Rate-Distortion Analysis of Minimum Excess Risk in Bayesian Learning

Hassan Hafez-Kolahi [1] Behrad Moniri [2] Shohreh Kasaei [1] Mahdieh Soleymani Baghshah [1]

Abstract

In parametric Bayesian learning, a prior is assumed on the parameter W which determines the distribution of samples. In this setting, Minimum Excess Risk (MER) is defined as the difference between the minimum expected loss achievable when learning from data and the minimum expected loss that could be achieved if W was observed. In this paper, we build upon and extend the recent results of (Xu & Raginsky, 2020) to analyze the MER in Bayesian learning and derive information-theoretic bounds on it. We formulate the problem as a (constrained) rate-distortion optimization and show how the solution can be bounded above and below by two other rate-distortion functions that are easier to study. The lower bound represents the minimum possible excess risk achievable by *any* process using R bits of information from the parameter W. For the upper bound, the optimization is further constrained to use R bits from the training set, a setting which relates MER to information-theoretic bounds on the generalization gap in frequentist learning. We derive information-theoretic bounds on the difference between these upper and lower bounds and show that they can provide order-wise tight rates for MER under certain conditions. This analysis gives more insight into the information-theoretic nature of Bayesian learning as well as providing novel bounds.

1. Introduction

One of the main problems studied in statistical learning theory is the excess risks of learning algorithms, which is the gap between the achieved error and the best possible error if the distribution was known (LeCam et al., 1973; Assouad, 1983; Keener, 2010). An interesting question in this regard is to study lower bounds on the excess risk which could be achieved by any algorithm. This concept is usually studied in the frequentist setting, in which a family of distributions is assumed and minimax bounds are derived to study if an algorithm which only has access to n samples from the distribution, can work well for all the distributions in the family.

Recently, (Xu & Raginsky, 2020) proposed a framework to define and study Minimum Excess Risk (MER) in Bayesian learning. In the Bayesian learning, it is assumed that the underlying distribution is described by a variable $W \in \mathcal{W}$ and a prior P_W is considered which describes the probability of any W before observing data. The joint distribution of W, the training set $Z^n = \{(X_i, Y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, and a test sample $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$, is described as $P_W \otimes (P_Z^W)^n \otimes P_Z^W$. The goal is to find a function $\hat{h} : \mathcal{X} \to \mathcal{Y}$ after observing the training set, in a way that $\mathbb{E}[\ell(Y, \hat{h}(X))]$ is small, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function.

In order to quantify the hardness of a problem, (Xu & Raginsky, 2020) define Minimum Excess Risk as the gap between the expected error of the best algorithm which only has access to the data and the minimum expected error if W was also observed. They show that for a variety of loss functions, the conditional mutual information $I(W; Y|Z^n, X)$ appears in the upper bounds on MER. When $W \in \mathbb{R}^p$, using information-theoretic results on the rate of $I(W; Z^n)$, they achieve bound of $O(\sqrt{p/n})$ on MER as $n \to \infty$, given that some assumptions on the distribution and loss hold (see Section 4). They also show that the bound can be improved to $O(p/n)$ for two specific losses: logarithmic loss and quadratic loss (for bounded $\mathcal{Y}$). They left the study of lower bounds on MER as an open problem.

In this work, we adapt a source coding view on learning and introduce a (variant of) rate-distortion optimization which captures the notion of MER. Then, we demonstrate how the constraint on this rate-distortion minimization can *naturally* be weakened and strengthened to achieve lower and upper bounds, respectively. These lower and upper bounds are easier to study and we derive a variety of results on them. In particular, we study the lower bound with tools from the source coding theory, and demonstrate that under some conditions, MER is lower bounded by $\Omega(p/n)$. This concludes

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran [2] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran. Correspondence to: Shohreh Kasaei <kasaei@sharif.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

the rate analysis of MER for the cases in which the matching upper bound $O(p/n)$ exists and show that both upper and lower bounds are order-wise tight as $n \to \infty$. As an important example, we show that the bounds are order-wise tight for quadratic loss when $\mathcal{Y}$ is bounded and the distribution is suitably *smooth* (see Section 6).

The studied rate distortion problems (original, upper bound, and lower bound) all have interesting interpretations and might be of interest by themselves.

1.1. An Appetizer for the Rate-Distortion View

Loosely speaking, the main idea behind the rate-distortion view developed in this paper is as follows. The variable W is first generated and the dataset Z^n is generated from W. Our goal is to observe Z^n and find $\hat{h}(x)$ which performs well compared to the case where W is known. Thus, if we could decode W from Z^n perfectly, an MER equal to zero would be achieved. However, in almost all applications of interest, it is impossible to find the exact value of W, since the information we can extract from W to build $\hat{h}$, is bounded above by $I(W; Z^n)$. In particular, if W is continuous, an infinite number of samples are needed for its exact decoding. But of course, we don't need full recovery of W to get a *good enough* $\hat{h}$. So the question is how good we can act, based on a suitable distortion function, if only $I(W; Z^n)$ nats of information about W is received *through* Z^n.

This line of reasoning makes it natural to study the problem as a rate-distortion optimization. But there is a challenge in using rate-distortion theory to study the learning problem: in an standard rate-distortion problem, we can decide how we encode W, but in learning problems, a (random) preprocess $W \to Z^n$ is also enforced. If we remove this constraint, we will have a standard rate-distortion problem. Since the feasible set is enlarged, this gives us a lower bound on the original minimization. It is not obvious how efficient this preprocess acts; i.e., if one is asked to use $R = I(W; Z^n)$ nats to represent W by an intermediate variable Ξ (in an arbitrary space of choice) in a way that it is possible to recover a good $\hat{h}$, is it a good idea to just generate n i.i.d. samples from P_{XY}^W; i.e., use $\Xi = Z^n$? We will try to answer such questions by quantifying and studying lower bounds.

On the other hand, there is an interesting question which is answered by studying an upper bound on the rate-distortion function. If we know that only $R = I(W; Z^n)$ nats about W are present in the dataset Z^n, we might hope to be able to just extract those nats and don't rely on Z^n more than necessary. To quantify this idea and study it, we can also restrict $I(Z^n; \hat{h}) \leq R = I(W; Z^n)$ and see if we can still find an $\hat{h}$ with a good performance? This scenario is similar to the frequentist approach of model compression in which the mutual information between the training set and the learned model is restricted to control the generalization gap. To better understand the information-theoretic properties of learning, we will study these rate-distortion functions and derive information-theoretic bounds on their difference. In particular we will show that (under some smoothness conditions) all three rate-distortion functions converge as $n \to \infty$. The rates of convergence are also derived which shows that the bounds are order-wise tight for quadratic loss under certain conditions. We also provide non-asymptotic information-theoretic bounds which explain the difference between these rate-distortion functions for finite samples.

