Information Theoretical Upper and Lower Bounds for Statistical Estimation

Tong Zhang, Yahoo Inc., New York City, USA

***Abstract*— We establish upper and lower bounds for some statistical estimation problems through concise information theoretical arguments. Our upper bound analysis is based on a simple yet general inequality which we call the information exponential inequality. We show that this inequality naturally leads to a general randomized estimation method, for which performance upper bounds can be obtained. The lower bounds, applicable for all statistical estimators, are obtained by original applications of some well known information theoretical inequalities, and approximately match the obtained upper bounds for various important problems. Moreover, our framework can be regarded as a natural generalization of the standard minimax framework, in that we allow the performance of the estimator to vary for different possible underlying distributions according to a pre-defined prior.**

***Index Terms*— Gibbs algorithm, lower bound, minimax, PAC-Bayes, randomized estimation, statistical estimation**

I. Introduction

The purpose of this paper is to develop upper and lower bounds for some prediction problems in statistical learning. The upper bound analysis is based on a simple yet very general inequality (information exponential inequality), which can be applied to statistical learning problems. Our lower bounds are obtained from some novel applications of well-known information theoretical inequalities (specifically, data-processing theorems). We show that the upper bounds and lower bounds have very similar forms, and approximately match under various conditions.

We shall first present a relatively abstract framework, in which we develop the information exponential inequality our upper bound analysis is based on. The abstraction is motivated from statistical prediction problems which this paper investigates. In statistical prediction, we have input space $\mathcal{X}$, output space $\mathcal{Y}$, and a space of predictors $\mathcal{G}$. For any $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, and any predictor $\theta \in \mathcal{G}$, we incur a loss $L_\theta(X,Y) = L_\theta(Z)$, where $Z = (X,Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Consider a probability measure D on $\mathcal{Z}$. Our goal is to find a parameter $\hat{\theta}(\hat{Z})$ from a random sample $\hat{Z}$ from D, such that the loss $\mathbf{E}_Z\, L_{\hat{\theta}(\hat{Z})}(Z)$ is small, where $\mathbf{E}_Z$ is the expectation with respect to D (and Z is independent of $\hat{Z}$).

In the standard learning theory, we consider n random samples instead of one sample. Although this appears at first to be more general, it can be handled using the one-sample formulation if our loss function is additive over the samples. To see this, consider $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_n\}$. Let $L_\theta(Z) = \sum_{i=1}^n L_{i,\theta}(X_i, Y_i)$. If $Z_i = (X_i, Y_i)$ are independent random variables, then it follows that $\mathbf{E}_Z L_\theta(Z) = \sum_{i=1}^n \mathbf{E}_{Z_i} L_{i,\theta}(X_1, Y_1)$. Therefore any result on the one-sample formulation immediately implies a result on the n-sample formulation. We shall thus focus on the one-sample case without loss of generality.

In this paper, we consider both deterministic and randomized estimators. Randomized estimators are defined with respect to a prior π on $\mathcal{G}$, which is a probability measure on $\mathcal{G}$ with the property that $\int_{\mathcal{G}} d\pi(\theta) = 1$. For a randomized estimation method, given sample $\hat{Z}$ from D, we select θ from $\mathcal{G}$ based on a sample-dependent probability measure $d\hat{\pi}_{\hat{Z}}(\theta)$ on $\mathcal{G}$, In this paper, we shall call such a sample-dependent probability measure as a *posterior randomization measure* (or simplified as posterior). The word posterior in this paper is not necessarily the Bayesian posterior distribution in the traditional sense. Whenever it appears, we shall always refer the latter as Bayesian posterior distribution to avoid confusion. For notational simplicity, we also use the symbol $\hat{\pi}$ to denote $\hat{\pi}_{\hat{Z}}$. The randomized estimator associated with a posterior randomization measure is thus completely determined by its posterior $\hat{\pi}$. Its *posterior averaging risk* is the averaged risk of the randomized estimator drawn from this posterior randomization measure, which can be defined as (assuming that the order of integration can be exchanged)

$$\mathbf{E}_{\theta\sim\hat{\pi}}\mathbf{E}_Z L_\theta(Z) = \mathbf{E}_Z \int L_\theta(Z) d\hat{\pi}(\theta).$$

In this paper, we are interested in estimating this average risk for an arbitrary posterior $\hat{\pi}$. The statistical complexity of this randomized estimator $\hat{\pi}$ will be measured by its *KL-complexity* respect to the prior, which is defined as:

$$D_{KL}(\hat{\pi}||\pi) = \int_{\mathcal{G}} \ln \frac{d\hat{\pi}(\theta)}{d\pi} d\hat{\pi}(\theta), \tag{1}$$

assuming it exists.

It should be pointed out that randomized estimators are often suboptimal. For example, in prediction problems, the mean of the posterior can be a better estimator (at least for convex loss functions). We focus on randomized estimation because the main techniques presented in this paper only apply to randomized estimation methods due to the complexity measure used in (1). The prior π in our framework reflects the statistician's conjecture about where the optimal parameter is more likely to happen. Our upper bound analysis reflects this belief: the bounds are good if the statistician puts a relatively large prior mass around the optimal parameter, and the bounds are poor if the statistician makes a wrong bet by giving a very small prior mass around the optimal parameter. In particular, the prior does not necessarily correspond to nature's prior in the Bayesian framework, in that nature does not have to pick the target parameter according to this prior.

With the above background in mind, we can now start to develop our main machinery, as well as its consequence in statistical learning theory. The remainder of this paper is organized as follows. Section II introduces the basic information exponential inequality and its direct consequence. In Section III, we present various generalizations of the basic inequality, and discuss their implications. Section IV applies our analysis to statistical estimation, and derive corresponding upper bounds. Examples for some important but more specific problems will be given in Section V. Lower bounds that have the same form of the upper bounds in Section IV will be developed in Section VI, implying that our analysis gives tight convergence rates. Moreover, in order to avoid clutter in the main text, we put all but very short proofs into Section VII. Implications of our analysis will be discussed in Section VIII.

Throughout the paper, we ignore measurability issues. Moreover, for any integration with respect to multiple variables, we assume that Fubini's theorem is applicable, so that the order of integration can be exchanged.

II. Information Exponential Inequality

The main technical result which forms the basis of our upper bound analysis is given by the following lemma, where we assume that $\hat{\pi} = \hat{\pi}_{\hat{Z}}$ is a posterior randomization measure on $\mathcal{G}$ that depends on the sample $\hat{Z}$. The lemma in this specific form appeared first in an earlier conference paper [1]. For completeness, the proof is included in Section VII-A. There were also a number of studies that employed similar ideas. For example, the early work of McAllester on Pac-Bayes learning [2] investigated the same framework, although his results were less general. Independent results that are quite similar to what we obtain here can also be found in [3]–[5].

Lemma 2.1 (Information Exponential Inequality): Consider a measurable real-valued function $L_\theta(Z)$ on $\mathcal{G} \times \mathcal{Z}$. Then for an arbitrary sample $\hat{Z}$-dependent randomized estimation method, with posterior $\hat{\pi}$, the following inequality holds

$$\mathbf{E}_{\hat{Z}} \exp \left[\mathbf{E}_{\theta\sim\hat{\pi}} \Delta_{\hat{Z}}(\theta) - D_{KL}(\hat{\pi}||\pi)\right] \leq 1.$$

where $\Delta_{\hat{Z}}(\theta) = -L_\theta(\hat{Z}) - \ln \mathbf{E}_Z \exp(-L_\theta(Z))$.

