}, P_{Y|x,z^n}) P(\mathrm{d}w,\mathrm{d}x,\mathrm{d}z^n) && (126)\\ &\le \mathbf{E}[d_{\mathrm{TV}}(K_{Y|X,W}, K_{Y|X,W'})] && (127)\end{aligned}$$

where (126) follows from Lemma 6, and (127) follows from Lemma 5. □

b) Regression: Next, we consider regression problems with $\mathsf{Y} \subset \mathbb{R}^p$ under the assumption that both the marginal and various conditional distributions of Y are absolutely continuous with respect to the Lebesgue measure. We consider both the soft prediction with the log loss, and the hard prediction with the quadratic loss. For the soft setting, $\mathrm{MER}_{\log}$ can be upper-bounded using the continuity of differential entropy with respect to the Wasserstein distance, as stated in the following lemma.

Lemma 7 ([56]): Let U be a random vector in $\mathbb{R}^p$ with finite $\mathbf{E}[\|U\|^2]$, and V be a Gaussian random vector in $\mathbb{R}^p$ with covariance matrix $\sigma^2\mathbf{I}_p$. Then

$$h(U) - h(V) \le \frac{1}{2\sigma^2}\big(3\sqrt{\mathbf{E}[\|U\|^2]} + 11\sqrt{\mathbf{E}[\|V\|^2]}\big)\cdot \mathcal{W}_2(P_U, P_V) \tag{128}$$

where $\mathcal{W}_2(P_U, P_V)$ is the 2-Wasserstein distance between P_U and P_V.

With Lemma 7 and Lemma 5, we have the following bound for regression with the log loss.

Theorem 12: If $\mathsf{Y} = \mathbb{R}^p$, and $K_{Y|x,w}$ is Gaussian with covariance matrix $\sigma^2\mathbf{I}_p$ for all (x, w), then for the log loss,

$$\mathrm{MER}_{\log} \le \frac{7}{\sigma^2}\sqrt{\mathbf{E}\big[\|Y\|^2\big]\mathbf{E}[\mathcal{W}_2^2(K_{Y|X,W'}, K_{Y|X,W})]}, \tag{129}$$

where W and W' are conditionally i.i.d. given (X, Z^n), and the expectation is with respect to $P_{X,W,W'}$.

Proof: For the log loss,

$$\begin{aligned}
&\mathrm{MER}_{\log} \\
&= h(Y|X, Z^n) - h(Y|X, W) && (130)\\
&= \int \big(h(Y|x, z^n) - h(Y|x, w)\big)P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n) \\
&\le \int \Big(\frac{3}{2\sigma^2}\sqrt{\mathbf{E}\big[\|Y\|^2|x, z^n\big]} + \frac{11}{2\sigma^2}\sqrt{\mathbf{E}\big[\|Y\|^2|x, w\big]}\Big) \\
&\qquad \mathcal{W}_2(P_{Y|x,z^n}, K_{Y|x,w})P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n) && (131)\\
&\le \bigg(\int \Big(\frac{3}{2\sigma^2}\sqrt{\mathbf{E}\big[\|Y\|^2|x, z^n\big]} + \\
&\qquad \frac{11}{2\sigma^2}\sqrt{\mathbf{E}\big[\|Y\|^2|x, w\big]}\Big)^2 P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n)\bigg)^{1/2} \\
&\qquad \bigg(\int \mathcal{W}_2^2(P_{Y|x,z^n}, K_{Y|x,w})P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n)\bigg)^{1/2} && (132)\\
&\le \frac{7}{\sigma^2}\big(\mathbf{E}\big[\|Y\|^2\big]\big)^{1/2} \\
&\qquad \bigg(\int \mathcal{W}_2^2(P_{Y|x,z^n}, K_{Y|x,w})P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n)\bigg)^{1/2} && (133)\\
&\le \frac{7}{\sigma^2}\sqrt{\mathbf{E}\big[\|Y\|^2\big]\mathbf{E}[\mathcal{W}_2^2(K_{Y|X,W'}, K_{Y|X,W})]}\big). && (134)
\end{aligned}$$

where (131) follows from Lemma 7; (132) follows from Cauchy-Schwarz inequality; (133) follows from the triangle inequality of the L_2 norm, which states that $\sqrt{\mathbf{E}[(U+V)^2]} \le \sqrt{\mathbf{E}[U^2]} + \sqrt{\mathbf{E}[V^2]}$; and (134) follows from Lemma 5 and Lemma A4. □

For the hard setting, in scalar regression problems with $\mathsf{Y} = \mathsf{A} = \mathbb{R}$ and the quadratic loss, the MER as given by (34) is the expected difference between two variances. The following results relate the variance difference between two probability distributions to their 2-Wasserstein distance and KL divergence respectively.

Lemma 8 ([57], proof given in Appendix A.7): Let U and V be random variables over a set $\mathsf{U} \subset \mathbb{R}$ with finite $\mathbf{E}[U^2]$ and $\mathbf{E}[V^2]$. Then,

$$|\mathsf{var}[U] - \mathsf{var}[V]| \le 2\big(\sqrt{\mathbf{E}[U^2]} + \sqrt{\mathbf{E}[V^2]}\big)\mathcal{W}_2(P_U, P_V). \tag{135}$$

When V is Gaussian with variance σ^2, Lemma 8 with Talagrand's inequality [58] states that

$$|\mathsf{var}[U] - \mathsf{var}[V]| \le 2\big(\sqrt{\mathbf{E}[U^2]} + \sqrt{\mathbf{E}[V^2]}\big)\sqrt{2\sigma^2 D_{\mathrm{KL}}(P_U\|P_V)}; \tag{136}$$

under the same condition, we also have a potentially tighter bound [55]:

$$|\mathsf{var}[U] - \mathsf{var}[V]| \le 2\sigma^2\big(\sqrt{D_{\mathrm{KL}}(P_U\|P_V)} + D_{\mathrm{KL}}(P_U\|P_V)\big). \tag{137}$$

With Lemma 8, we can derive the following upper bounds for MER_2.

Theorem 13: For regression problems with $\mathsf{Y} = \mathbb{R}$, if $\mathbf{E}[Y^2|x, w]$ is finite for all (x, w), then for the quadratic loss,

$$\mathrm{MER}_2 \le 4\sqrt{\mathbf{E}[Y^2]\mathbf{E}[\mathcal{W}_2^2(K_{Y|X,W'}, K_{Y|X,W})]}, \tag{138}$$

where W' is a sample from $P_{W|X,Z^n}$, conditionally independent of everything given (X, Z^n).

Proof: Similar to the proof of Theorem 12, for the quadratic loss, we have

$$\begin{aligned}
&\mathrm{MER}_2 \\
&= \int \big(\mathsf{var}[Y|x, z^n] - \mathsf{var}[Y|x, w]\big)P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n) \\
&\le 2\int \big(\sqrt{\mathbf{E}[Y^2|x, z^n]} + \sqrt{\mathbf{E}[Y^2|x, w]}\big) \\
&\qquad \mathcal{W}_2(P_{Y|x,z^n}, K_{Y|x,w})P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n) && (139)\\
&\le 2\bigg(\int \big(\sqrt{\mathbf{E}[Y^2|x, z^n]} + \sqrt{\mathbf{E}[Y^2|x, w]}\big)^2 \\
&\qquad P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n)\bigg)^{1/2} \\
&\qquad \bigg(\int \mathcal{W}_2^2(P_{Y|x,z^n}, K_{Y|x,w})P(\mathrm{d}w, \mathrm{d}x, \mathrm{d}z^n)\bigg)^{1/2} && (140)\\
&\le 4\sqrt{\mathbf{E}[Y^2]\mathbf{E}[\mathcal{W}_2^2(P_{Y|X,Z^n}, K_{Y|X,W})]} && (141)\\
&\le 4\sqrt{\mathbf{E}[Y^2]\mathbf{E}[\mathcal{W}_2^2(P_{Y|X,W'}, K_{Y|X,W})]}. && (142)
\end{aligned}$$

where (139) follows from Lemma 8; (140) follows from Cauchy-Schwarz inequality; (141) follows from the triangle inequality of the L_2 norm; and (142) follows from Lemma 5 and Lemma A4.