While rate-distortion theory was used before in learning theoretic settings (see Section 2), the systematic view developed in this paper as well as the derived bounds are novel to the best of our knowledge.

1.2. Outline of the Paper

In Section 2, the related literature is discussed. The notations are introduced in Section 3. Section 4 is devoted to information-theoretic upper bounds on MER. In Section 5, the main results of the paper on the rate-distortion view of MER are presented. In Section 6, some applications of the developed tools are studied. Finally, conclusions and future works are presented in Section 7. The proofs of theorems are presented in the supplementary materials.

2. Related Work

Deriving minimax bounds on excess risk in the frequentist setting is a well studied problem. The Le Cam's and Assouad's methods are two of the most widely used approaches for deriving lower bounds on minimax risk (LeCam et al., 1973; Assouad, 1983). Fano's method is also a popular method based on Fano's lower bound on the error probability in an M-ary testing problem (Yang & Barron, 1999).

In Bayesian learning, one of the tracks which is related to MER, is the convergence of posterior to the true parameter (Ghosal et al., 2000; Shen et al., 2001; Ghosal et al., 2007; Le Cam & Yang, 2012). The main difference between this line of works and MER studied by (Xu & Raginsky, 2020) is that the former tries to analyze Bayesian inference from a frequentist perspective, while in latter, W is still considered as random in the analysis. The information-theoretic results used in this new setting as well as the definition of MER itself (which is based on an expectation on P_W) are influenced by this view on the problem. Moreover, the subject of study in previous works is usually estimation of the parameter while in MER, the Bayes risk for the random variable of interest is directly studied. Results from convergence of posterior are useful to derive bounds on MER (e.g. see Theorem 7 of current paper as well as Section 4 of (Xu & Raginsky, 2020)).

These recent results of Xu & Raginsky (2020) can be seen as extensions to the universal prediction of (Merhav & Feder, 1998). In universal prediction, the accumulated loss on a sequence of samples is studied (using an approach similar to universal source coding). In contrast, this new treatment allows the analysis of the supervised setting with general subgaussian loss. Moreover, they utilize a refined treatment which yields direct bounds on the error of estimating the label of the test sample (the last sample) when the training set is given.

Another track which heavily influences the current work, is the recent series of results in frequentist learning which use mutual information between the learned model and the dataset to control the generalization gap (Russo & Zou, 2015; 2016; Xu & Raginsky, 2017; Bassily et al., 2018; Asadi et al., 2018; Steinke & Zakynthinou, 2020; Hafez-Kolahi et al., 2020). While the setting in these works is different, the mathematical tools developed to derive information-theoretic bounds are similar. In particular, we use ideas from (Steinke & Zakynthinou, 2020) to derive new bounds on MER. Moreover, the upper bound on rate-distortion function studied in this paper, which is based on $I(Z^n;\hat{h})$, is directly related to this frequentist setting (see Section 5.4). A relevant work on this setting is (Bu et al., 2020) which used model compression to produce $\tilde{h}$ from $\hat{h}$ to control the generalization gap.

In (Gao et al., 2019), a rate-distortion optimization is utilized to understand the limits of model compression, and results for linear models are derived. Their setting is similar to a loosened variant of the rate-distortion lower bound studied in this paper where the solution is restricted to be from the parametric family (see Section 6).

In (Nokleby et al., 2016), upper bounds on the expected excess risk are derived for certain class of problems which satisfy a notion of "interpolation set". Their Bayesian treatment of the parameter makes their results comparable to the setting of (Xu & Raginsky, 2020) when studying zero-one loss.

3. Notation and Preliminaries

Random variables and their realizations are represented with uppercase and lowercase letters respectively; e.g., $x \in \mathcal{X}$ is a realization of random variable X. Conditional distributions and expectations are identified with superscripts; e.g., P_X^z indicates the conditional distribution of X given $Z = z$ and $\mathbb{E}_X^z[f(X,z)]$ indicate the expectation of $f(X,z)$ based on this distribution. $\mathrm{KL}(P_X \parallel Q_X) = \int \log \frac{P(X)}{Q(X)} dP$ is the KL divergence of distribution P_X from Q_X. Mutual information is defined as $I(X;Y) = \mathrm{KL}(P_{XY} \parallel P_X \otimes P_Y)$ where P_X and P_Y are marginal distributions of P_{XY}. The conditional mutual information is defined as $I(X;Y|Z) = \mathbb{E}_Z[I^Z(X;Y)]$ in which for all z, $I^z(X;Y)$ is the mutual information on the conditioned distributions P_{XY}^z, i.e. $I^z(X;Y) = \mathrm{KL}(P_{XY}^z \parallel P_X^z \otimes P_Y^z)$. Throughout the paper, all logarithms are in natural base and all information-theoretic quantities are in nats.

Given a distribution P_{XY} on a set $\mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the Bayes risk of estimating Y from X is denoted by

$$R_\ell(Y|X) = \inf_{\psi:\mathcal{X}\to\mathcal{Y}} \mathbb{E}[\ell(Y, \psi(X)].$$

It is assumed that the infimum is attained, and the optimal decision for a given x is represented by $\psi_{\ell,Y|X}^*(x)$, where we omit ℓ, if it is clear from the context.

4. MER and the Information-Theoretic Upper Bounds

As described in Section 1, a common scenario in Bayesian learning is to consider a prior distribution P_W on W and consider the joint distribution $P_W \otimes (P_{XY}^W)^n \otimes P_{XY}^W$ which generates W, Z^n, Z. Here, $Z^n = \{(X_i, Y_i)\}_{i=1}^n$ is the training set and $Z = (X, Y)$ is the test sample. Usually, it is also assumed that X is independent of W, and we have the distribution $P_{XY}^W = P_X \otimes P_Y^{XW}$, i.e., the unknown parameter is just used in describing the relation between X and Y. The goal is to predict Y when Z^n and X are given. The Bayes risk for this task is $R_\ell(Y|Z^n, X)$. If the parameter W was known, we could do better and achieve $R_\ell(Y|W, X)$. MER is defined as the expected extra price we should pay as a result of not knowing W; i.e.,

$$\mathrm{MER}_\ell^n = R_\ell(Y|Z^n, X) - R_\ell(Y|W, X). \tag{1}$$

Upper bounds on MER for various loss functions are studied in (Xu & Raginsky, 2020). Note that if we have a Markov chain $Y - U - V$ then there is a data processing inequality for Bayes risk; i.e.,

$$R_\ell(Y|U) \leq R_\ell(Y|V)$$

(see Lemma 1 of (Xu & Raginsky, 2020)). The following lemma, which is due to Theorem 4 of (Xu & Raginsky, 2020), gives an upper bound on the looseness of this inequality when the loss function is bounded.