The importance of this bound is that the left hand side is a quantity that involves an arbitrary randomized estimator, and the right hand side is a numerical constant independent of the estimator. Therefore the inequality is a result that can be applied to an arbitrary randomized estimator. The remaining issue is merely how to interpret this bound.

The following theorem gives a sample-dependent generalization bound of an arbitrary randomized estimator which is easier to interpret then Lemma 2.1. The proof, which uses Markov's and Jensen's inequality, is left to Section VII-A.

Theorem 2.1 (Information Posterior Bound): Consider randomized estimation, where we select posterior $\hat{\pi}$ on $\mathcal{G}$ based on $\hat{Z}$, with π a prior. Consider a real-valued function $L_\theta(Z)$ on $\mathcal{G} \times \mathcal{Z}$. Then $\forall t$, the following event holds with probability at least $1 - \exp(-t)$:

$$-\mathbf{E}_{\theta\sim\hat{\pi}} \ln \mathbf{E}_Z\, e^{-L_\theta(Z)} \leq \mathbf{E}_{\theta\sim\hat{\pi}} L_\theta(\hat{Z}) + D_{KL}(\hat{\pi}||\pi) + t.$$

Moreover, we have the following expected risk bound:

$$\begin{aligned} &-\mathbf{E}_{\hat{Z}}\mathbf{E}_{\theta\sim\hat{\pi}} \ln \mathbf{E}_Z\, e^{-L_\theta(Z)} \\ \leq &\mathbf{E}_{\hat{Z}} \left[\mathbf{E}_{\theta\sim\hat{\pi}} L_\theta(\hat{Z}) + D_{KL}(\hat{\pi}||\pi)\right]. \end{aligned}$$

Theorem 2.1 applies to unbounded loss functions. The right hand side of the theorem (see the first inequality) is the empirical loss plus complexity. The left-hand-side quantity is expressed as the minus logarithm of the true expected exponential of the negative loss. Using Jensen's inequality, we can see that it is smaller than the true expected loss. Bounds in terms of the true expected loss can also be obtained (see Section IV and Section V).

In general, there are two ways to use Theorem 2.1. Given a fixed prior π, one can obtain an estimator by minimizing the right hand side of the first inequality which is observation-dependent (and the minimizer is independent of t). The left-hand side (generalization error) is bounded according to the theorem. This approach, which we may call *Information Risk Minimization* (IRM), is the focus of Section IV and Section V. Another possible application of Theorem 2.1 is to estimate the generalization performance of any predefined randomized (and some deterministic) estimator by choosing a prior π which optimizes the bound.

III. Extensions of Information Exponential Inequality

The basic information exponential inequality and information posterior bound can be extended in various ways. Although seemingly more general at first, these extensions are essentially equivalent to the basic formulations.

A. Randomized model section

One important problem we may encounter in practice is the issue of prior selection, which corresponds to model-selection in deterministic statistical modeling. The purpose of this discussion is to illustrate that model selection can be naturally integrated into our framework without any additional analysis. Because of this, the main results of the paper can be applied to model selection, although we do not explicitly state them in such context. Therefore although this section is not needed for the technical development of the paper, we still include it here for completeness.

Mathematically, we may consider countably many models on $\mathcal{G}_k \subset \mathcal{G}$, which are indexed by $k \in \mathcal{M} = \{1, 2, \ldots\}$ (the countability assumption is more of a notational convenience than a requirement). We shall call $\mathcal{M}$ the *model space*. Each model $k \in \mathcal{M}$ is associated with a prior π_k on $\mathcal{G}_k \subset \mathcal{G}$, together with a belief $\mu_k \geq 0$ such that $\sum_{k=1}^{\infty} \mu_k = 1$. The probability measure $\mu = \{\mu_k\}$ can be regarded as a prior belief put on the models.

Now we may consider a randomized model-selection scheme where we randomly select model k based on a data-dependent probability distribution $\hat{\mu} = \{\hat{\mu}_k\} = \{\hat{\mu}^k_{\hat{Z}}\}$ on the model space $\mathcal{M}$, and then a random estimator $\hat{\pi}_k = \hat{\pi}^k_{\hat{Z}}$ with respect to model k. In this framework, the resulting randomized estimator is equivalent to a posterior randomization measure of the following form

$$\sum_{k=1}^{\infty} \hat{\mu}_k d\hat{\pi}_k \tag{2}$$

on $\mathcal{G}$. The posterior $\hat{\pi}^k$ is a randomization measure with respect to model k, with its complexity measured by $D_{KL}(d\hat{\pi}_k || d\pi_k)$. Letting $\{f^k(\theta)\}$ be a sequence of functions indexed by $k \in \mathcal{M}$, we are interested in the quantity:

$$\sum_{k=1}^{\infty} \hat{\mu}_k \mathbf{E}_{\hat{\pi}_k} f^k(\theta).$$

Note that for notational simplicity, throughout this section, we shall drop the $\hat{Z}$ dependence in the randomized estimators $\hat{\mu}_k$ and $\hat{\pi}_k$.

It can be seen immediately that this framework can be mapped into the framework without model-selection simply by changing the parameter space from $\mathcal{G}$ to $\mathcal{G} \times \mathcal{M}$. The prior sequence π_k becomes a prior π' on $\mathcal{G} \times \mathcal{M}$ defined as $d\pi'(\theta, k) = \mu_k d\pi_k(\theta)$. This implies that mathematically, from the randomized estimation point of view, the model-selection framework considered above is not truly necessary. This is consistent with Bayesian statistics where prior selection (would be the counterpart of model selection in non-Bayesian statistical inferencing) is not necessary since one can always introduce hidden parameters and use a prior mixture (with respect to the hidden variable). This is similar to what happens here. We believe that the ability of our framework to treat model selection and estimation under the same setting is very attractive and conceptually important. It also simplifies our theoretical analysis.

Based on this discussion, the statistical complexity of the randomized model-selection method in (2) is measured by the KL-complexity of the resulting randomized estimator on the $\mathcal{G} \times \mathcal{M}$ space, which can now be conveniently expressed as:

$$\begin{aligned} & D_{KL}(\hat{\mu}, \hat{\pi} || \mu, \pi) \\ = & \sum_{k=1}^{n} \hat{\mu}_k D_{KL}(d\hat{\pi}_k || d\pi_k) + D_{KL}(\hat{\mu} || \mu), \end{aligned} \tag{3}$$

where

$$D_{KL}(\hat{\mu} || \mu) = \sum_{k=1}^{\infty} \hat{\mu}^k \ln \frac{\hat{\mu}^k}{\mu_k}.$$

It follows that any KL-complexity based bound for randomized estimation without model selection also holds for randomized estimation with model selection. We simply replace the KL-complexity definition in (1) by the complexity measure (3). The case of information posterior bound is listed below.