□

In Section IV-C.3 we make use of Theorem 11, 12 and 13 with Lemma 4 to bound the MER in concrete learning problems in terms of the MMSE of estimating W from data.

3) Examples:

a) Logistic regression (continued): We continue with the example of logistic regression discussed in Section IV-A.1, where $\mathsf{Y}=\{0,1\}$, $\mathsf{W}\subset\mathbb{R}^d$, $K_{Y|x,w}(1)=\sigma(w^\top\phi(x))$ with $\sigma(a):=1/(1+e^{-a})$, and X is assumed to be independent of W. As $\|\nabla_w\sigma(w^\top\phi(x))\|\le\|\phi(x)\|/4$, from Lemma A5 we know that $\sigma(w^\top\phi(x))$ is $\|\phi(x)\|/4$-Lipschitz in w, hence

$$d_{\mathrm{TV}}(K_{Y|x,w'},K_{Y|x,w})=\big|\sigma(w'^\top\phi(x))-\sigma(w^\top\phi(x))\big|\le\frac{1}{4}\|\phi(x)\|\|w'-w\|. \tag{143}$$

Then, from Theorem 10, Theorem 11 and Lemma 4, the following bounds hold for the log loss and the zero-one loss.

Corollary 6: In binary logistic regression, for the log loss,

$$\mathrm{MER}_{\log}\le\log\Big(1+\frac{1}{2}s_\phi^2e^{s_\phi s_{\mathsf{W}}}R_2(W|Z^n)\Big) \tag{144}$$

where $s_\phi:=\sup_{x\in\mathsf{X}}\|\phi(x)\|$ and $s_{\mathsf{W}}:=\sup_{w\in\mathsf{W}}\|w\|$; while for the zero-one loss,

$$\mathrm{MER}_{01}\le\frac{1}{4}\mathbf{E}[\|\phi(X)\|]\sqrt{2R_2(W|Z^n)}. \tag{145}$$

Proof: For the log loss, from Theorem 10,

$$\begin{aligned}
&\mathrm{MER}_{\log}\\
&\le\mathbf{E}\big[D_{\mathrm{KL}}(K_{Y|X,W}\|K_{Y|X,W'})\big] &(146)\\
&\le\mathbf{E}\Big[\log\Big(1+\frac{2d_{\mathrm{TV}}^2(K_{Y|X,W},K_{Y|X,W'})}{\sigma(W'^\top\phi(X))\wedge(1-\sigma(W'^\top\phi(X)))}\Big)\Big] &(147)\\
&\le\mathbf{E}\Big[\log\Big(1+4e^{|W'^\top\phi(X)|}d_{\mathrm{TV}}^2(K_{Y|X,W},K_{Y|X,W'})\Big)\Big] &(148)\\
&\le\mathbf{E}\Big[\log\Big(1+\frac{1}{4}e^{s_\phi s_{\mathsf{W}}}s_\phi^2\|W-W'\|^2\Big)\Big] &(149)\\
&\le\log\Big(1+\frac{1}{2}e^{s_\phi s_{\mathsf{W}}}s_\phi^2R_2(W|Z^n)\Big)\Big) &(150)\\
&\le\frac{1}{2}s_\phi^2e^{s_\phi s_{\mathsf{W}}}R_2(W|Z^n) &(151)
\end{aligned}$$

where we used a reverse Pinsker's inequality as stated in [59, Theorem 28], the fact that $\sigma(w^\top\phi(x))\wedge(1-\sigma(w^\top\phi(x)))\ge\exp\{-|w^\top\phi(x)|\}/2$, $d_{\mathrm{TV}}(K_{Y|x,w'},K_{Y|x,w})\le\|\phi(x)\|\|w'-w\|/4$ from (143), Jensen's inequality, and Lemma 4.

For the zero-one loss, from Theorem 11,

$$\begin{aligned}
\mathrm{MER}_{01}&\le\mathbf{E}[d_{\mathrm{TV}}(K_{Y|X,W'},K_{Y|X,W})] &(152)\\
&\le\frac{1}{4}\mathbf{E}\big[\|\phi(X)\|\|W'-W\|\big] &(153)\\
&\le\frac{1}{4}\mathbf{E}[\|\phi(X)\|]\sqrt{\mathbf{E}\big[\|W'-W\|^2\big]} &(154)\\
&=\frac{1}{4}\mathbf{E}[\|\phi(X)\|]\sqrt{2R_2(W|Z^n)} &(155)
\end{aligned}$$

where we used (143) and Lemma 4. □

The upper bound in (144) shows that the rate of convergence of $\mathrm{MER}_{\log}$ in n for logistic regression is the same as that for $R_2(W|Z^n)$, as $\log(1+u)\le u$ for $u>0$. This improves the upper bound given in (87) when $R_2(W|Z^n)$ is small, e.g., when n is large.

b) Nonlinear and linear regression (continued): We also continue with the discussion on the nonlinear and linear regression problems in Section IV-A.2, where $Y=g(X,W)+V$, $\mathsf{W}=\mathbb{R}^d$, X and W are independent, and $V\sim\mathcal{N}(0,\sigma^2)$ is independent of (X,W). Under this model,

$$\begin{aligned}
\mathcal{W}_2^2(K_{Y|x,w'},K_{Y|x,w})&=2\sigma^2D_{\mathrm{KL}}(K_{Y|x,w'}\|K_{Y|x,w})\\
&=(g(x,w)-g(x,w'))^2. &(156)
\end{aligned}$$

From Theorem 10, Theorem 13 and Lemma 4, we obtain the following upper bounds for nonlinear regression.

Corollary 7: For the above nonlinear regression problem, let $s_g:=\mathbf{E}\big[\sup_{w\in\mathsf{W}}\|\nabla_wg(X,w)\|^2\big]$. Then for the log loss,

$$\mathrm{MER}_{\log}\le\frac{1}{\sigma^2}R_2(g(X,W)|X,Z^n)\le\frac{s_g}{\sigma^2}R_2(W|Z^n), \tag{157}$$

while for the quadratic loss,

$$\begin{aligned}
\mathrm{MER}_2&\le4\sqrt{2(\mathbf{E}[g(X,W)^2]+\sigma^2)R_2(g(X,W)|X,Z^n)}\\
&\le4\sqrt{2(\mathbf{E}[g(X,W)^2]+\sigma^2)s_gR_2(W|Z^n)}. &(158)
\end{aligned}$$

Proof: For the log loss,

$$\begin{aligned}
&\mathrm{MER}_{\log}\\
&\le\mathbf{E}[D_{\mathrm{KL}}(K_{Y|X,W}\|K_{Y|X,W'})] &(159)\\
&=\frac{1}{2\sigma^2}\mathbf{E}\big[\big(g(X,W)-g(X,W')\big)^2\big] &(160)\\
&=\frac{1}{\sigma^2}R_2(g(X,W)|X,Z^n) &(161)\\
&\le\frac{1}{2\sigma^2}\mathbf{E}\big[(\sup_{w\in\mathsf{W}}\|\nabla_wg(X,w)\|)^2\|W-W'\|^2\big] &(162)\\
&=\frac{1}{\sigma^2}\mathbf{E}\big[\sup_{w\in\mathsf{W}}\|\nabla_wg(X,w)\|^2\big]R_2(W|Z^n). &(163)
\end{aligned}$$

where (159) follows from Theorem 10; (160) is from (156); (161) is due to Lemma 4 for the quadratic loss; (162) is due to (160) and Lemma A5; and (163) follows from Lemma 4.