Lemma 1. *Consider random variables Y, U and V forming Markov chain $Y - U - V$ and an arbitrary non-negative bounded function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, b]$. We have*

$$R_\ell(Y|V) - R_\ell(Y|U) \leq \sqrt{\frac{b^2}{2} I(Y;U|V)}. \tag{2}$$

Using this lemma, it is straightforward to derive upper bounds on MER_ℓ^n as

$$\begin{aligned}\mathrm{MER}_\ell^n &= R_\ell(Y|Z^n,X) - R_\ell(Y|W,X) \\ &= R_\ell(Y|Z^n,X) - R_\ell(Y|W,Z^n,X) \\ &\leq \sqrt{\frac{b^2}{2} I(Y;W|Z^n,X)} \qquad (3) \\ &\leq \sqrt{\frac{b^2}{2n} I(W;Z^n)}, \qquad (4)\end{aligned}$$

where the second equality is due to the fact that $Y \perp\!\!\!\perp Z^n|W,X$. The final inequality is proved by noting that $I(Z;W|Z^n)$ is a decreasing function of n and applying chain rule on $I(Z^n;W)$ (see the Proof of Theorem 2 in (Xu & Raginsky, 2020)). Note that the bound (3) could be much better than (4). The main reason is that it does not depend on $I(W;X)$ which could be large. The improvement one can achieve when using (3) instead of (4) is similar to using the *conditioning* technique to improve the information-theoretic generalization bounds in frequentist learning (Hafez-Kolahi et al., 2020). Actually, the same conditioning technique is at the heart of deriving the first bound (see proof of Lemma 1 in the supplementary materials). It is also worth noting that if the distribution on W is not known, but the capacity of channel $P_{Z^n}^W$ is limited, then $I(W;Z^n)$ is controlled and Equation (4) can be used to derive a redundancy-capacity result similar to universal prediction (Merhav & Feder, 1998).

The following lemma can be used along Lemma 1 to achieve convergence rates. This is a classic result on growth rate of mutual information between observations and the parameter which can be found in (Clarke & Barron, 1990; 1994).

Lemma 2. *If W is taking values in a p-dimensional compact subspace of $\mathbb{R}^p$, and the model P_Z^w is smooth in w, then as $n \to \infty$, we have*

$$I(W;Z^n) = \frac{p}{2}\log\big(\frac{n}{2\pi e}\big) + h(W) + \frac{\mathbb{E}\big[\log|J_Z^W(W)|\big]}{2} + o(1),$$

in which $|J_Z^W(w)|$ is the determinant of the Fisher information matrix about W contained in Z. Rigorous statement of the smoothness conditions can be found in the appendix.

Using this lemma, it can be shown that $I(Y;W|Z^n) = O(1/n)$ as $n \to \infty$. This gives us a rate of $O(\sqrt{1/n})$ on MER_ℓ^n for any bounded loss. In (Xu & Raginsky, 2020), it is also proved that for bounded quadratic loss and logarithmic loss the square root can be removed which improves the rate to $O(1/n)$.

4.1. Dropping the Square Root

Whether it is possible to drop the square root for a general bounded loss is an open problem. In this section we demonstrate that this is possible for the realizable case.

Lemma 3. *Consider random variables Y, U, and V forming Markov chain $Y - U - V$ and an arbitrary non-negative bounded function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,b]$. We have*

$$R_\ell(Y|V) \leq 2R_\ell(Y|U) + 3bI(Y;U|V). \qquad (5)$$

To prove this bound, a symmetrization technique that is used in (Steinke & Zakynthinou, 2020) to derive a variety of bounds on generalization gap, is adapted. For the case where $R_\ell(Y|U)$ is close to zero, this bound can give better results compared to Lemma 1. The mutual information $I(Y;U|V)$ can be unbounded in certain problems. In particular, in realizable setting, this can happen when the random variables are continuous and the relation between them is deterministic. Informally, this is due to the fact that $I(Y;U|V)$ quantifies the nats necessary for full recovery of Y (which could be unbounded for continuous random variables). This can be solved by covering the space of $\mathcal{Y}$ at different levels and adopting the chaining technique to acquire sharper bounds (e.g., see (Asadi et al., 2018) for an application of the chaining technique on information-theoretic generalization bounds). This is discussed in the supplementary materials.

5. Rate-Distortion Analysis of MER

Inequalities (4) and (3) give us information-theoretic *upper* bounds on MER. Lower bounding MER in Bayesian learning has been remained as an open problem (Xu & Raginsky, 2020). The tools we develop in this section let us study the lower bounds as well.

5.1. The Challenge of Finding a Lower Bound

First, it should be noted that it is not possible to have a matching lower bound in the form of $(\mathrm{MER})_\ell^n \geq \alpha\sqrt{I(Y;W|Z^n,X)}$ for some $\alpha > 0$. Loosely speaking, the reason is that it is possible that Y and W share many bits but those bits are not used in the loss function. In other words, the information contained in W about Y is not necessarily related to loss. As an example, consider the toy problem where $\mathcal{W} = \mathcal{Y} = [-2,2]^2$, and

$$\begin{cases} W_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(-1,1), \\ \epsilon_i \overset{\text{indep.}}{\sim} \mathrm{Unif}(-a_i,a_i), \\ Y = (Y_1,Y_2) = (W_1+\epsilon_1, W_2+\epsilon_2). \end{cases}$$

Define $\ell((y_1,y_2),(\hat{y}_1,\hat{y}_2)) = c_1(y_1-\hat{y}_1)^2 + c_2(y_2-\hat{y}_2)^2$. Here, a_1, a_2, c_1, and c_2 are hyper-parameters defining the problem. Now consider the extreme case where $a_1 = 0, a_2 = 1, c_1 = 1$, and $c_2 = 0$. In this case, the loss function is ignoring the second dimension of Y, which is the harder one to estimate. Actually, by observing a single sample, W_1 is found and $\forall n > 1, \mathrm{MER}_\ell^n = 0$. However, the mutual information $I(W;(Y_1,Y_2)|Z^n) \geq I(W;Y_2|Z^n)$ approaches zero only as $n \to \infty$.