Corollary 3.1 (Model Selection Bound): Consider a sequence of measurable real-valued functions $L^k_\theta(Z)$ on $\mathcal{G} \times \mathcal{Z}$ indexed by k. Then for an arbitrary $\hat{Z}$-dependent randomized estimation method (with randomized model-selection) of the form (2), and for all t, the following event holds with probability at least $1-\exp(-t)$:

$$-\sum_{k=1}^{\infty}\hat{\mu}_k\mathbf{E}_{\theta\sim\hat{\pi}_k}\ln\mathbf{E}_Z\, e^{-L_\theta^k(Z)}$$
$$\leq\sum_{k=1}^{\infty}\hat{\mu}_k\mathbf{E}_{\theta\sim\hat{\pi}_k}L_\theta^k(\hat{Z})+D_{KL}(\hat{\mu},\hat{\pi}||\mu,\pi)+t.$$

Moreover, we have the following expected risk bound:

$$-\mathbf{E}_{\hat{Z}}\sum_{k=1}^{\infty}\hat{\mu}_k\mathbf{E}_{\theta\sim\hat{\pi}_k}\ln\mathbf{E}_Z\, e^{-L_\theta^k(Z)}$$
$$\leq\mathbf{E}_{\hat{Z}}\left[\sum_{k=1}^{\infty}\hat{\mu}_k\mathbf{E}_{\theta\sim\hat{\pi}_k}L_\theta^k(\hat{Z})+D_{KL}(\hat{\mu},\hat{\pi}||\mu,\pi)\right].$$

Although not more general in principle, the introduction of the hidden component k maybe useful in various applications. For example, consider a decomposition of the posterior randomization measure of the form in (2) with $\hat{\mu}_k=\mu_k$. The corresponding prior decomposition is $d\pi=\sum_{k=1}^{\infty}\mu_k d\pi_k$. From (3), the statistical complexity of a posterior $d\hat{\pi}_{\hat{Z}}=\sum_{k=1}^{\infty}\mu_k d\hat{\pi}_k$ can be measured by $\sum_{k=1}^{\infty}\mu_k D_{KL}(d\hat{\pi}_k||d\pi_k)$. With an appropriate choice of prior decomposition, we may obtain a more refined complexity estimate of a posterior $d\hat{\pi}$, which is better than the complexity measure $D_{KL}(d\hat{\pi}||d\pi)$. In fact, such a decomposition is necessary to obtain non-trivial results for deterministic estimators under a continuous prior π, where $D_{KL}(d\hat{\pi}||d\pi)=\infty$ for a delta-function like posterior $d\hat{\pi}$.

B. Prior transformation and localization

Since we are free to choose any prior in the complexity measure (1), a bound using this quantity can be turned into another bound with another complexity measure under a transformed prior. In particular, the following transformation is useful. The result can be verified easily using direct calculation.

Proposition 3.1: Given a real-valued functions $r(\theta)$ on $\mathcal{G}$, and define

$$d\pi^{\rightarrow r}(\theta)=\frac{e^{r(\theta)}d\pi(\theta)}{\mathbf{E}_{\theta\sim\pi}e^{r(\theta)}},$$

then

$$D_{KL}(\hat{\pi}||\pi^{\rightarrow r})=D_{KL}(\hat{\pi}||\pi)-\mathbf{E}_{\theta\sim\hat{\pi}}r(\theta)+\ln\mathbf{E}_{\theta\sim\pi}e^{r(\theta)}.$$

In particular, we may choose $r(\theta)=\alpha\ln\mathbf{E}_Z\, e^{-L_\theta(Z)}$, which leads to the following corollary.

Corollary 3.2 (Localized KL-complexity): Use the notation of Theorem 2.1 and consider $\alpha\in[0,1]$. Let

$$c(\alpha)=\ln\mathbf{E}_{\theta\sim\pi}\mathbf{E}_Z^{\alpha}e^{-L_\theta(Z)},$$

then for all t, the following event holds with probability at least $1-\exp(-t)$:

$$-(1-\alpha)\mathbf{E}_{\theta\sim\hat{\pi}}\ln\mathbf{E}_Z\, e^{-L_\theta(Z)}$$
$$\leq\mathbf{E}_{\theta\sim\hat{\pi}}L_\theta(\hat{Z})+D_{KL}(\hat{\pi}||\pi)+c(\alpha)+t.$$

Moreover, we have the following expected risk bound:

$$-(1-\alpha)\mathbf{E}_{\hat{Z}}\mathbf{E}_{\theta\sim\hat{\pi}}\ln\mathbf{E}_Z\, e^{-L_\theta(Z)}$$
$$\leq\mathbf{E}_{\hat{Z}}\left[\mathbf{E}_{\theta\sim\hat{\pi}}L_\theta(\hat{Z})+D_{KL}(\hat{\pi}||\pi)\right]+c(\alpha).$$

Although the corollary is obtained merely through a change of prior in Theorem 2.1, it is important to understand that this prior change itself depends on the function to be estimated. In order to obtain a statistical estimation procedure (see Section IV), one wants to minimize the right hand side of Theorem 2.1. Hence the prior used in the estimation procedure (which is part of the estimator) cannot depend on the function to be estimated. Therefore the statistical consequences of Theorem 2.1 and Corollary 3.2, when applied to randomized estimation methods, are different.

As we shall see later, for some interesting estimation problems, the right hand side of Corollary 3.2 will be significantly better than that of Theorem 2.1. The $c(\alpha)$ term in Corollary 3.2 is quite useful for some parametric problems, where we would like to obtain a convergence rate of the order $O(1/n)$. In such cases, the choice of $\alpha=0$ (as in Theorem 2.1) would lead to a rate of $O(\ln n/n)$, which is suboptimal. This localization technique has also appeared in [3]–[6].

IV. Information Risk Minimization

Given any prior π, by minimizing the right hand side of Theorem 2.1, we obtain an estimation procedure, which we call *Information Risk Minimization* (IRM).

We shall consider the method for the case of n iid samples $\hat{Z}=(\hat{Z}_1,\ldots,\hat{Z}_n)\in\mathcal{Z}=\mathcal{Z}_1\times\cdots\times\mathcal{Z}_n$, where $\mathcal{Z}_1=\cdots=\mathcal{Z}_n$. The loss function is $L_\theta(\hat{Z})=\rho\sum_{i=1}^{n}\ell_\theta(\hat{Z}_i)$, where ℓ is a function on $\mathcal{G}\times\mathcal{Z}_1$ and $\rho>0$ is a constant. Given a subset S of probability distributions on $\mathcal{G}$, the IRM method finds a randomized estimator $\hat{\pi}_S$ which minimizes the empirical information complexity:

$$\hat{\pi}_{\rho,S}=\arg\inf_{\hat{\pi}\in S}\left[\mathbf{E}_{\theta\sim\hat{\pi}}\sum_{i=1}^{n}\ell_\theta(\hat{Z}_i)+\frac{D_{KL}(\hat{\pi}||\pi)}{\rho}\right]. \tag{4}$$

There are two cases of S that are most interesting. One is to take S as the set of all possible posterior randomization measures. In this case, the corresponding estimator (often referred to as the Gibbs algorithm) can be solved and becomes a randomized estimator which we simply denote as $\hat{\pi}_\rho$:

$$d\hat{\pi}_\rho = \frac{\exp(-\rho\sum_{i=1}^n \ell_\theta(\hat{Z}_i))}{\mathbf{E}_{\theta\sim\pi}\exp(-\rho\sum_{i=1}^n \ell_\theta(\hat{Z}_i))}\, d\pi. \tag{5}$$

In order to see that $d\hat{\pi}_\rho$ minimizes (4), we simply note that the right hand side of (4) can be rewritten as:

$$\left[\mathbf{E}_{\theta\sim\hat{\pi}}\sum_{i=1}^n \ell_\theta(\hat{Z}_i) + \frac{D_{KL}(\hat{\pi}||\pi)}{\rho}\right] = \frac{D_{KL}(\hat{\pi}||\hat{\pi}_\rho)}{\rho} + c,$$

where c is a quantity that is independent of $\hat{\pi}$. It follows that the optimal solution is achieved at $\hat{\pi} = \hat{\pi}_\rho$. The Gibbs method is closely related to the Bayesian method. Specifically, the Bayesian posterior distribution is given by the posterior randomization measure $\hat{\pi}_\rho$ of the Gibbs algorithm at $\rho = 1$ with the log-loss.