For the quadratic loss, from Theorem 13, (156) and Lemma 4, and a similar reasoning as above,

$$\begin{aligned}
&\mathrm{MER}_2\\
&\le4\sqrt{\mathbf{E}[Y^2]\mathbf{E}[\big(g(X,W)-g(X,W')\big)^2]} &(164)\\
&=4\sqrt{2\mathbf{E}[Y^2]R_2(g(X,W)|X,Z^n)} &(165)\\
&\le4\sqrt{2\mathbf{E}[Y^2]\mathbf{E}\big[\sup_{w\in\mathsf{W}}\|\nabla_wg(X,w)\|^2\big]R_2(W|Z^n)} &(166)\\
&=4\sqrt{2\big(\mathbf{E}[g(W,X)^2]+\sigma^2\big)s_gR_2(W|Z^n)}, &(167)
\end{aligned}$$

which proves the second upper bound. □

For the special case of linear regression with Gaussian prior $P_W=\mathcal{N}(0,\sigma_W^2\mathbf{I}_d)$, we have $g(x,w)=w^\top\phi(x)$, $s_g=\mathbf{E}[\|\phi(X)\|^2]$, and $R_2(W|Z^n)=\mathbf{E}[\mathrm{tr}(C_{W|Z^n})]$ with $C_{W|Z^n}$ given in (98); Corollary 7 in this case gives

$$\mathrm{MER}_{\log}\le\frac{1}{\sigma^2}\mathbf{E}[\|\phi(X)\|^2]\mathbf{E}[\mathrm{tr}(C_{W|Z^n})], \tag{168}$$

and

$$\mathrm{MER}_2 \le 4\Big(2\big(\sigma_W^2\mathbf{E}[\|\phi(X)\|^2]+\sigma^2\big)\mathbf{E}[\|\phi(X)\|^2]\mathbf{E}[\mathrm{tr}(C_{W|X^n,Y^n})]\Big)^{1/2}. \quad (169)$$

From the exact expressions of MER given in (100) and (101), we see that the upper bound for $\mathrm{MER}_{\log}$ in (168) is order-optimal for vanishing MER; while the upper bound for MER_2 in (169) is not, unlike the upper bound (99) for MER_2 given by Theorem 9. In Appendix B, we derive an alternative upper bound for MER_2 based on Theorem 6 in Section III-D, however it is not order-optimal either. Nevertheless, the upper bound in (169) can be tighter than the order-optimal upper bound in (99) when $R_2(W|Z^n)$ is large, e.g., when n is small.

We also see from Theorem 9 and Corollary 7 that the MER upper bounds for the general nonlinear regression problem depend on n only through $R_2(g(X,W)|X,Z^n)$ or $R_2(W|Z^n)$. Although closed-form expressions of these quantities are generally intractable, the upper bounds explicitly show how the epistemic part of the overall prediction uncertainty depends on the model uncertainty, which can be quantified by the corresponding MMSE. Moreover, the upper bounds obtained in terms of $R_2(g(X,W)|X,Z^n)$ can be much tighter than those in terms of $R_2(W|Z^n)$, especially when multiple values of W map to the same function $g(\cdot,W)$, e.g., when $g:\mathsf{X}\times\mathsf{W}\to\mathsf{Y}$ can be represented by over-parametrized neural networks [60].

V. Extensions

A. Multiple Model Families

Instead of being described by a single model family, in many cases the joint distribution of X and Y can be better represented by a finite class of model families $\mathbb{M}=\{\mathcal{M}_m, m\in\mathsf{M}\}$ all together, where each family $\mathcal{M}_m=\{P_{X,Y|w,m}, w\in\mathsf{W}_m\}$ is a collection of parametrized joint distributions of (X,Y). The class of model families $\mathbb{M}$ is also known as the *model class*, and the index m of each family is also known as the *model index* [61]. In the Bayesian formulation, the model index M is treated as a random element of M with prior P_M; given a model index m, the model parameters are represented as a random vector in W_m with prior $P_{W|m}$. As before, denoting $Z_i := (X_i,Y_i)$, $i=1,\dots,n$, as the observed data and $Z:=(X,Y)$ as a fresh pair, the quantities under consideration are assumed to be generated from the joint distribution

$$P_{M,W,Z^n,Z}=P_MP_{W|M}\Big(\prod_{i=1}^n P_{Z_i|W,M}\Big)P_{Z|W,M} \quad (170)$$

where $P_{Z_i|W,M}=P_{Z|W,M}$ for $i=1,\dots,n$. In the same spirit in the single model family setting, we can define the MER in the above multi-model family setting as follows.

Definition 4: In the multi-model family setting, the fundamental limit of the Bayes risk with respect to the loss function ℓ is defined as

$$R_\ell(Y|X,W,M)=\inf_{\Psi:\mathsf{X}\times\mathsf{W}\times\mathsf{M}\to\mathsf{A}}\mathbf{E}[\ell(Y,\Psi(X,W,M))]. \quad (171)$$

Definition 5: In the multi-model family setting, the minimum excess risk with respect to the loss function ℓ is defined as

$$\mathrm{MER}_\ell = R_\ell(Y|X,Z^n)-R_\ell(Y|X,W,M). \quad (172)$$

Similar to Lemma 2 and Theorem 10 in the single model family setting, for the log loss, the MER in the multi-model family setting can be related to the conditional mutual information $I(M,W;Y|X,Z^n)$ and its upper bounds.

Theorem 14: In the multi-model family setting, with the log loss,

$$\mathrm{MER}_{\log}=I(M,W;Y|X,Z^n), \quad (173)$$

which can be upper-bounded by $\frac{1}{n}I(M,W;Y^n|X^n)$. Further, if $P_{X,Y|w,m}=P_{X|w,m}K_{Y|X,w,m}$ for all $(m,w)\in\mathsf{M}\times\mathsf{W}_m$, then

$$\mathrm{MER}_{\log}\le\mathbf{E}[D_{\mathrm{KL}}(K_{Y|X,W,M}\|K_{Y|X,W',M'})], \quad (174)$$

where (M',W') is a sample from the posterior distribution $P_{M,W|X,Z^n}$ such that (M,W) and (M',W') are conditionally i.i.d. given (X,Z^n).

In addition, we can still bound the MER in terms of the deviation of the posterior predictive distribution from the true predictive model, similar to the results in Section IV-C. As in the predictive modeling framework, suppose that for each model family $\mathcal{M}_m\in\mathbb{M}$, $P_{X,Y|w,m}=P_{X|w,m}K_{Y|X,w,m}$ for all $w\in\mathsf{W}_m$. Then for any statistical distance D, a diameter-like quantity of the model class $\mathbb{M}$ with respect to D can be defined as

$$\mathrm{diam}(\mathbb{M},D)=\max_{m\ne m'\in\mathsf{M}}\sup_{w\in\mathsf{W}_m,w'\in\mathsf{W}_{m'}}\sup_{x\in\mathsf{X}}D(K_{Y|x,w',m'},K_{Y|x,w,m}). \quad (175)$$

With the above definition we have the following general upper bound on the deviation of the posterior predictive distribution from the true predictive model. The proof is given in Appendix C.