Based on such observations, we argue that in order to derive MER lower bounds, the relation between the used rate and the loss function ℓ should be considered more carefully. This was one of the main motivations to define the MER as a rate-distortion problem. But, it is also insightful in itself to study Bayesian learning from a source coding perspective, as was discussed in Section 1.1.

5.2. Rate-Distortion Optimization

Rate-Distortion theory was introduced by (Shannon, 1948; 1959) to quantify the minimum average number of bits needed to transmit a random variable with a given maximum distortion. Let P_X be a distribution over $\mathcal{X}$ and $X^n = \{X_i\}_{i=1}^n$ be n i.i.d. samples of P_X. An encoder $f_n : \mathcal{X}^n \to \{1,2,\dots,2^{nR}\}$ maps the message into a codeword, and the decoder $g_n : \{1,2,\dots,2^{nR}\} \to \hat{\mathcal{X}}^n$, decodes the codeword. The distortion function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, measures the distortion and $d(X^n, \hat{X}^n)$ is the average distortion of X_i and $\hat{X}_i$s. For a given rate R, the rate-distortion function $D(R)$ is the infimum of all distortions D, such that there exists a sequence (f_n, g_n) with codeword size 2^{nR}, that $\lim_{n\to\infty} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D$. It is shown that

$$D(R) = \inf_{P_{\hat{X}}^{X}} \mathbb{E}[d(X,\hat{X})] \ \text{ s.t. } \ I(X;\hat{X}) \leq R.$$

We denote this optimization as the rate-distortion minimization, and the function $D(R)$ as the the rate-distortion function (some authors call this the distortion-rate function to contrast with another function $R(D)$ which maps distortion to rate). For an overview of the classic rate distortion theory, see Chapter 10 of (Cover & Thomas, 2012).

Now, we are ready to precisely define the rate-distortion problem describing the MER. To do so, let's define the distortion function as the excess risk of $\hat{h}$ compared to the Bayes decision $h_w^*(x)$, i.e.

$$d(w,\hat{h}) = \mathbb{E}_{XY}^{w}[\ell(Y,\hat{h}(X)) - \ell(Y,h_w^*(X))]. \tag{6}$$

Note that this definition is consistent with our final goal which is to study MER: if we consider the optimal learning algorithm which generates $\hat{h}(.) = \psi^*_{Y|Z^nX}(z^n,.)$ for any given dataset $Z^n = z^n$, the expected distortion is

$$\begin{aligned}\mathbb{E}_{WZ^n}&[d(W,\psi^*_{Y|Z^nX}(Z^n,\cdot))]\\ &= \mathbb{E}_{WZ^nXY}[\ell(Y,\psi^*_{Y|Z^nX}(Z^n,X))\\ &\qquad\qquad - \ell(Y,\psi^*_{Y|WX}(W,X))]\\ &= R_\ell(Y|Z^n,X) - R_\ell(Y|W,X)\\ &= \mathrm{MER}_\ell^n.\end{aligned} \tag{7}$$

Now, we define the (constrained) rate-distortion optimization as

$$\begin{aligned} D_n(R) = &\inf_{P_{\hat{h}}^{Z^n}} \mathbb{E}[d(W,\hat{h})], \\ &\text{s.t. } I(W;\hat{h}) \leq R, \end{aligned} \tag{8}$$

in which the expectation and mutual information are evaluated with respect to $P_{W\hat{h}}$ which is the marginal distribution of $P_W \otimes P_{Z^n}^W \otimes P_{\hat{h}}^{Z^n}$. Note that in standard rate-distortion problems, we are allowed to directly optimize $P_{\hat{h}}^W$. However, here there is an extra constraint that $P_{\hat{h}}^W = P_{Z^n}^W \otimes P_{\hat{h}}^{Z^n}$. Also note the dependence of $D_n(R)$ on n: for each n there is a different rate-distortion optimization which yields D_n. Thus, we have a series of optimization problems. It is easy to verify that $D_n(R)$ is non-increasing in both n and R.

The following theorem states the relation between $D_n(R)$ and MER_ℓ^n.

Theorem 4. *For a given training set size n, for all rates $R \geq I(W;Z^n)$, we have*

$$D_n(R) = \mathrm{MER}_\ell^n.$$

Note that since a Markov chain $W - Z^n - \hat{h}$ holds, having $R \geq I(W;Z^n)$ actually removes the constraint on optimization problem (8). Thus, Eq. (7) can be used to prove this theorem.

We have seen that $I(W;Z^n)$ appeared in an upper bound on MER in Eq. (4). Combining this fact and Theorem 4, for a bounded loss we have

$$D_n(I(W;Z^n)) = \mathrm{MER}_\ell^n \leq \sqrt{\frac{b^2}{2n} I(W;Z^n)}. \tag{9}$$

5.3. Lower Bound

As discussed in Section 1.1, to have a standard rate-distortion problem, one can remove the constraint that $\hat{h}$ is generated only using the samples Z^n; i.e.

$$\begin{aligned} D^L(R) = &\inf_{P_{\hat{h}}^{W}} \mathbb{E}[d(W,\hat{h})], \\ &\text{s.t. } I(W;\hat{h}) \leq R. \end{aligned} \tag{10}$$

Note that since the feasible set is enlarged, the solution to this minimization will be a lower bound on the optimization of (8):

$$\forall R,\ \forall n;\ \ D^L(R) \leq D_n(R). \tag{11}$$

Function $D^L(R)$ is much easier to study than $D_n(R)$, since the corresponding optimization problem (10) is independent of n.

In the next sections, we will first derive an upper bound on $D_n(R)$. Then by studying the gap between the upper and lower bounds, we shed light on the behavior of $D_n(R)$.