The other most interesting choice is when $\mathcal{G}$ is a discrete net which is consisted of countably many discrete parameters: $\mathcal{G} = \{\theta_1, \theta_2, \ldots\}$. The prior π can be represented as $\pi_k(k = 1, 2, \ldots)$. We consider the set S of distributions concentrated on a point $\hat{k}$: $\hat{\pi}_{\hat{k}} = 1$ and $\hat{\pi}_k = 0$ when $k \neq \hat{k}$. IRM in this case becomes a deterministic estimator which estimates a parameter $\hat{k}_\rho$ as follows:

$$\hat{k}_\rho = \arg\inf_k \left[\rho\sum_{i=1}^n \ell_{\theta_k}(\hat{Z}_i) + \ln\frac{1}{\pi_k}\right]. \tag{6}$$

We then use the deterministic parameter $\theta_{\hat{k}_\rho}$ as our estimator. This method minimizes a penalized empirical risk on a discrete net, which is closely related to the minimum description length method for density estimation. In general, estimators on a discrete net have been frequently used in theoretical studies, although they can be quite difficult to apply in practice. From a practical point of view, the Gibbs algorithm (5) with respect to a continuous prior might be easier to implement since one can use Monte Carlo techniques to draw samples from the Gibbs posterior. Moreover, in our analysis, the Gibbs algorithm has better generalization guarantee.

A. A general convergence bound

For the IRM method (4), we assume that ρ is a given constant. As is typical in statistical estimation, one may choose an optimal ρ by cross-validation. It is possible to allow data-dependent ρ in our analysis using the model selection idea discussed in Section III-A. In this case, we may define a discrete set of $\{\rho_k\}$, each with loss function $L_\theta^k(Z) = \rho_k L_\theta(Z)$. In our analysis, we can then pick a specific model $\hat{k}$ deterministically, which then leads to a data-dependent choice of $\rho_{\hat{k}}$.

However, for simplicity, in the following discussion, we focus on IRM with a pre-specified (possibly non-optimal) ρ. A more general extension of (4) is to allow θ-dependent scaling factor ρ, which is also possible to study (see related discussions in Section IV-B).

The following theorem gives a general bound for the IRM method. The proof can be found in Section VII-B.

Theorem 4.1: Define resolvability

$$r_{\rho,S} = \inf_{\pi'\in S}\left[\mathbf{E}_{\theta\sim\pi'}\mathbf{E}_{Z_1}\ell_\theta(Z_1) + \frac{1}{\rho n}D_{KL}(\pi'||\pi)\right].$$

Then $\forall\alpha \in [0, 1)$, the expected generalization performance of IRM method (4) can be bounded as

$$\begin{aligned}&-\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}}\ln\mathbf{E}_{Z_1}e^{-\rho\ell_\theta(Z_1)}\\ \leq&\frac{\rho}{1-\alpha}\left[r_{\rho,S} + \frac{1}{\rho n}\ln\mathbf{E}_{\theta\sim\pi}\mathbf{E}_{Z_1}^{\alpha n}e^{-\rho\ell_\theta(Z_1)}\right].\end{aligned}$$

Note for the Gibbs estimator (5), where S consists of all possible posterior randomization measures, the corresponding resolvability is given by

$$r_{\rho,S} = -\frac{1}{\rho n}\ln\mathbf{E}_{\theta\sim\pi}e^{-\rho n\mathbf{E}_{Z_1}\ell_\theta(Z_1)}. \tag{7}$$

For estimation on discrete net (6), the associated resolvability is

$$r_{\rho,S} = \arg\inf_k\left[\mathbf{E}_{Z_1}\ell_{\theta_k}(Z_1) + \frac{1}{\rho n}\ln\frac{1}{\pi_k}\right]. \tag{8}$$

B. Refined convergence bounds

In order to apply Theorem 4.1 to statistical estimation problems, we need to write the left hand side in a more informative form. For some specialized problems, the left hand side may have intuitive interpretations. For example, for density estimation with log-loss, the right hand side (of the second inequality) involves the KL-distance between the true density and the estimated density, and the left hand side can be interpreted as a different distance between the true density and the estimated density. Analysis following this line can be found in [6].

For general statistical estimation problems, it is useful to replace the left hand side by the true generalization error with respect to the randomized estimator, which is the quantity we are interested in. In order to achieve this, one can use a small value of ρ and bounds for the logarithmic moment generating functions that are given in Appendix I. Before going into technical details more rigorously, we first outline the basic idea (non-rigorously). Based on Proposition 1.1 and the remark thereafter, for small positive ρ, we have for each θ:

$$\begin{aligned}&-\ln \mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)}\\ &=\rho\mathbf{E}_{Z_1}\ell_\theta(Z_1) - \frac{\rho^2}{2}\mathbf{Var}_{Z_1}\ell_\theta(Z_1) + O(\rho^3).\end{aligned}$$

Now substituting into Theorem 4.1 with $\alpha = 0$, we obtain

$$\begin{aligned}&\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}}\mathbf{E}_{Z_1}\ell_\theta(Z_1)\\ \leq& r_{\rho,S} + \frac{\rho}{2}\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}}\mathbf{Var}_{Z_1}\ell_\theta(Z_1) + O(\rho^2).\end{aligned}$$

To further develop the idea (non-rigorously at the moment), we can ignore the higher order term in ρ, and obtain a bound for small ρ as long as we can replace $\mathbf{Var}_{Z_1}\ell_\theta(Z_1)$ by a bound of the form

$$\mathbf{Var}_{Z_1}\ell_\theta(Z_1) \leq K\mathbf{E}_{Z_1}\ell_\theta(Z_1) + v_\theta, \tag{9}$$

where $K > 0$ is a constant and v_θ are θ-dependent constants. Up to the leading order, this gives a bound of the form

$$\begin{aligned}&\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}}\mathbf{E}_{Z_1}\ell_\theta(Z_1)\\ \leq&\frac{1}{1-0.5K\rho}r_{\rho,S} + \frac{\rho}{2-K\rho}\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}}v_\theta + O(\rho^2).\end{aligned}$$

Interesting observations can be made from this bound. In order to optimize the bound on the right hand side, one may make a first-order adjustment to (4) by taking the variance into consideration:

$$\begin{aligned}\hat{\pi}_{\rho,S} = \arg\inf_{\hat{\pi}\in S}&\left[\mathbf{E}_{\theta\sim\hat{\pi}}\sum_{i=1}^{n}(\ell_\theta(\hat{Z}_i) + \frac{\rho}{2}v_\theta)\right.\\ &\left.+\frac{D_{KL}(\hat{\pi}||\pi)}{\rho}\right].\end{aligned}$$

Note that for small ρ, the effect of the v_θ adjustment is also relatively small. If v_θ cannot be reliably estimated, then one may also consider using the empirical variance of ℓ_θ as a plug-in estimator. Practically, the variance adjustment is quite reasonable. This is because if ℓ_θ has a large variance, then its empirical risk is a less reliable estimate of its true risk, and thus more likely to overfit the data. It is therefore useful to penalize the unreliability according to the estimated variances. A related method is to allow different scaling factor ρ for different θ, which can also be used to compensate different variance at different θ. This can be achieved through a redefinition of the prior weighted by θ dependent ρ and then normalize (also see [4]).