Theorem 15: In the multi-model setting, for any statistical distance D that is convex in the first argument,

$$\mathbf{E}[D(P_{Y|X,Z^n},K_{Y|X,W,M})]\le\mathbf{E}[D(K_{Y|X,W',M'},K_{Y|X,W,M})], \quad (176)$$

where (M',W') is a sample from the posterior distribution $P_{M,W|X,Z^n}$ such that (M,W) and (M',W') are conditionally i.i.d. given (X,Z^n). The right side of (176) can be further upper-bounded by

$$\mathbf{E}[D(K_{Y|X,W',M},K_{Y|X,W,M})]+2\mathrm{diam}(\mathbb{M},D)R_{01}(M|X,Z^n), \quad (177)$$

where W' is a sample from the posterior distribution $P_{W|X,Z^n,M}$ such that W' and W are conditionally i.i.d. given (X,Z^n,M). If D is convex in the second argument, we obtain another set of upper bounds by exchanging the order of the arguments of D in the results above.

Theorem 15 shows that under the multi-model family setting, the expected deviation consists of two parts: the first part can be related to the estimation error of the model parameters

when the model index is correctly identified, which depends on the complexity of each model family; the second part is related to the penalty when the model index is wrongly identified, which depends on the overall complexity of the model class and the error probability of model index estimation.

As an example, for linear regression with multiple model families, the predictive model in the mth family can be described as $K_{Y|x,w,m} = \mathcal{N}(w^\top \phi(x,m), \sigma^2)$, where $w \in \mathsf{W}_m \subset \mathbb{R}^{d_m}$ is the model parameter vector and $\phi(x,m) \in \mathbb{R}^{d_m}$ is the feature vector of the observation x. We also assume that X is independent of (M,W). In this case,

$$D_{\mathrm{KL}}(K_{Y|x,w',m'} \| K_{Y|x,w,m}) = \frac{1}{2\sigma^2}\big(w'^\top \phi(x,m') - w^\top \phi(x,m)\big)^2, \tag{178}$$

and

$$\mathrm{diam}(\mathbb{M}, D_{\mathrm{KL}}) = \frac{1}{2\sigma^2} \max_{m \neq m' \in \mathsf{M}} \sup_{w \in \mathsf{W}_m, w' \in \mathsf{W}_{m'}} \sup_{x \in \mathsf{X}} \big(w'^\top \phi(x,m') - w^\top \phi(x,m)\big)^2. \tag{179}$$

From Theorem 14, the chain rule of mutual information, Theorem 15, and the previous results on linear regression in the single model family, we have the following upper bounds for $\mathrm{MER}_{\log}$ for linear regression with multiple models:

$$\mathrm{MER}_{\log} \leq \frac{1}{2\sigma^2}\mathbf{E}[\|\phi_M(X)\|^2] R_2(W|Z^n, M) + H(M|Z^n) \tag{180}$$

and

$$\mathrm{MER}_{\log} \leq \frac{1}{\sigma^2}\mathbf{E}[\|\phi_M(X)\|^2] R_2(W|Z^n, M) + 2\mathrm{diam}(\mathbb{M}, D_{\mathrm{KL}}) R_{01}(M|Z^n) \tag{181}$$

where

$$R_2(W|Z^n, M) = \sum_{m \in \mathsf{M}} P_M(m) \mathbf{E}[\mathrm{tr}(C_{W|Z^n,m})] \tag{182}$$

with $C_{W|Z^n,m} = (\sigma_{W,m}^{-2}\mathbf{I}_{d_m} + \sigma^{-2}\mathbf{\Phi}_m\mathbf{\Phi}_m^\top)^{-1}$ and $\mathbf{\Phi}_m = [\phi(X_1,m), \ldots, \phi(X_n,m)]$ being the $d_m \times n$ feature matrix for the mth model family. We see that the MER consists of a part that depends on the minimum achievable model parameter estimation error given each model index, and a part that depends on the uncertainty of model index estimation and the "diameter" of the model class.

B. MER in Nonparametric Models

The definition of MER can also be extended to Bayesian learning under a nonparametric predictive model that can be specified in terms of a random process. Formally, consider the case where F is a real-valued random process indexed by $x \in \mathsf{X}$, and the predictive model is a probability transition kernel $K_{Y|F(X)}$. It is further assumed that F is *a priori* independent of X. The observed data and the fresh pair are assumed to be generated from the joint distribution

$$P_{F,Z^n,Z} = P_F\Big(\prod_{i=1}^n P_{Z_i|F}\Big) P_{Z|F} \tag{183}$$

where $P_{Z_i|F} = P_{Z|F} = P_X K_{Y|F(X)}$ for $i = 1, \ldots, n$. Two simple examples of the above model are 1) noiseless Gaussian process regression model, where F is a Gaussian process with a mean function $m : \mathsf{X} \to \mathbb{R}$ and a covariance function $k : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$, and $Y = F(X)$; and 2) binary classification model with Gaussian process as a latent function [8], where F can be the same Gaussian process, and $K_{Y|F(X)}(1|f(x)) = \sigma(f(x))$ with $\sigma(\cdot)$ being the logistic sigmoid function.

In the same spirit in the parametric case, the MER under the above nonparametric model can be defined as

$$\mathrm{MER}_\ell = R_\ell(Y|X, Z^n) - R_\ell(Y|F(X)), \tag{184}$$

where $R_\ell(Y|X,Z^n)$ and $R_\ell(Y|F(X))$ are defined according to the general definition of the Bayes risk in (3), and correspond to (2) and (11) respectively. For the log loss, using the fact that $H_\mu(Y|F(X), Z^n) = H_\mu(Y|F(X))$ and following the same argument as in Corollary 1, we have

$$\mathrm{MER}_{\log} = I(F(X); Y|X, Z^n) \leq \frac{1}{n} I(F(X); Y^n|X^n). \tag{185}$$

For the quadratic loss,

$$\mathrm{MER}_2 = R_2(Y|X, Z^n) - R_2(Y|F(X)). \tag{186}$$

In the special case of noiseless Gaussian process regression model, $R_2(Y|F(X)) = 0$, which implies

$$\mathrm{MER}_2 = R_2(F(X)|F(X_1), \ldots, F(X_n)) \tag{187}$$

$$= \mathbf{E}\big[k(X,X) - k(X,X^n)^\top \mathbf{\Sigma}(X^n)^{-1} k(X,X^n)\big] \tag{188}$$

where $k(X,X^n)$ is the covariance vector between $F(X)$ and $(F(X_1), \ldots, F(X_n))$, and $\mathbf{\Sigma}(X^n)$ is the covariance matrix of $(F(X_1), \ldots, F(X_n))$. The above expression may be further analyzed using the eigenfunction expansion of the covariance function k [8]. For the binary classification model with Gaussian process as the latent function, or more general models specified with non-Gaussian processes, the MER may not have a simple close-form expression.

VI. Summary and Discussion

We have defined the minimum excess risk in Bayesian learning with respect to general loss functions, and presented general methods for obtaining upper bounds for this quantity. How to lower-bound this quantity is left as an open problem. We would like to close the paper by discussing the following two aspects.

A. Tightness and Utility of the Results

Two methods for deriving upper bounds on the MER have been presented: one method relates the MER to $I(W;Y|X,Z^n)$; the other one relates it to $R_2(W|X,Z^n)$ via various continuity arguments.

With the precise asymptotic expansion of $I(W;Z^n)$, the first method is suitable for asymptotic analysis for a wide range of loss functions. Using this method, we have shown that for any bounded loss function, the MER scales with the data size n as $O(\sqrt{1/n})$ in general (Corollary 3 and the discussion thereafter), while for the log loss (Theorem 3), the quadratic

loss with bounded Y (Theorem 4), and bounded loss under realizable binary classification models (Corollary 5), this convergence rate can be improved to $O(1/n)$. When the model parameter lies in a compact subset of $\mathbb{R}^d$, or when the VC dimension of the generative function class is d, the MER bounds can also capture the dependence on d, as $O(\sqrt{d/n})$ or $O(d/n)$ in different settings. An MER lower bound of $\Omega(d/n)$ is derived in a follow-up work [47, Theorem 10] for the cases where the excess risk of using $\Psi^*(X, W')$ as the plug-in decision rule is lower bounded by $\|W - W'\|^2$, which matches upper bounds in certain settings derived in this work.