5.4. Upper Bound

To define the upper bound, we add another constraint to the optimization problem (8): the mutual information between the dataset and the learned model $\hat{h}$ should also be constrained by R. More precisely, we define

$$D_n^U(R) = \inf_{P_{\hat{h}}^{Z^n}} \mathbb{E}[d(W, \hat{h})], \qquad (12)$$
$$\text{s.t. } I(Z^n; \hat{h}) \leq R.$$

Note that the constraint in (12) is more strict than the constraint in (8), since by data processing inequality we have

$$I(Z^n; \hat{h}) \leq R \Longrightarrow I(W; \hat{h}) \leq R.$$

Thus, we can write

$$\forall R,\ \forall n; D_n(R) \leq D_n^U(R). \qquad (13)$$

This rate-distortion problem is of interest by itself. Note that an increasingly popular approach in controlling the generalization gap in frequentist setting by information-theoretic tools, is to guarantee that mutual information between dataset and the model is small (Xu & Raginsky, 2017; Russo & Zou, 2015; 2016; Bassily et al., 2018). To translate the frequentist setting to the Bayesian setting of our discussion, consider the same form of parametric learning in which the unknown distribution is assumed to be described by the parameter w. But no distribution is assumed on the value of w, and an algorithm should work for any w, in a minimax fashion. Thus, by bounding the mutual information, we mean that for all w, $I^w(Z^n; \hat{h}) \leq R$ and we have $I(Z^n; \hat{h}|W) = \mathbb{E}_W[I^w(Z^n; \hat{h})] \leq R$. On the other hand since there is the Markov chain $W - Z^n - \hat{h}$, we have $I(Z^n; \hat{h}|W) \leq I(Z^n; \hat{h})$. Therefore, understanding the effect of the constraint $I(Z^n; \hat{h}) \leq R$ in the Bayesian setting could be illuminative also for the frequentist setting. In particular, if $I(Z^n; \hat{h}) \leq R$ is satisfied for all P_W, we have $I^w(Z^n; \hat{h}) \leq R; \forall w \in \mathcal{W}$.

A natural question that should be studied is whether the equality $D_n(I(W; Z^n)) \overset{?}{=} D_n^U(I(W; Z^n))$ holds. The informal rational behind this question is as follows: Intuitively, if we know that only $R = I(W; Z^n)$ nats of information about W is present in the dataset Z^n, it should be possible to just extract those nats without relying more on the dataset. Unfortunately, this equality does not hold in general. Actually, often an unbounded $I(\hat{h}; Z^n)$ is needed in order to achieve $D_n(R)$. To see this, consider the simple problem where $W \sim \mathcal{N}(0, 1)$, $Z^n = (Y_i)_{i=1}^n$, $Y_i \sim \mathcal{N}(W, 1)$, and $\ell(y, \hat{y}) = (y - \hat{y})^2$. In this case, it is easy to verify that there is a unique optimal Bayes decision rule $\psi^*_{Y|Z^n}(z^n)$ (expected value of the posterior distribution of W given Z^n), which is a deterministic function of Z^n. Thus, while $I(W; Z^n)$ is finite (as we know from well-known results on Gaussian channels), $I(Z^n; \hat{h})$ should be infinite to achieve the best performance.

Despite this unsatisfactory observation, it is actually possible to do quite well with a limited rate if we don't persist in using exactly the optimal decision rule. This is made precise in the next theorem.

Theorem 5. *For any bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, b]$, and for all $n \geq 1$, we have*

$$D_n^U(I(W; Z^n)) \leq \sqrt{\frac{b^2}{2} I(W; Y|Z^n, X)} \qquad (14)$$
$$\leq \sqrt{\frac{b^2}{2n} I(W; Z^n)}. \qquad (15)$$

5.5. Relation between Lower and Upper Bounds

The following theorem states the relation between the upper bound $D_n^U(R)$ and the lower bound $D_L(R)$.

Theorem 6. *For any bounded loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, b]$, we have*

$$D_n^U(R) \leq D^L(R) + \sqrt{\frac{b^2}{2} I(W; \hat{h}_R|Z^n)}, \qquad (16)$$

where the mutual information is based on the distribution $P_{W,\hat{h}_R Z^n} = P_W \otimes P^{*W}_{\hat{h}_R} \otimes P^W_{Z^n}$ and $P^{*W}_{\hat{h}_R}$ is a solution to the optimization of $D^L(R)$.

This theorem states that to understand the difference between $D^L(R)$ and $D_n^U(R)$, one can solve the optimization associated to $D^L(R)$ to find $P^{*W}_{\hat{h}_R}$. Then, the mutual information $I(W; \hat{h}_R|Z^n)$ controls the gap between the upper bound and lower bound. While this nonasymptotic bound provides an intuitive understanding of the interplay between $D^L(R)$ and $D_n^U(R)$ for all n and R, it is hard to be evaluated. But as $n \to \infty$, if the posterior is concentrated to the true realization, it is reasonable to expect that $I(W; \hat{h}_R|Z^n) \to 0$ and all of the rate-distortion functions converge. This is made precise in the next theorem.

Theorem 7. *Suppose the distortion $d(W, \hat{h})$ defined in Eq. (6) can be represented as a distance $d'(h^*_W, \hat{h})$. Let W and W' be two samples independently generated from $P_W^{Z^n}$. If we have*

$$\lim_{n \to \infty} \mathbb{E}[d'(h^*_W, h^*_{W'})] = 0,$$

then

$$\forall R \geq 0;\ D^L(R) = \lim_{n \to \infty} D_n(R) = \lim_{n \to \infty} D_n^U(R). \qquad (17)$$

Note that the condition $\lim_{n \to \infty} \mathbb{E}[d'(h^*_W, h^*_{W'})] = 0$ is usually satisfied as a result of the convergence of the posterior distribution. In Section 6, we will see cases for which the distortion can be represented as a distance.

In Figure 1, the relation between all the introduced rate-distortion functions is presented. This figure also summarizes some of the presented results.

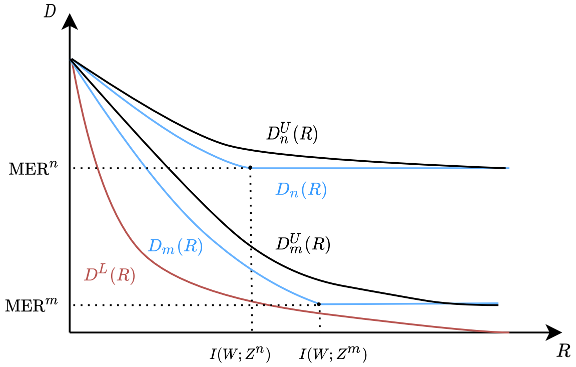

Figure 1. A schematic view of the relation between rate-distortion functions studied in this paper. The original rate-distortion function $D_n(R)$ and its upper bound $D_n^U(R)$ are presented for two sample sizes n and m, where $n < m$. The lower bound $D^L(R)$ is also illustrated. As discussed in Theorem 4, $D_n(R)$ is equal to MER^n for $R \geq I(W;Z^n)$. Also note that the upper bound approaches MER^n as $R \to \infty$. Both $D_n(R)$ and $D_n^U(R)$ approach the $D(R)$ as $n \to \infty$.