In the following examples, we pay special attention to the case of $v_\theta = 0$ in (9). The following theorem is a direct consequence of Theorem 2.1, where the technical idea discussed above is rigorously applied. See Section VII-B for its proof.

Theorem 4.2: Consider the IRM method in (4). Assume there exists a positive constant K such that $\forall\theta$:

$$\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \leq K\mathbf{E}_{Z_1}\ell_\theta(Z_1).$$

Assume either of the following conditions:

- Bounded loss: $\exists M \geq 0$ s.t. $-\inf_{\theta,Z_1}\ell_\theta(Z_1) \leq M$; let $\beta_\rho = 1 - K(e^{\rho M} - \rho M - 1)/(\rho M^2)$.
- Bernstein loss: $\exists M, b > 0$ s.t. $\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \leq m!M^{m-2}Kb\mathbf{E}_{Z_1}\ell_\theta(Z_1)$ for all $\theta \in \mathcal{G}$ and integer $m \geq 3$; let $\beta_\rho = 1 - K\rho(1-\rho M + 2b\rho M)/(2-2\rho M)$.

Assume we choose a sufficiently small ρ such that $\beta_\rho > 0$. Let the (true) expected loss of θ be $R(\theta) = \mathbf{E}_{Z_1}\ell_\theta(Z_1)$, then the expected generalization performance of the IRM method is bounded by

$$\begin{aligned}&\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}}R(\theta) \\ \leq&\frac{1}{(1-\alpha)\beta_\rho}\left[r_{\rho,S} + \frac{1}{\rho n}\ln\mathbf{E}_{\theta\sim\pi}e^{-\alpha\rho\beta_\rho nR(\theta)}\right],\end{aligned}\tag{10}$$

where $r_{\rho,S} = \inf_{\pi'\in S}\left[\mathbf{E}_{\theta\sim\pi'}R(\theta) + \frac{1}{\rho n}D_{KL}(\pi'||\pi)\right]$.

Remark 4.1: Since $(e^x - x - 1)/x \to 0$ as $x \to 0$, we know that the first condition $\beta_\rho > 0$ can be satisfied as long as we pick a sufficiently small ρ. In fact, using the inequality $(e^x - x - 1)/x \leq 0.5xe^x$ (when $x \geq 0$), we may also take $\beta_\rho = 1 - 0.5\rho Ke^{\rho M}$ in the first condition of Theorem 4.2.

C. Convergence bounds under various local prior decay conditions

We study consequences of (10) under some general conditions on the local prior structure $\pi(\epsilon) = \pi(\{\theta : R(\theta) \leq \epsilon\})$ around the best achievable parameter. For some specific forms of local prior conditions, convergence rates can be stated very explicitly.

For simplicity, we shall focus only on the Gibbs algorithm (5). Similar results can also be obtained for (6) under an appropriate discrete net. We first present a theorem which contains the general result, followed by more specific consequences. The proofs are left to Section VII-B. Related results can also be found in Chapter 3 of [5].

Theorem 4.3: Consider the Gibbs algorithm (5). If (10) holds with a non-negative function $R(\theta)$, then

$$\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_\rho}R(\theta) \leq \frac{\Delta(\alpha\beta_\rho, \rho n)}{(1-\alpha)\beta_\rho\rho n},$$

where

$$\begin{aligned}\Delta(a,b) =& \ln \frac{\mathbf{E}_{\theta\sim\pi} e^{-abR(\theta)}}{\mathbf{E}_{\theta\sim\pi} e^{-bR(\theta)}}\\ \le& \ln \inf_{u,v}\left[\sup_{\epsilon\le u} \frac{\max(0,\pi(\epsilon/a)-v)}{\pi(\epsilon)}\right.\\ &\left. + \inf_{\epsilon} \frac{v+(1-v)\exp(-bu)}{\pi(\epsilon)e^{-b\epsilon}}\right],\end{aligned}$$

and $\pi(\epsilon) = \pi(\{\theta : R(\theta) \le \epsilon\})$.

In the following, we give two simplified bounds, one with global entropy, which gives the correct rate of convergence for non-parametric problems. The other bound is a refinement that uses localized entropy, useful for parametric problems.

Corollary 4.1 (Global Entropy Bound): Consider the Gibbs algorithm (5). If (10) holds, then

$$\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \le \frac{\inf_\epsilon[\rho n\epsilon - \ln\pi(\epsilon)]}{\beta_\rho \rho n} \le \frac{2\bar{\epsilon}_{global}}{\beta_\rho},$$

where $\pi(\epsilon) = \pi(\{\theta : R(\theta) \le \epsilon\})$ and $\bar{\epsilon}_{global} = \inf\left\{\epsilon : \epsilon \ge \frac{1}{\rho n}\ln\frac{1}{\pi(\epsilon)}\right\}$.

Corollary 4.2 (Local Entropy Bound): Consider the Gibbs algorithm (5). If (10) holds, then

$$\begin{aligned}&\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \le \frac{1}{(1-\alpha)\beta_\rho\rho n}\inf_{u_1\le u_2}\\ &\ln\left[\sup_{\epsilon\in[u_1,u_2]}\frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)} + \exp(\frac{\rho n u_1}{\alpha\beta_\rho}) + \frac{e^{-\rho n u_2/2}}{\pi(u_2/2)}\right]\\ \le& \frac{\bar{\epsilon}_{local}}{(1-\alpha)\beta_\rho},\end{aligned}$$

where $\pi(\epsilon) = \pi(\{\theta : R(\theta) \le \epsilon\})$, and

$$\begin{aligned}\bar{\epsilon}_{local} =& \frac{2}{\rho n} + \inf\left\{\frac{\epsilon}{\alpha\beta_\rho} : \epsilon \ge\right.\\ &\left.\sup_{\epsilon'\in[\epsilon,2u]}\frac{\alpha\beta_\rho}{\rho n}\ln\left[\frac{\pi(\epsilon'/(\alpha\beta_\rho))}{\pi(\epsilon')} + \frac{e^{-\rho n u}}{\pi(u)}\right]\right\}.\end{aligned}$$

Remark 4.2: By letting $u \to \infty$ in the definition of $\bar{\epsilon}_{local}$, we can see easily that $\bar{\epsilon}_{local} \le \ln 2/(\rho n) + \bar{\epsilon}_{global}/(\alpha\beta_\rho)$. Therefore using the localized complexity $\bar{\epsilon}_{local}$ is always better (up to a constant) than using $\bar{\epsilon}_{global}$. If the ratio $\pi(\epsilon/(\alpha\beta_\rho))/\pi(\epsilon)$ is much smaller than $1/\pi(\epsilon)$, the localized complexity can be much better than the global complexity.

In the following, we consider three cases of local prior structures, and derive the corresponding rates of convergence. Comparable lower-bounds are given in Section VI.