The second method has the potential to provide us with nonasymptotic upper bounds. The only explicit expression for $R_2(W|X, Z^n)$ we have so far is for linear regression, for which we have derived order-optimal upper bound for both $\mathrm{MER}_{\log}$ (via Corollary 7) and MER_2 (via Theorem 9). In order to obtain explicit upper bounds for problems beyond linear regression, e.g. logistic regression or nonlinear regression, we would need upper bounds on $R_2(W|X, Z^n)$, or other forms of minimum model parameter estimation error in these settings. Nevertheless, from the examples on logistic regression ((87) from Theorem 8, Corollary 6), linear regression (Theorem 9), and nonlinear regression (Corollary 7), we see that the MER upper bounds obtained from the second method depend on n only through $R_2(W|X, Z^n)$. This explicitly shows how the model uncertainty translates to the epistemic uncertainty and contributes to the overall prediction uncertainty. The definition of MER provides such a principled way to define different notions of uncertainties in Bayesian learning, and its study guides the analysis and estimation of these uncertainties, which is an increasingly important direction of research with wide applications.

B. MER in Bayesian Learning vs. Excess Risk in Frequentist Learning

The distinguishing feature of Bayesian learning is that the generative model of data is assumed to be drawn from a known model family according to some prior distribution, while there is virtually no restrictions on the admissible decision rules. As a result, the MER in Bayesian learning is determined by how accurate the model can be estimated, and there is no notion of *approximation error* unless the model family or the prior distribution is misspecified. This stands in contrast to the frequentist formulation of statistical learning where the data-generating model is assumed to be completely unknown, but the set of admissible decision rules is restricted, and the excess risk consists of an estimation error part and an approximation error part [62], [63]. In the discussion on multiple model families in Section V-A, it is shown that the MER there not only depends on the accuracy of the model parameter estimation within a fixed model, but also on a diameter-like term that upper-bounds the penalty incurred by a wrong estimate of the model index. The latter quantity may be viewed as an analogue of the approximation error in the frequentist setting. Its impact on the MER vanishes as the data size increases though, as its prefactor, which is the error probability of model index estimation, would eventually go to zero.

Despite of the different problem formulations, some MER upper bounds obtained in this paper share a similar form with the excess risk bounds in frequentist learning. One example is when the model is realizable, it is shown in Corollary 5 that the MER for Bayesian binary classification is $O(d/n)$, where d is the VC dimension of the generative function class. This result shares the same form as the "fast rate" results in frequentistic learning [64], where the distribution is completely unknown, but the hypothesis space, which is the set of decision rules, has a VC dimension d. Another example would be the identical expressions shared by the MER-information relationship in Corollary 4 and the generalization-information relationship in the frequentist setting of [14]. These results show the important roles played by information-theoretic quantities in the theory of statistical learning.

APPENDIX A
MISCELLANEOUS LEMMAS

1. Regularity Conditions for (30)

The regularity conditions for (30) to hold are listed here for completeness. These conditions are drawn from in [40, Section 2]. Let $\mathsf{W} \subset \mathbb{R}^d$ and assume that the densities of $P_{Z|w}$ exist with respect to the Lebesgue measure for all $w \in \mathsf{W}$. Also assume the parameter space W has a non-void interior and its boundary has a d-dimensional Lebesgue measure zero.

1) The density $p_{Z|W}(z|w)$ is twice continuously differentiable in w for almost every z; there exists $\delta(w)$ such that for every $j, k \in \{1, \ldots, d\}$,
$$\mathbf{E}\Big[\sup_{w':\|w'-w\|\le\delta(w)} \Big|\frac{\partial^2}{\partial w'_j \partial w'_k} \log p_{Z|W}(Z|w')\Big|^2\Big]$$
is finite and continuous in w; and for some $\xi > 0$, for each $j \in \{1, \ldots, d\}$,
$$\mathbf{E}\Big[\Big|\frac{\partial}{\partial w_j} \log p_{Z|W}(Z|w)\Big|^{2+\xi}\Big]$$
is finite and continuous as a function of w.
2) The following two definitions of Fisher information matrix are equal and positive definite:
$$[I_{Z|w}]_{j,k} = \mathbf{E}\Big[\frac{\partial}{\partial w_j} \log p_{Z|W}(Z|w) \frac{\partial}{\partial w_k} \log p_{Z|W}(Z|w)\Big],$$
and
$$[J_{Z|w}]_{j,k} = \Big[\frac{\partial^2}{\partial w'_j \partial w'_k} D_{\mathrm{KL}}(P_{Z|w}\|P_{Z|w'})\Big|_{w'=w}\Big].$$
When Condition 1) is satisfied, $[J_{Z|w}]_{j,k} = -\mathbf{E}\big[\frac{\partial^2}{\partial w_i \partial w_j} \log p_{Z|W}(Z|w)\big|W = w\big]$, and Condition 2) will be satisfied if $\int \frac{\partial^2}{\partial_i \partial_j} p_{Z|W}(z|w)\mathrm{d}z = 0$.
3) For $w \neq w'$, we have $P_{Z|w} \neq P_{Z|w'}$.
4) The prior on W is continuous and is supported on a compact subset of the interior of W.

The following lemma is from [40].

Lemma A1: Under the above conditions,

$$I(W;Z^n) = \frac{d}{2}\log\frac{n}{2\pi e} + h(W) + \frac{1}{2}\mathbf{E}\big[\log\det J_{Z|W}\big] + o(1) \quad \text{as } n\to\infty, \tag{189}$$

where $h(W)$ is the differential entropy of W, and the expectation is taken with respect to P_W.

2. A Transportation Inequality

The following lemma is adapted from [65, Lemma 4.18] and [46, Theorem 2].

Lemma A2: For distributions P and Q on a set U and a function $f:\mathsf{U}\to\mathbb{R}$, suppose there exists a function φ over $(0,b)$ with some $b\in(0,\infty]$ such that

$$\log\mathbf{E}_Q\big[e^{-\lambda(f(U)-\mathbf{E}_Q f(U))}\big] \le \varphi(\lambda), \quad \forall\, 0<\lambda<b. \tag{190}$$

Then

$$\mathbf{E}_Q[f(U)] - \mathbf{E}_P[f(U)] \le \varphi^{*-1}(D_{\mathrm{KL}}(P\|Q)), \tag{191}$$

where

$$\varphi^*(\gamma) = \sup_{0<\lambda<b} \lambda\gamma - \varphi(\lambda), \quad \gamma\in\mathbb{R} \tag{192}$$

is the Legendre dual of φ and φ^{*-1} is the inverse of φ^*, defined as

$$\varphi^{*-1}(x) = \sup\{\gamma\in\mathbb{R}:\varphi^*(\gamma)\le x\}, \quad x\in\mathbb{R}. \tag{193}$$

Proof of Lemma A2: The Donsker-Varadhan theorem states that

$$D_{\mathrm{KL}}(P\|Q) = \sup_{g:\mathsf{U}\to\mathbb{R}} \mathbf{E}_P[g(U)] - \log\mathbf{E}_Q[e^{g(U)}], \tag{194}$$

which implies that

$$\begin{aligned} &D_{\mathrm{KL}}(P\|Q)\\ &\ge \sup_{0<\lambda<b} \lambda(\mathbf{E}_Q[f(U)] - \mathbf{E}_P[f(U)]) - \\ &\qquad \log\mathbf{E}_Q[e^{-\lambda(f(U)-\mathbf{E}_Q f(U))}] &(195)\\ &\ge \sup_{0<\lambda<b} \lambda(\mathbf{E}_Q[f(U)] - \mathbf{E}_P[f(U)]) - \varphi(\lambda) &(196)\\ &= \varphi^*(\mathbf{E}_Q[f(U)] - \mathbf{E}_P[f(U)]). &(197)\end{aligned}$$

Consequently, from the definition in (193),

$$\mathbf{E}_Q[f(U)] - \mathbf{E}_P[f(U)] \le \varphi^{*-1}(D_{\mathrm{KL}}(P\|Q)), \tag{198}$$

which proves (191). □

3. Series With Growth Rate $\log n$

The following lemma is a restatement of [41, Lemma 6].