6. Applications

In the framework developed in previous sections, the distortion function $d(W,\hat{h})$ has a quite general form: it measure the distortion between a variable W and a function $\hat{h}$. In practice it is usually easier to represent the problem in a way that both input and output are elements of a shared metric space and the distance of that space is the distortion measure.

In order to achieve this, we first need a lemma which allows us to reformulate the rate-distortion functions.

Lemma 8. *(Reparameterization Lemma) Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$ be a distortion function. Assume that there exists mappings $f : \mathcal{X} \to \mathcal{V}$ and $g : \hat{\mathcal{X}} \to \hat{\mathcal{V}}$, and a distortion function $d' : \mathcal{V} \times \hat{\mathcal{V}} \to \mathbb{R}$, such that for all $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$, we have $d(x,\hat{x}) = d'(f(x), g(\hat{x}))$. If $\hat{\mathcal{V}} = f(\hat{\mathcal{X}})$, it follows that*

$$\begin{aligned} \min_{P^X_{\hat{X}}} \ \mathbb{E}_{X,\hat{X}}\left[d(X,\hat{X})\right] \ &= \ \min_{P^V_{\hat{V}}} \ \mathbb{E}_{V,\hat{V}}\left[d'(V,\hat{V})\right], \\ \text{s.t.} \ \ I(X;\hat{X}) \leq R \qquad &\qquad \text{s.t.} \ \ I(V;\hat{V}) \leq R \end{aligned}$$

where the second minimization is the rate-distortion function for random variable $V = f(X)$.

When the reparameterization is applicable, we might abuse the notation and write $d(V,\hat{V})$ instead of $d'(V,\hat{V})$.

For quadratic loss, the reparameterization lemma can be used to represent $d(W,\hat{h})$ as a norm on a suitable function space. Let $\mathcal{Y} \subseteq \mathbb{R}$ and consider $l(y,\hat{y}) = (y-\hat{y})^2$, for all $y, \hat{y}, \in \mathcal{Y}$. Based on Equation (6), we have

$$\begin{aligned} d(w,\hat{h}) &= \mathbb{E}^w_{XY}\left[\left|Y - h^*_w(X)\right|^2 - \left|Y - \hat{h}(X)\right|^2\right] \\ &= \mathbb{E}^w_X\left[\left|h^*_w(X) - \hat{h}(X)\right|^2\right], \end{aligned}$$

which is the norm of $L^2(P_X)$. Thus using the reparameterization lemma, the rate distortion problems can be restated for the distortion function $d(h^*_w,\hat{h}) = ||h^*_w - \hat{h}||_{L^2(P_X)}$.

It would be helpful if we could represent the distortion function by a distance on the parameter space, but this is not always possible. To be precise, define the hypothesis class

$$\mathcal{H} = \{h_w(.) = \psi^*_{Y|WX}(w,\cdot)|w \in \mathcal{W}\} \qquad (18)$$

where $\mathcal{W}$ is the set of all possible Ws. Note that the optimal function learned from the samples z^n, $\psi^*_{Y|Z^nX}(z^n,.)$, does not necessarily lie in $\mathcal{H}$. In other words it might not be parameterizable using W. But it might still be possible to derive lower bounds by projecting on the set $\mathcal{H}$.

Assume that $\mathcal{H}$ is a convex subset of the Hilbert space $L^2(P_X)$. For a given $f \in L^2(P_X)$, define $\mathrm{proj}_{\mathcal{H}}(f)$ as the projection of f on the convex set $\mathcal{H} \subseteq L^2(P_X)$. As a result of the contraction property of projections on convex sets in Hilbert spaces, we have $d(\mathrm{proj}_{\mathcal{H}}(\hat{h}), h^*_w) \leq d(\hat{h}, h^*_w)$. Therefore,

$$\begin{aligned} D_L(R) \geq \ &\min_{P^W_{\hat{h}}} \ \mathbb{E}\left[d\left(h^*_W, \mathrm{proj}_{\mathcal{H}}(\hat{h})\right)\right], \\ &\text{s.t.} \ \ I\left(h^*_W; \mathrm{proj}_{\mathcal{H}}(\hat{h})\right) \leq R. \end{aligned}$$

By the application of Lemma 8, we arrive at the following lower bound

$$\begin{aligned} D_L(R) \geq \ &\min_{P^W_{\hat{W}}} \ \mathbb{E}\left[d'\left(W,\hat{W}\right)\right]. \\ &\text{s.t.} \ \ I\left(W;\hat{W}\right) \leq R. \end{aligned}$$

in which $d'(w,\hat{w}) = d(h^*_w, h_{\hat{w}})$ where $h_{\hat{w}} = \mathrm{proj}_{\mathcal{H}}(\hat{h})$. This process of projection and reparameterization is summarized in Fig. 2.

Based on Theorem 4 and Equation (11), we know that $\mathrm{MER}^n_\ell \geq D^L(I(W;Z^n))$. Under the conditions that the space $\mathcal{W}$ is finite-dimensional, and that some regularity conditions on P^w_Z hold (see Lemma 2), we can use this fact and the following lemma to lower bound MER.

Lemma 9. *(Shannon Lower Bound (Shannon, 1959)) Let W and $\hat{W}$ be random variables taking values in $\mathbb{R}^p$, $||.||$ be an arbitrary norm on $\mathbb{R}^p$, and r be a positive real number.*

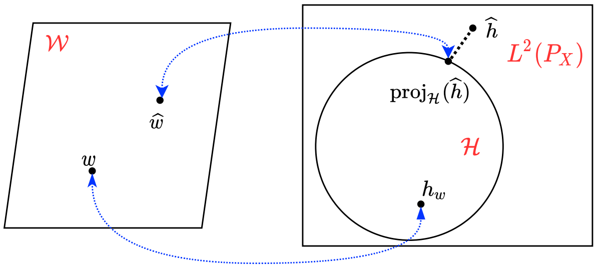

Figure 2. Projection to parametric family of functions. In the left, we have the set of all parameters $\mathcal{W}$ and in the right, the set of all functions in $L^2(P_X)$. $\mathcal{H}$ is the set of functions which are associated to a parameter in $\mathcal{W}$. For a function $\hat{h}$ which is not in $\mathcal{H}$, first it is projected to $\mathcal{H}$ using the $L^2(P_X)$ geometry, and the parameter $\hat{w}$ for $\text{proj}_{\mathcal{H}}(\hat{h})$ is used. Then the reparameterization lemma can be used to restate this refined problem in $\mathcal{W}$.