1) Non-parametric type local prior: It is well known that for standard nonparametric families such as smoothing splines, etc, the ϵ-entropy often grows at the order of $O(\epsilon^{-r})$ for some $r > 0$. We shall not list detailed examples here, and simply refer the readers to [7]–[9] and references therein. Similarly, we assume that there exists constants C and r such that the prior $\pi(\epsilon)$ satisfies the condition:

$$C_1\epsilon^{-r} \le \ln\frac{1}{\pi(\epsilon)} \le C_2\epsilon^{-r}.$$

This implies that $\bar{\epsilon}_{global} \le (C_2/(\rho n))^{1/(1+r)}$. It is easy to check that $\bar{\epsilon}_{local}$ is the same order of $\bar{\epsilon}_{global}$ when $C_1 > 0$. Therefore, for a prior that behaves non-parametrically around the truth, it does not matter whether we use global complexity or local complexity.

2) Parametric type local prior: For standard parametric families, the prior π has a density with an underlying dimensionality d: $\pi(\epsilon) = O(\epsilon^{-d})$. We may assume that the following condition holds:

$$C_1 + d\ln\frac{1}{\epsilon} \le \ln\frac{1}{\pi(\epsilon)} \le C_2 + d\ln\frac{1}{\epsilon}.$$

This implies that $\bar{\epsilon}_{global}$ is of the order $d\ln n/n$. However, we have

$$\bar{\epsilon}_{local} \le \frac{\ln 2 + C_2 - C_1 - d\ln(\alpha\beta_\rho)}{\rho n},$$

which is of the order $O(d/n)$ for large d. In this case, we obtain a better rate of convergence using the localized complexity measure.

3) Singular local prior: It is possible to obtain a rate of convergence faster than $O(1/n)$. This cannot be obtained with either $\bar{\epsilon}_{global}$ or $\bar{\epsilon}_{local}$, which are of the order no better than n^{-1}. The phenomenon of faster than $O(1/n)$ convergence rate is related to super-efficiency and hence can only appear at a countably many isolated points.

To see that it is possible to obtain faster than $1/n$ convergence rate (super efficiency) in our framework, we only consider the simple case where there exists $u > 0$ such that

$$\pi(2u/(\alpha\beta_\rho)) = \pi(0) > 0.$$

That is, we have a point-like prior mass at the truth with zero density around it. In this case, we can apply Corollary 4.2 with $u_1 = -\infty$ and $u_2 = 2u$, and obtain

$$\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \le \frac{\ln\left[1 + \frac{\exp(-\rho n u)}{\pi(u)}\right]}{(1-\alpha)\beta_\rho\rho n}.$$

This gives an exponential rate of convergence. Clearly this example can be generalized to the case that a point is not completely isolated from its neighbor.

V. Some Examples

We focus on consequences and applications of Theorem 4.2. Specifically, we give a few examples of some important IRM formulations for which (10) hold with some positive constants α, ρ, and β_ρ. It follows that for these problems, the convergence rate analysis in Section IV-C applies.

A. Conditional density estimation

Conditional density estimation is very useful in practical applications. It includes the standard density estimation problem widely studied in statistics as a special case. Moreover, many classification algorithms (such as decision trees or logistic regression) can be considered as conditional density estimators.

Let $Z_1 = (X_1, Y_1)$, where X_1 is the input variable, and Y_1 is the output variable. We are interested in estimating the conditional density $p(Y_1|X_1)$. In this framework, we assume (with a slight abuse of notation) that each parameter θ corresponds to a conditional density function: $\theta(Z_1) = p(Y_1|\theta, X_1)$. In density estimation, we consider the negative log loss function $-\ln\theta(Z_1)$. Our goal is to find a randomized conditional density estimator θ from the data, such that the expected log-loss $-\mathbf{E}_{Z_1}\ln\theta(Z_1)$ is as small as possible.

In this case, the IRM estimator in (4) becomes

$$\hat{\pi}_{\rho,S} = \arg\inf_{\hat{\pi}\in S}\left[\rho\mathbf{E}_{\theta\sim\hat{\pi}}\sum_{i=1}^{n}\ln\frac{1}{\theta(\hat{Z}_i)} + D_{KL}(\hat{\pi}||\pi)\right]. \tag{11}$$

If we let S be the set of all possible measures, then the optimal solution is $d\hat{\pi}_{\rho,S} \propto \prod_{i=1}^{n}\theta(\hat{Z}_i)^\rho d\pi$, which corresponds to the Bayesian posterior distribution when $\rho = 1$. For discrete S, the method is essentially a two-part code MDL estimator. As mentioned earlier, Theorem 4.1 can be directly applied since the left hand side can be interpreted as a (Hellinger-like) distance between distributions. This approach has been taken in [1], [6] (also see [10] for related results). However, in this section, we are interested in using the log-loss on the left-hand side. This requires rewriting the left-hand side logarithmic moment generating function of Theorem 4.1 using ideas illustrated at the beginning of the section.

We further assume that θ is defined on a domain $\mathcal{G}$ which is a closed convex density class. However, we do not assume that $\mathcal{G}$ contains the true conditional density. We also let $\theta_{\mathcal{G}}$ be the optimal density in $\mathcal{G}$ with respect to the log loss:

$$\mathbf{E}_{Z_1}\ln\frac{1}{\theta_{\mathcal{G}}(Z_1)} = \inf_{\theta\in\mathcal{G}}\mathbf{E}_{Z_1}\ln\frac{1}{\theta(Z_1)}.$$

In the following, we are interested in a bound which compare the performance of the randomized estimator (11) to the best possible predictor $\theta_{\mathcal{G}} \in \mathcal{G}$, and thus define

$$\ell_\theta(Z_1) = \ln\frac{\theta_{\mathcal{G}}(Z_1)}{\theta(Z_1)}.$$

In order to apply Theorem 4.1, we need the following variance bound which is of the form (9) (see Section VII-C).

Proposition 5.1: If there exists a constant $M_{\mathcal{G}} \geq 2/3$ such that $\mathbf{E}_{Z_1}\ell_\theta(Z_1)^3 \leq M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2$, then $\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \leq \frac{8M_{\mathcal{G}}}{3}\mathbf{E}_{Z_1}\ell_\theta(Z_1)$.
Using this result, we obtain the following theorem from Theorem 4.2.

Theorem 5.1: Consider the IRM estimator (11) for conditional density estimation (under log-loss). Then $\forall\alpha \in [0,1)$, inequality (10) holds with $R(\theta) = \mathbf{E}_{Z_1}\ln\frac{\theta_{\mathcal{G}}(Z_1)}{\theta(Z_1)}$ under either of the following two conditions:

- $\sup_{\theta_1,\theta_2\in\mathcal{G},Z_1}\ln\frac{\theta_1(Z_1)}{\theta_2(Z_1)} \leq M_{\mathcal{G}}$: we pick ρ such that $\beta_\rho = (11\rho M_{\mathcal{G}} + 8 - 8e^{\rho M_{\mathcal{G}}})/(3\rho M_{\mathcal{G}}) > 0$.
- There exist $M_{\mathcal{G}}, b > 0$ such that $M_{\mathcal{G}}b \geq 1/9$, and $\forall\theta \in \mathcal{G}$ and $m \geq 3$, $\mathbf{E}_{Z_1}|\ln\frac{\theta(Z_1)}{\theta_{\mathcal{G}}(Z_1)}|^m \leq m!M_{\mathcal{G}}^{m-2}b\mathbf{E}_{Z_1}(\ln\frac{\theta(Z_1)}{\theta_{\mathcal{G}}(Z_1)})^2$: we pick ρ such that $\beta_\rho = 1-8b\rho M_{\mathcal{G}}(1-\rho M_{\mathcal{G}}+2b\rho M_{\mathcal{G}})/(1-\rho M_{\mathcal{G}}) > 0$.