Lemma A3: Suppose $(a_1,a_2,\ldots)$ and $(b_1,b_2,\ldots)$ are two sequences of real numbers such that $a_n=\sum_{i=1}^n b_i$ for all n. Then

$$\lim_{n\to\infty}\frac{a_n}{\log n} = \lim_{n\to\infty} nb_n, \tag{199}$$

whenever both limits exist.

4. Convexity of $\mathcal{W}_p^p(P,Q)$

Lemma A4: The pth power of the p-Wasserstein distance is jointly convex in its two arguments, i.e. $\mathcal{W}_p^p(P,Q)$ is convex in (P,Q).

Proof: By definition,

$$\mathcal{W}_p^p(P,Q) = \inf_{\Pi(P,Q)} \mathbf{E}_{(X,Y)\sim\Pi}[\|X-Y\|^p]. \tag{200}$$

For arbitrary (P_1,Q_1), (P_2,Q_2), and $\gamma\in[0,1]$, let Π_1 and Π_2 be the optimal couplings for $\mathcal{W}_p^p(P_1,Q_1)$ and $\mathcal{W}_p^p(P_2,Q_2)$ respectively. Then

$$\begin{aligned} &\mathcal{W}_p^p(\gamma P_1+(1-\gamma)P_2, \gamma Q_1+(1-\gamma)Q_2) &(201)\\ \le&\mathbf{E}_{(X,Y)\sim\gamma\Pi_1+(1-\gamma)\Pi_2}[\|X-Y\|^p] &(202)\\ =&\gamma\mathbf{E}_{(X,Y)\sim\Pi_1}[\|X-Y\|^p] + \\ &(1-\gamma)\mathbf{E}_{(X,Y)\sim\Pi_2}[\|X-Y\|^p] &(203)\\ =&\gamma\mathcal{W}_p^p(P_1,Q_1) + (1-\gamma)\mathcal{W}_p^p(P_2,Q_2), &(204)\end{aligned}$$

where the first inequality is because $\gamma\Pi_1+(1-\gamma)\Pi_2$ is a coupling of $\gamma P_1+(1-\gamma)P_2$ and $\gamma Q_1+(1-\gamma)Q_2$. This shows the convexity of $\mathcal{W}_p^p(P,Q)$ in (P,Q). □

5. Lipschitz Continuity of Multivariate Function

The following lemma states a sufficient condition for a multivariate function to be Lipschitz continuous [66].

Lemma A5: Suppose a function $f:\mathbb{R}^n\to\mathbb{R}$ is continuously differentiable everywhere in a convex set $\mathsf{X}\subset\mathbb{R}^n$. If $c>0$ is such that $\|\nabla f(x)\|\le c$ for all $x\in\mathsf{X}$, then $|f(y)-f(x)|\le c\|y-x\|$ for all $x,y\in\mathsf{X}$.

6. Proof of Lemma 6

With the log loss, the generalized entropy of discrete Y is the Shannon entropy. We have

$$\begin{aligned} &H(P)-H(Q)\\ &= \mathbf{E}_P[-\log P(U)] - \mathbf{E}_Q[-\log Q(U)] &(205)\\ &\le \mathbf{E}_P[-\log Q(U)] - \mathbf{E}_Q[-\log Q(U)] &(206)\\ &= \sum_{u\in\mathsf{U}}(P(u)-Q(u))(-\log Q(u)) &(207)\\ &\le (-\log\min_{u\in\mathsf{U}} Q(u)) d_{\mathrm{TV}}(P,Q), &(208)\end{aligned}$$

where the first inequality follows from the fact that $H(P)=\inf_Q \mathbf{E}_P[-\log Q(U)]$, and the last inequality follows from the fact that $-\log Q(u)\in[0,-\log\min_{u\in\mathsf{U}}Q(u)]$ and the dual representation of the total variation distance.

For the zero-one loss, the generalized entropy of discrete Y is one minus the maximum probability. We have

$$\begin{aligned} &(1-\max_{u\in\mathsf{Y}}P(u)) - (1-\max_{u\in\mathsf{U}}Q(u))\\ &= \max_{u\in\mathsf{Y}}Q(u) - \max_{u\in\mathsf{Y}}P(u) &(209)\\ &\le Q(u_{\max}) - P(u_{\max}) &(210)\\ &\le d_{\mathrm{TV}}(Q,P) &(211)\end{aligned}$$

where in (210), $u_{\max}:=\arg\max_{u\in\mathsf{Y}}Q(u)$; (211) follows from the fact that $d_{\mathrm{TV}}(Q,P)=\sup_{E\subset\mathsf{U}}Q[E]-P[E]$ for any pair of distributions on U. The claim follows from the fact that the total variation distance is symmetric.

7. Proof of Lemma 8

First note that according to the definition of the $\mathcal{W}_2$ distance, $\mathbf{E}[U^2] = \mathcal{W}_2^2(P_U, \delta_0)$ and $\mathbf{E}[V^2] = \mathcal{W}_2^2(P_V, \delta_0)$, where δ_0 denotes the point mass at 0. Then

$$\begin{aligned}
&\mathsf{var}[U] - \mathsf{var}[V] \\
&= \mathbf{E}[U^2] - \mathbf{E}[V^2] + (\mathbf{E}[U] + \mathbf{E}[V])(\mathbf{E}[V] - \mathbf{E}[U]) \\
&\le (\mathcal{W}_2^2(P_U, \delta_0) - \mathcal{W}_2^2(P_V, \delta_0)) + \\
&\quad |\mathbf{E}[U] + \mathbf{E}[V]| \mathcal{W}_1(P_U, P_V) \\
&\le (\mathcal{W}_2(P_U, \delta_0) + \mathcal{W}_2(P_V, \delta_0)) |\mathcal{W}_2(P_U, \delta_0) - \\
&\quad \mathcal{W}_2(P_V, \delta_0)| + |\mathbf{E}[U] + \mathbf{E}[V]| \mathcal{W}_2(P_U, P_V) \qquad (212) \\
&\le (\sqrt{\mathbf{E}[U^2]} + \sqrt{\mathbf{E}[V^2]}) \mathcal{W}_2(P_U, P_V) + \\
&\quad |\mathbf{E}[U] + \mathbf{E}[V]| \mathcal{W}_2(P_U, P_V) \qquad (213) \\
&\le 2(\sqrt{\mathbf{E}[U^2]} + \sqrt{\mathbf{E}[V^2]}) \mathcal{W}_2(P_U, P_V) \qquad (214)
\end{aligned}$$

where we have used the triangle inequality for the $\mathcal{W}_2$ distance and the fact that

$$|\mathbf{E}[U] - \mathbf{E}[V]| \le \mathcal{W}_1(P_U, P_V) \le \mathcal{W}_2(P_U, P_V). \qquad (215)$$

Appendix B
MER Upper Bound for Linear Regression Based on Theorem 6

To make use of Theorem 6 for linear regression with the quadratic loss, let $Y = W^\top \phi(X) + V$ where $V \sim \mathcal{N}(0, \sigma^2)$, and assume X is independent of W. In addition, let W' be sampled from $P_{W|Z^n}$ independently of everything else. Since $\psi^*(X, W') = W'^\top \phi(X)$, we have

$$\begin{aligned}
(Y - \psi^*(X, W'))^2 &= (Y - W'^\top \phi(X))^2 \\
&= ((W - W')^\top \phi(X) + V)^2. \qquad (216)
\end{aligned}$$