For any $D \geq 0$, define the rate-distortion function

$$R(D) = \inf_{P^W_{\hat{W}}} I(W; \hat{W}),$$
$$\text{s.t. } \mathbb{E}_{W,\hat{W}} ||W - \hat{W}||^r \leq D.$$

We have

$$R(D) \geq h(W) - \log\left(V_p \left(\frac{Dre}{p}\right)^{\frac{p}{r}} \Gamma(1 + \frac{p}{r})\right),$$

in which $h(W)$ is the differential entropy of W and V_p is the volume of $\{x \in \mathbb{R}^p : ||x|| \leq 1\}$.

It is known that Shannon Lower Bound is asymptotically tight as $D \to 0$ (Wu, 2020; Koch, 2016).

Under the conditions of Lemma 2, we can prove the following lower bound for MER:

Theorem 10. *Let $\mathcal{W}$ be a p-dimensional compact subspace of $\mathbb{R}^p$, the hypothesis set defined in Eq.* (18) *be convex, and assume that the regularity conditions of Lemma 2 hold. If there exists a norm $||.||$ such that $||W - \hat{W}||^2 \leq d'(W, \hat{W})$, we have*

$$\text{MER}^n_\ell \geq \frac{p}{n} \cdot \frac{\pi}{\left(V_p\, \Gamma(1 + \frac{p}{2})\right)^{\frac{2}{p}}} \exp\left(-\frac{\mathbb{E}\log|J^W_Z(W)|}{p}\right),$$

as $n \to \infty$, where V_p is the volume of $\{x \in \mathbb{R}^p : ||x|| \leq 1\}$.

This theorem formalizes the intuition that a lower MER might be achieved in problems in which the expected Fisher information of W contained in Z is large. Note that this theorem implies $\text{MER}^n_\ell = \Omega(1/n)$.

If there exists a constant c such that for every w, $\left(J^W_Z(w)\right)_{ii} \leq c$, we can write

$$\mathbb{E}_W \log|J^W_Z(W)| \leq \mathbb{E}_W \log\prod_{i=1}^p \left(J^W_Z(W)\right)_{ii} \leq p\log(c),$$

where the first inequality follows from Hadamard's inequality. Also note that given a positive-definite matrix A, the volume of the ellipsoid $\{x \in \mathbb{R}^p : ||x||_A \leq 1\}$ is given by $V_p = (\det A)^{-\frac{1}{2}} \frac{\pi^{p/2}}{\Gamma(1+\frac{p}{2})}$. Using these facts, the lower bound of $\Omega(p/n)$ can be obtained for MER, as stated in the following corollary.

Corollary 10.1. *Under the conditions of Theorem 10, if $||\cdot|| = ||\cdot||_A$ for some positive-definite matrix A, and by assuming that for all w; $(J^W_Z(w))_{ii} \leq c$, we have*

$$\text{MER}^n_\ell \geq \frac{\gamma p}{nc} = \Omega\left(\frac{p}{n}\right),$$

as $n \to \infty$, in which γ is the smallest eigenvalue of A.

6.1. Gaussian Location Model

Let $Y_i = W + V_i$; $\forall i \leq n+1$, where $V_i \sim \mathcal{N}(0, \sigma^2)$ and a prior $W \sim \text{Unif}(0, 1)$ be assumed on W. Given $\{Y_i\}^n_{i=1}$, the goal is to predict Y_{n+1}. In this problem, we have $J^W_Z(w) = J^W_Y(w) = \frac{1}{\sigma^2}$. Let $\ell(a, b) = (a - b)^2$ which implies $d(w, \hat{h}) = (w - \hat{h})^2$. The distance $d(w, \hat{h})$ is a norm in the space $\mathbb{R}$, and it satisfies the conditions of Theorem 10. Thus, $\text{MER}^n_2 = \Omega(\frac{1}{n})$.

6.2. Linear Regression

Let $Y = W^\top X + \sigma\nu$, where $W \sim P_W$, $X \sim \mathcal{N}(0, \Sigma_X)$, and $\nu \sim \mathcal{N}(0, 1)$, in which P_W is supported on a compact subspace $\mathcal{W}$ of $\mathbb{R}^p$. Also assume that W, X, and ν are independent, the matrix Σ_X is full-rank, and the space $\mathcal{W}$ is convex. Consider $\ell(a, b) = (a - b)^2$. Note that $h^*_w(x) = w^\top x$ and that the hypothesis class $\mathcal{H} = \{h_w(x) = w^\top x | w \in \mathcal{W}\}$ is convex. We have

$$\begin{aligned} d'(w, \hat{w}) &= d(h_w, h_{\hat{w}}) \\ &= \mathbb{E}_X[(w^\top X - \hat{w}^\top X)^2] \\ &= (w - \hat{w})^\top \Sigma_X (w - \hat{w}), \end{aligned}$$

meaning that the distance $d'(w, \hat{w})$ can be formulated by a norm in the space $\mathbb{R}^p$, i.e. $d'(w, \hat{w}) = ||w - \hat{w}||^2_{\Sigma_X}$. This problem satisfies the assumptions of Corollary 10.1, and we have $\text{MER}^n_\ell = \Omega(\frac{p}{n})$. A similar rate-distortion problem has been studied in the context of compression of linear models in (Gao et al., 2019).

Note that in this case the loss is unbounded. While in Lemma 1 upper bounds for bounded loss are provided, similar results for the general case of unbounded loss can be derived if the tails of the distribution are suitably controlled; e.g. the distribution is subgaussian (see Theorem 4 of (Xu & Raginsky, 2020)). Moreover, for the case of quadratic loss, if the Gaussian noise ν is replaced by a bounded random variable, the loss would be bounded and there exists upper bounds with the same rate of $O(\frac{p}{n})$ (see Theorem 3 of (Xu & Raginsky, 2020)).