Proof: Under the first condition, using Proposition 5.1, we may take $K = 8/3M_{\mathcal{G}}$ and $M = M_{\mathcal{G}}$ in Theorem 4.2 (bounded loss case). Under the second condition, using Proposition 5.1, we may take $K = 16M_{\mathcal{G}}b$ and $M = M_{\mathcal{G}}$ in Theorem 4.2 (Bernstein loss case). ■

Similar to the remark after Theorem 4.2, we may also let $\beta_\rho = (3 - 4\rho M_{\mathcal{G}}e^{\rho M_{\mathcal{G}}})/3$ under the first condition of Theorem 5.1. The second condition involves moment inequalities that needs to be verified for specific problems. It applies to certain unbounded conditional density families such as conditional Gaussian models with bounded variance. We shall discuss a related scenario in the least squares regression case. Note that under Gaussian noise with identical variance, the conditional density estimation using the log-loss is equivalent to the estimation of conditional mean using least squares regression.

Since for log-loss, (10) holds under appropriate boundedness or moment assumptions on the density family, the consequences in Section IV-C apply to the Gibbs algorithm (which gives the Bayesian posterior distribution when $\rho = 1$). As we shall show in Section VI, similar lower bounds can be derived.

B. Least squares regression

Let $Z_1 = (X_1, Y_1)$, where X_1 is the input variable, and Y_1 is the output variable. We are interested in predicting Y_1 based on X_1. We assume that each parameter θ corresponds to a predictor: $\theta(X_1)$. The quality of the predictor is measured by the mean squared error $\mathbf{E}_{Z_1}(\theta(X_1)-Y_1)^2$. In this framework, the IRM estimator in (4) becomes

$$\hat{\pi}_{\rho,S} = \arg\inf_{\hat{\pi}\in S}\left[\mathbf{E}_{\theta\sim\hat{\pi}}\sum_{i=1}^{n}(\theta(X_i)-Y_i)^2 + \frac{D_{KL}(\hat{\pi}||\pi)}{\rho}\right]. \qquad (12)$$

We further assume that θ is defined on a domain $\mathcal{G}$, which is a closed convex function class. Let $\theta_{\mathcal{G}}$ be the optimal predictor in $\mathcal{G}$ with respect to the least squares loss:

$$\mathbf{E}_{Z_1}(\theta_{\mathcal{G}}(X_1)-Y_1)^2 = \min_{\theta\in\mathcal{G}}\mathbf{E}_{Z_1}(\theta(X_1)-Y_1)^2.$$

In the following, we are interested in comparing the performance of the randomized estimator (11) to the best possible predictor $\theta_{\mathcal{G}} \in \mathcal{G}$. Define

$$\ell_\theta(Z_1) = (\theta(X_1)-Y_1)^2 - (\theta_{\mathcal{G}}(X_1)-Y_1)^2.$$

We have the following proposition. The proof is in Section VII-C.

Proposition 5.2: Let

$$M_{\mathcal{G}} = \sup_{X_1,\theta\in\mathcal{G}}\mathbf{E}_{Y_1|X_1}(\theta(X_1)-Y_1)^2.$$

Then $\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \le 4M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1)$.

The following refined bound estimates the moment of the loss in terms of the moment of the noise $Y_1 - \theta(X_1)$. The proof is left to Section VII-C.

Proposition 5.3: Let $A_{\mathcal{G}} = \sup_{X_1,\theta\in\mathcal{G}}|\theta(X_1) - \theta_{\mathcal{G}}(X_1)|$ and $\sup_{X_1,\theta\in\mathcal{G}}\mathbf{E}_{Y_1|X_1}|\theta(X_1) - Y_1|^m \le m!B_{\mathcal{G}}^{m-2}M_{\mathcal{G}}$ for $m \ge 2$. Then we have:

$$\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \le m!(2A_{\mathcal{G}}B_{\mathcal{G}})^{m-2}4M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1).$$

Although Proposition 5.2 is a special case of Proposition 5.3, we include it separately for convenience. These moment estimates can be combined with Theorem 4.2, and we obtain the following theorem.

Theorem 5.2: Consider the IRM estimator (12) for least squares regression. Then $\forall\alpha \in [0,1)$, inequality (10) holds with $R(\theta) = \mathbf{E}_{Z_1}(\theta(X_1) - Y_1)^2 - \mathbf{E}_{Z_1}(\theta_{\mathcal{G}}(X_1) - Y_1Z)^2$, under either of the following conditions:

- $\sup_{\theta\in\mathcal{G},Z_1}(\theta(X_1) - Y_1)^2 \le M_{\mathcal{G}}$: we pick ρ such that $\beta_\rho = (5\rho M_{\mathcal{G}} + 4 - 4e^{\rho M_{\mathcal{G}}})/(\rho M_{\mathcal{G}}) > 0$.
- Proposition 5.3 holds for all integer $m \ge 2$: we pick small ρ such that $\beta_\rho = 1 - 4M_{\mathcal{G}}\rho/(1-2A_{\mathcal{G}}B_{\mathcal{G}}\rho) > 0$.

Proof: Under the first condition, using Proposition 5.2, we have $M_{\mathcal{G}} \le \sup_{\theta\in\mathcal{G},Z_1}(\theta(X_1) - Y_1)^2$. We may thus take $K = 4M_{\mathcal{G}}$ and $M = M_{\mathcal{G}}$ in Theorem 4.2 (bounded loss case). Under the second condition, using Proposition 5.3, we can let $K = 8M_{\mathcal{G}}$, $M = 2A_{\mathcal{G}}B_{\mathcal{G}}$ and $b = 1/2$ in Theorem 4.2 (Bernstein loss case). ■

Remark 5.1: Similar to the remark after Theorem 4.2, we may also let $\beta_\rho = 1 - 2\rho M_{\mathcal{G}}e^{\rho M_{\mathcal{G}}}$ under the first (bounded loss) condition.

The theorem applies to unbounded regression problems with exponentially decaying noise such as Gaussian noise (also see [3]). For example, we have the following result (see Section VII-C for its proof).

Corollary 5.1: Assume that there exists a function $y_0(X)$ such that

- For all X_1, the random variable $|Y_1 - y_0(X_1)|$, conditioned on X_1, is dominated by the absolute value of a zero-mean Gaussian random variable[1] with standard deviation σ.
- $\exists$ constant $b > 0$ such that $\sup_{X_1}|y_0(X) - \theta(X_1)| \le b$.

If we also choose A such that $A \ge \sup_{X_1,\theta\in\mathcal{G}}|\theta(X_1) - \theta_{\mathcal{G}}(X_1)|$, then (10) holds with $\beta_\rho = 1 - 4\rho(b+\sigma)^2/(1 - 2A(b+\sigma)\rho) > 0$.

Although we have only discussed least squares regression, similar results can also be obtained for some other loss functions in regression and classification.

C. Non-negative loss functions

Consider a non-negative loss function $\ell_\theta(Z_i) \ge 0$. We are interested in developing a bound under the assumption that the best error $\mathbf{E}_{\theta\in\mathcal{G}}\ell_\theta(Z_i)$ can approach zero. A straight-forward application of Theorem 4.2 leads to the following bound.

Theorem 5.3: Consider the IRM estimator (4). Assume $\ell_\theta(Z_i) \ge 0$, and there exist positive constants K such that $\forall\theta$:

$$\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \le K\mathbf{E}_{Z_1}\ell_\theta(Z_1).$$

Then $\forall\alpha \in [0,1)$, inequality (10) holds with $R(\theta) = \mathbf{E}_{Z_1}\ell_\theta(Z_1)$ and $\beta_\rho = 1 - 0.5\rho K$.