Since W' is a conditionally i.i.d. copy of W given (X, Z^n), it can be seen that the conditional distribution of $(W - W')^\top \phi(X)$ given $(X, Z^n) = (x, z^n)$ is Gaussian with zero mean and variance $2\phi(x)^\top C_{W|z^n} \phi(x)$. It follows that conditional on $(X, Z^n) = (x, z^n)$, $(Y - \psi^*(x, W'))^2$ has the same distribution as $\big(2\phi(x)^\top C_{W|z^n} \phi(x) + \sigma^2\big) U^2$, where $U \sim \mathcal{N}(0, 1)$. As a consequence of the fact that

$$\begin{aligned}
&\log \mathbf{E}[e^{-\lambda(\sigma_\chi^2 U^2 - \mathbf{E}[\sigma_\chi^2 U^2])}] \\
&= \lambda \sigma_\chi^2 - \frac{1}{2} \log(1 + 2\sigma_\chi^2 \lambda) \\
&\le \sigma_\chi^4 \lambda^2 := \varphi(\lambda) \quad \text{for } \lambda > 0, \qquad (217)
\end{aligned}$$

the fact that $\varphi^{*-1}(\gamma) = 2\sigma_\chi^2 \sqrt{\gamma}$, the assumption that $\sup_{x, x^n} \phi(x)^\top C_{W|z^n} \phi(x) \le b$, the fact that $I(W; Y|X, Z^n) = \mathbf{E}[\frac{1}{2} \log(1 + \phi(X)^\top C_{W|Z^n} \phi(X)/\sigma^2)]$, and Theorem 6, we have

$$\begin{aligned}
&\mathrm{MER}_2 \\
&\le 2(2b + \sigma^2) \sqrt{\frac{1}{2} \log\Big(1 + \frac{1}{\sigma^2} \mathbf{E}\big[\phi(X) C_{W|Z^n} \phi(X)\big]\Big)} \qquad (218) \\
&\le 2(2b + \sigma^2) \\
&\quad \sqrt{\frac{1}{2} \log\Big(1 + \frac{1}{\sigma^2} \mathbf{E}[\|\phi(X)\|^2] \mathbf{E}[\mathrm{tr}(C_{W|X^n, Y^n})]\Big)}. \qquad (219)
\end{aligned}$$

Appendix C
Proof of Theorem 15

From the fact that

$$\begin{aligned}
P_{Y|x,z^n} = &\sum_{m' \in \mathsf{M}} P_{M|X,Z^n}(m'|x, z^n) \\
&\int_{\mathsf{W}_{m'}} P_{W|M,X,Z^n}(\mathrm{d}w'|m', x, z^n) K_{Y|x,w',m'}
\end{aligned}$$

and the convexity assumption of the statistical distance under consideration, the proof of the first inequality essentially follows the same steps of the proof of Lemma 5.

The second inequality is based on the first one, and can be shown as

$$\begin{aligned}
&\mathbf{E}\big[D(P_{Y|X,Z^n}, K_{Y|X,W,M})\big] \qquad (220) \\
&\le \mathbf{E}\big[D(K_{Y|X,W',M'}, K_{Y|X,W,M})\big] \qquad (221) \\
&= \sum_{m \in \mathsf{M}} P_M(m) \int_{\mathsf{W}_m} P_{W|M}(\mathrm{d}w|m) \cdot \\
&\quad \int_{\mathsf{X} \times \mathsf{Z}^n} P_{X,Z^n|W,M}(\mathrm{d}x, \mathrm{d}z^n|w, m) \cdot \\
&\quad \sum_{m' \in \mathsf{M}} P_{M|X,Z^n}(m'|x, z^n) \cdot \\
&\quad \int_{\mathsf{W}_{m'}} P_{W|X,Z^n,M}(\mathrm{d}w'|x, z^n, m') \cdot \\
&\quad D(K_{Y|x,w',m'}, K_{Y|x,w,m}) \qquad (222) \\
&= S_1 + S_2 \qquad (223)
\end{aligned}$$

where the last step is to split the summation over m' such that

$$\begin{aligned}
S_1 = &\sum_{m \in \mathsf{M}} P_M(m) \int_{\mathsf{W}_m} P_{W|M}(\mathrm{d}w|m) \\
&\int_{\mathsf{X} \times \mathsf{Z}^n} P_{X,Z^n|W,M}(\mathrm{d}x, \mathrm{d}z^n|w, m) \\
&P_{M|X,Z^n}(m|x, z^n) \int_{\mathsf{W}_m} P_{W|X,Z^n,M}(\mathrm{d}w'|x, z^n, m) \\
&D(K_{Y|x,w',m}, K_{Y|x,w,m}) \qquad (224) \\
\le &\sum_{m \in \mathsf{M}} P_M(m) \int_{\mathsf{W}_m} P_{W|M}(\mathrm{d}w|m) \\
&\int_{\mathsf{X} \times \mathsf{Z}^n} P_{X,Z^n|W,M}(\mathrm{d}x, \mathrm{d}z^n|w, m) \\
&\int_{\mathsf{W}_m} P_{W|X,Z^n,M}(\mathrm{d}w'|x, z^n, m) \\
&D(K_{Y|x,w',m}, K_{Y|x,w,m}) \qquad (225) \\
= &\mathbf{E}[D(K_{Y|X,W',M}, K_{Y|X,W,M})] \qquad (226)
\end{aligned}$$

and

$$\begin{aligned}
S_2 = &\sum_{m \in \mathsf{M}} P_M(m) \int_{\mathsf{W}_m} P_{W|M}(\mathrm{d}w|m) \\
&\int_{\mathsf{X} \times \mathsf{Z}^n} P_{X,Z^n|W,M}(\mathrm{d}x, \mathrm{d}z^n|w, m) \cdot \\
&\sum_{m' \ne m} P_{M|X,Z^n}(m'|x, z^n) \\
&\int_{\mathsf{W}_{m'}} P_{W|X,Z^n,M}(\mathrm{d}w'|x, z^n, m')
\end{aligned}$$

$$D(K_{Y|x,w',m'}, K_{Y|x,w,m}) \tag{227}$$
$$\leq \Big(\max_{m,m' \in \mathsf{M}, m \neq m'} \sup_{w \in \mathsf{W}_m, w' \in \mathsf{W}_{m'}} \sup_{x \in \mathsf{X}} D(K_{Y|x,w',m'}, K_{Y|x,w,m}) \Big) \sum_{m \in \mathsf{M}} P_M(m) \int_{\mathsf{W}_m} P_{W|M}(\mathrm{d}w|m) \int_{\mathsf{X},\mathsf{Z}^n} P_{X,Z^n|W,M}(\mathrm{d}x, \mathrm{d}z^n|w,m) \sum_{m' \neq m} P_{M|X,Z^n}(m'|x,z^n) \tag{228}$$
$$= \mathrm{diam}(\mathbb{M}, D) \mathbb{P}[M' \neq M] \tag{229}$$
$$\leq 2\mathrm{diam}(\mathbb{M}, D) R_{01}(M|X, Z^n) \tag{230}$$

where the last step follows from Lemma 4 applied to the zero-one loss.