6.3. Certain Classes of Non-Linearities

Fix $w_0 \in \mathbb{R}^p$ and a function $\Phi_{w_0}(\cdot) : \mathcal{X} \to \mathbb{R}^p$. Consider the set of nonlinear functions

$$f(\cdot, w) = f(\cdot, w_0) + \Phi_{w_0}^\top(\cdot)(w - w_0),$$

for $w \in \mathcal{W} \subseteq \mathbb{R}^p$. The class resembles Neural Tangent Kernels (Jacot et al., 2018). Assume that $Y = f(X, W) + \sigma\nu$, where $X \sim P_X$, $W \sim P_W$ in which P_W is supported on a compact subset $\mathcal{W}$ of $\mathbb{R}^p$, and $\nu \sim \mathcal{N}(0, 1)$. Also assume that $\mathcal{W}$ is convex. Consider the loss function $\ell(a, b) = (a - b)^2$. We have $h_w^*(x) = f(x, w)$ and the hypothesis set $\mathcal{H} = \{f(\cdot, w) \,|\, w \in \mathcal{W}\}$ is convex. If $\mathbb{E}[\Phi_{w_0}(X)\Phi_{w_0}^\top(X)]$ is full-rank, and Φ_{w_0} is smooth such that the smoothness conditions of Lemma 2 hold, then following the same line of reasoning as the linear regression example, we have $\mathrm{MER}_\ell^n = \Omega(\frac{p}{n})$.

6.4. Usability for More General Cases

There are various aspects in Theorem 10 which one should take care of when dealing with more complicated problems. For example, while the previous section provided analysis for a very simplified neural network, there are some difficulties to apply such analysis for a more general (Bayesian) neural network.

One difficulty is the apparent dependence of the bound $\Omega(p/n)$ on p in the over-parameterized regime where we could have $p \gg n$. In particular, in many high dimensional problems, there are just a few dimensions for which the covariance matrix has large eigenvalues; i.e. data mostly resides in a lower dimensional space. In such scenarios, a more precise treatment is needed. To see that the bound does not depend on p, note that the constant hidden in the rate actually depends on the determinant of the covariance matrix, and having small eigenvalues potentially allows one to achieve a smaller MER.

Moreover, if the Fisher matrix is singular, better (non-singular) parameterization of the problem exists and the Reparameterization Lemma (Lemma 8) can be used to take advantage of this fact and then apply the lower bounds. The same technique might work for some cases where the mapping is not injective (a requirement which was enforced by the conditions of Lemma 2). For example, in neural networks, one source of complexity is that permuting the order of neurons and their corresponding weights in a hidden layer of a fully connected NN does not change the function. It might be possible to define a standard ordering of neurons to tackle this problem, though it might be challenging as other conditions should also be met simultaneously.

7. Conclusion and Future Work

In this paper, the recent framework of (Xu & Raginsky, 2020) for studying MER was studied and a source coding view on MER was suggested. In this view, the variable W is considered as the input, and the generated hypothesis $\hat{h}$ as the output. A suitable distortion measure $d(w, \hat{h})$ was defined to capture the notion of excess risk. This view was used to find fundamental limits on learning with limited amount of information. Since in learning from dataset Z^n, the information is inherently bounded by $I(W; Z^n)$, this view provides a natural methodology to study the limits of learning. Using this view, a rate-distortion function $D_n(R)$ was introduced and it was proved that it is equal to MER for large enough R. Then it was demonstrated how $D_n(R)$ is bounded bellow and above by two other rate-distortion functions $D^L(R)$ and $D_n^U(R)$ respectively, which were generated by two natural modifications of the original optimization. The lower bound indicated the limits on the ability of any process generating a hypothesis $\hat{h}$ from W while having a limited rate (not restricted to use a training set). The upper bound indicated the price one should pay if a bound on the $I(Z^n; \hat{h})$ is also enforced, a setting related to model compression. These three rate-distortion functions where studied and various upper and lower bounds on them were derived. In particular, it was demonstrated that (under certain conditions) the lower bound has the right rate matching the upper bound, proving that all of the bounds are order-wise tight, and the rate for MER_ℓ^n is $\Theta(p/n)$. Finally some applications of these results were discussed.

Some problems remained open for future studies. One of the limitations of the current work, is that Theorem 10 requires some technical conditions for the $\Omega(p/n)$ to be guaranteed. Analyzing lower rates for more general classes of problems remains an open problem. In particular, it is interesting to study MER lower bounds for non-parametric problems. The challenge in this setting is that the underlying results which were used to derive lower bound require a finite dimensional parameter space. Another interesting direction for future studies is to find conditions which guarantee $O(1/n)$ upper bounds for general bounded (or subgaussian) losses. While such rates are well studied from the frequentist standpoint (minimax setting), they are less understood in the Bayesian learning.

References

Asadi, A. R., Abbe, E., and Verdú, S. Chaining Mutual Information and Tightening Generalization Bounds. *arXiv preprint arXiv:1806.03803*, 2018.

Assouad, P. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Bassily, R., Moran, S., Nachum, I., Shafer, J., and Yehudayoff, A. Learners that Use Little Information. In *Algorithmic Learning Theory*, pp. 25–55, 2018.

Bu, Y., Gao, W., Zou, S., and Veeravalli, V. Information-theoretic understanding of population risk improvement with model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3300–3307, 2020.

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

Clarke, B. S. and Barron, A. R. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012.

Gao, W., Liu, Y.-H., Wang, C., and Oh, S. Rate distortion for model compression: From theory to practice. In *International Conference on Machine Learning*, pp. 2102–2111. PMLR, 2019.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. Convergence rates of posterior distributions. *Annals of Statistics*, pp. 500–531, 2000.

Ghosal, S., Van Der Vaart, A., et al. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007.

Hafez-Kolahi, H., Golgooni, Z., Kasaei, S., and Soleymani, M. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Keener, R. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer New York, 2010.

Koch, T. The shannon lower bound is asymptotically tight. *IEEE Transactions on Information Theory*, 62(11):6155–6161, 2016.

Le Cam, L. and Yang, G. L. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2012.

LeCam, L. et al. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Merhav, N. and Feder, M. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

Nokleby, M., Beirami, A., and Calderbank, R. Rate-distortion bounds on bayes risk in supervised learning. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 2099–2103. IEEE, 2016.

Russo, D. and Zou, J. How much does your data exploration overfit? Controlling bias via information usage. *arXiv preprint arXiv:1511.05219*, 2015.

Russo, D. and Zou, J. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240, 2016.

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

Shen, X., Wasserman, L., et al. Rates of convergence of posterior distributions. *Annals of Statistics*, 29(3):687–714, 2001.

Steinke, T. and Zakynthinou, L. Reasoning About Generalization via Conditional Mutual Information. *arXiv preprint arXiv:2001.09122*, 2020.

Wu, Y. Lecture notes on information-theoretic methods for high-dimensional statistics. `http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf`, 2020.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2521–2530, 2017.

Xu, A. and Raginsky, M. Minimum excess risk in bayesian learning. *arXiv preprint arXiv:2012.14868*, 2020.

Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pp. 1564–1599, 1999.