ACKNOWLEDGMENT

The authors would like to thank Yihong Wu for insightful comments on an early draft of this work; Lemma 8 is given by him. They are also thankful to Max Welling and Auke Wiggers for discussions on different notions of uncertainties in Bayesian learning. The comments from the anonymous reviewers of the IEEE TRANSACTIONS ON INFORMATION THEORY greatly improved the quality of this article, they are grateful to the reviewers and the area chair.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Germany: Springer, 2006.
[2] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*. London, U.K.: Chapman & Hall, 2011.
[3] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 681–688.
[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.
[5] R. M. Neal, *Bayesian Learning for Neural Networks*. Berlin, Germany: Springer, 1996.
[6] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.
[7] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016.
[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge, MA, USA: MIT Press, 2006.
[9] C. Louizos, K. Ullrich, and M. Welling, "Bayesian compression for deep learning," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017.
[10] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017.
[11] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *Proc. ICML*, 2018, pp. 1184–1193.
[12] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction," 2019, *arXiv:1910.09457*.
[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017.
[14] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017.
[15] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.
[16] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.
[17] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, pp. 2451–2492, 1997.
[18] J. Baxter, "A Bayesian/information theoretic model of learning to learn via multiple task sampling," *Mach. Learn.*, vol. 28, no. 1, pp. 7–39, Jul. 1997.
[19] D. Haussler, M. Kearns, and R. E. Schapire, "Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension," *Mach. Learn.*, vol. 14, no. 1, pp. 83–113, Jan. 1994.
[20] L. L. Cam and G. L. Yang, *Asymptotics in Statistics Some Basic Concepts*, 2nd ed. New York, NY, USA: Springer, 2000.
[21] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart, "Convergence rates of posterior distributions," *Ann. Statist.*, vol. 28, no. 2, pp. 500–531, Apr. 2000.
[22] S. Ghosal and A. van der Vaart, "Convergence rates of posterior distributions for noniid observations," *Ann. Statist.*, vol. 35, no. 1, pp. 192–223, Feb. 2007.
[23] N. G. Polson and V. Ročková, "Posterior concentration for sparse deep learning," in *Proc. Conf. Neural Inf. Process. Syst.*, 2018.
[24] A. R. Barron, "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems," in *Bayesian Statistics 6*. London, U.K.: Oxford Univ. Press, 1998.
[25] D. McAllester, "PAC-Bayesian stochastic model selection," *Mach. Learn.*, vol. 51, no. 1, pp. 5–21, 2003.
[26] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proc. 10th Annu. Conf. Comput. Learn. Theory (COLT)*, 1997, pp. 2–9.
[27] T. Zhang, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1307–1321, Apr. 2006.
[28] R. Meir and T. Zhang, "Generalization error bounds for Bayesian mixture algorithms," *J. Mach. Learn. Res.*, vol. 4, pp. 839–860, Oct. 2003.
[29] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *Ann. Statist.*, vol. 32, no. 4, pp. 1367–1433, Aug. 2004.
[30] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *Ann. Math. Stat.*, vol. 33, no. 2, pp. 404–419, 1962.
[31] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Proc. Conf. Neural Inf. Process. Syst.*, 2016.
[32] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. Cham, Switzerland: Springer, 2002.
[33] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.
[34] Y. Wu and S. Verdu, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1289–1301, Mar. 2012.
[35] I. J. Good, "On the principle of total evidence," *Brit. J. Philosophy Sci.*, vol. 17, no. 4, pp. 319–321, Feb. 1967.
[36] B. Skyrms, *The Value of Knowledge*. Minneapolis, MN, USA: Univ. Minnesota Press, 1990.
[37] C. S. Qazaz, C. K. I. Williams, and C. M. Bishop, "An upper bound on the Bayesian error bars for generalized linear regression," in *Mathematics of Neural Networks: Models, Algorithms and Applications*. Boston, MA, USA: Springer, 1997, pp. 295–299.
[38] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.
[39] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[40] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Stat. Planning Inference*, vol. 41, no. 1, pp. 37–60, Aug. 1994.
[41] D. Haussler and M. Opper, "General bounds on the mutual information between a parameter and n conditionally independent observations," in *Proc. 8th Annu. Conf. Comput. Learn. Theory (COLT)*, 1995, pp. 402–411.
[42] B. Hajek, *Random Processes for Engineers*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[43] N. Sauer, "On the density of families of sets," *J. Combinat. Theory A*, vol. 13, no. 1, pp. 145–147, Jul. 1972.

[44] S. Shelah, "A combinatorial problem; stability and order for models and theories in infinitary languages," *Pacific J. Math.*, vol. 41, no. 1, pp. 247–261, Apr. 1972.
[45] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, 2016, pp. 1232–1240.
[46] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1475–1479.
[47] H. Hafez-Kolahi, B. Moniri, S. Kasaei, and M. S. Baghshah, "Rate-distortion analysis of minimum excess risk in Bayesian learning," in *Proc. Int. Conf. Mach. Learn.*, 2021.
[48] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Proc. Conf. Learn. Theory*, 2020.
[49] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020.
[50] J. Liu, P. Cuff, and S. Verdu, "On α-decodability and α-likelihood decoder," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017.
[51] A. Bhatt, J.-T. Huang, Y.-H. Kim, J. J. Ryu, and P. Sen, "Variations on a theme by Liu, Cuff, and Verdú: The power of posterior sampling," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.
[52] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.
[53] C. Villani, *Topics Optim. Transp.* (Graduate Studies in Mathematics), vol. 58. Providence, RI, USA: American Mathematical Society, 2003.
[54] A. Xu, "Continuity of generalized entropy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2246–2251.
[55] A. Xu, "Continuity of generalized entropy and statistical learning," *IEEE Trans. Inf. Theory*, 2021.
[56] Y. Polyanskiy and Y. Wu, "Wasserstein continuity of entropy and outer bounds for interference channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3992–4002, Jul. 2016.
[57] Y. Wu, personal communication, 2019.
[58] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geometric Funct. Anal.*, vol. 6, no. 3, pp. 587–600, May 1996.
[59] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[60] C. Fang, H. Dong, and T. Zhang, "Mathematical models of overparameterized neural networks," *Proc. IEEE*, vol. 109, no. 5, pp. 683–703, May 2021.
[61] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques-an overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, Nov. 2018.
[62] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Cham, Switzerland: Springer, 1996.
[63] S. Shalev-Shwartz and S. Ben-David, *Understand. Mach. Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[64] P. D. Grünwald and N. A. Mehta, "Fast rates for general unbounded loss functions: From ERM to generalized Bayes," *J. Mach. Learn. Res.*, vol. 21, no. 56, pp. 1–80, 2020.
[65] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.
[66] H. F. Walker, *Lecture Notes of MA500: Basic Real Analysis*. Worcester, MA, USA: Worcester Polytechnic Institute, 2013.

Aolin Xu received the B.S. degree from Beijing University of Posts and Telecommunications in 2007, the M.S. degree from Tsinghua University in 2010, and the Ph.D. degree from the University of Illinois at Urbana-Champaign in 2016, all in electrical engineering. His research interests include statistical inference and learning, information theory, and decision making under uncertainty.

Maxim Raginsky (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Northwestern University, Evanston, IL, USA, in 2000, 2000, and 2002, respectively. He has held research positions with Northwestern University and the University of Illinois at Urbana-Champaign, Urbana, IL, USA, where he was a Beckman Foundation Fellow from 2004 to 2007, and Duke University, Durham, NC, USA. In 2012, he returned to UIUC, where he is currently a William L. Everitt Fellow and an Associate Professor with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory. His research interests are in probability and stochastic processes, deterministic and stochastic control, machine learning, optimization, and information theory. He was a recipient of the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2013. He has served on editorial boards of IEEE TRANSACTIONS ON INFORMATION THEORY, *Foundations and Trends in Communications and Information Theory*, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. He is currently a member of the editorial boards of *SIAM Journal on Mathematics of Data Science*, *Journal of Machine Learning Research*, and *Mathematics of Control, Signals, and Systems*.