Proof: We can apply Theorem 4.2, under the bounded loss condition $-\ell_\theta(Z_1) \le M$, which holds

[1] That is, conditioned on X_1, the moments of $|Y_1 - y_0(X_1)|$ with respect to Y_1 are no larger than the corresponding moments of the dominating Gaussian random variable.

for any $M > 0$. Letting $M \to 0$, we have $\beta_\rho = 1 - K(e^{\rho M} - \rho M - 1)/(\rho M^2) \to 1 - 0.5K\rho$. ■

The condition in Theorem 5.3 is automatically satisfied if $\ell_\theta(Z_1)$ is uniformly bounded: $0 \le \ell_\theta(Z_1) \le K$.

VI. Lower Bounds

In previous sections, we obtained performance bounds for the IRM method (4) using the information exponential inequality. The purpose of this section is to prove some lower bounds which hold for arbitrary statistical estimators. Our goal is to match these lower bounds to the upper bounds proved earlier for the IRM method (at least for certain problems), which implies that the IRM method is near optimal.

The upper bounds we obtained in previous sections are for every possible realization of the underlying distribution. It is not possible to obtain a lower bound for any specific realization since we can always design an estimator that picks a parameter that achieves the best possible performance under this particular distribution. However, such an estimator will not work well for a different distribution. Therefore as far as lower bounds are concerned, we are interested in the performance averaged over a set of underlying distributions.

In order to obtain lower bounds, we associate each parameter θ with a probability distribution $P_\theta(x, y)$ so that we can take samples $Z_i = (X_i, Y_i)$ from this distribution. In addition, we shall design the map in such a way that the optimal parameter under this distribution is θ. For (conditional) density estimation, the map is the density itself. For regression, we associate each predictor θ with a conditional Gaussian distribution with constant variance and the conditional mean given by the prediction $\theta(X_1)$ of each input X_1.

We consider the following scenario: we put a prior π on θ, which becomes a prior on the distribution $P_\theta(x, y)$. Assume that we are interested in estimating θ, under a loss function $\ell_\theta(Z_1)$, then the quantity

$$R_\theta(\theta') = \mathbf{E}_{Z_1 \sim P_\theta} \ell_{\theta'}(Z_1)$$

is the true risk between an estimated parameter θ' and the true distribution parameter θ. Let Z be n independent samples $Z = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from the underlying distribution P_θ. Denote by P_θ^n the product distribution: $dP_\theta^n(Z) = \prod_{i=1}^n dP_\theta(Z_i)$. Let $\hat{\pi}(Z)$ be an arbitrary randomized estimator, and consider a random variable $\hat{\theta}$ that is drawn from $\hat{\pi}(Z)$. The average performance of $\hat{\theta}$ can thus be expressed as

$$\mathbf{E}_{\theta \sim \pi} \mathbf{E}_{Z \sim P_\theta^n} \mathbf{E}_{\hat{\theta} \sim \hat{\pi}(Z)} R_\theta(\hat{\theta}), \tag{13}$$

In this section, we are mainly interested in obtaining a lower bound for any possible estimator, so that we can compare this lower bound to the upper bound for the IRM method developed earlier.

Note that (13) only gives one performance measure, while the upper bound for the IRM method is specific for every possible truth P_θ. It is thus useful to study the best local performance around any possible θ with respect to the underlying prior π. To address this issue, we observe that for every partition of the θ space into the union of disjoint small balls B_k, we may rewrite (13) as

$$\sum_k \pi(B_k) \mathbf{E}_{\theta \sim \pi_{B_k}} \mathbf{E}_{Z \sim P_\theta^n} \mathbf{E}_{\hat{\theta} \sim \hat{\pi}(Z)} R_\theta(\hat{\theta}),$$

where for each small ball B_k, the localized prior is defined as:

$$\pi_{B_k}(A) = \frac{\pi(A \cap B_k)}{\pi(B_k)}.$$

Therefore, instead of bounding the optimal Bayes risk with respect to the global prior π in (13), we shall bound the optimal risk with respect to a local prior π_B for a small ball B around any specific parameter θ, which gives a more refined performance measure. In this framework, if for some small local ball π_B, the IRM method has performance not much worse than the best possible estimator, then we can say that it is *locally near optimal.*

The main theorem in our lower bound analysis is presented below. The proof, which relies on a simple information theoretical argument, is given in Section VII-D.

Theorem 6.1: Consider an arbitrary randomized estimator $\hat{\pi}(Z)$ on $B' \subset \mathcal{G}$, where Z consists of n independent samples $Z = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from some underlying density P_θ, then for all non-negative function $R_\theta(\theta')$, we have for all B that contains θ:

$$\mathbf{E}_{\theta \sim \pi_B} \mathbf{E}_{Z \sim P_\theta^n} \mathbf{E}_{\hat{\theta} \sim \hat{\pi}(Z)} R_\theta(\hat{\theta}) \ge 0.5$$
$$\sup \left\{ \epsilon : \inf_{\theta' \in B'} \ln \frac{1}{\pi_B(\{\theta : R_\theta(\theta') < \epsilon\})} \ge 2n\Delta_B + \ln 4 \right\},$$

where $\Delta_B = \mathbf{E}_{\theta \sim \pi_B} \mathbf{E}_{\theta' \sim \pi_B} D_{KL}(P_\theta(Z_1) || P_{\theta'}(Z_1))$.

Theorem 6.1 has a form that resembles Corollary 4.2. In the following, we state a result which shows the relationship more explicitly.

Corollary 6.1 (Local Entropy Lower Bound): Under the notations of Theorem 6.1, consider a reference point $\theta_0 \in \mathcal{G}$, and a set of balls $B(\theta_0, \epsilon) \subset \mathcal{G}$ containing θ_0 (by definition) and indexed by $\epsilon > 0$, such that

$$\sup_{\theta, \theta' \in B(\theta_0, \epsilon)} D_{KL}(P_\theta(Z_1) || P_{\theta'}(Z_1)) \le \epsilon.$$

Given $u > 0$, and consider $\underline{\epsilon}(\theta_0, u)$ which satisfies:

$$\underline{\epsilon}(\theta_0, u) = \sup_{\epsilon > 0} \{\epsilon : \\ \inf_{\theta' \in B'} \ln \frac{\pi(B(\theta_0, u\epsilon))}{\pi(\{\theta : R_\theta(\theta') < \epsilon\} \cap B(\theta_0, u\epsilon))} \geq 2nu\epsilon + \ln 4 \Big\},$$

then locally around θ_0, we have

$$\mathbf{E}_{\theta \sim \pi_{B(\theta_0, u\underline{\epsilon}(\theta_0,u))}} \mathbf{E}_{Z \sim P_\theta^n} \mathbf{E}_{\hat{\theta} \sim \hat{\pi}(Z)} R_\theta(\hat{\theta}) \geq 0.5 \underline{\epsilon}(\theta_0, u).$$

Proof: The corollary is a simple application of Theorem 6.1. Just let $B = B(\theta_0, u\underline{\epsilon}(\theta_0, u))$ and note that $\Delta_B \leq u\underline{\epsilon}(\theta_0, u)$. ■

The definition of $B(\theta_0, \epsilon)$ requires that within the $B(\theta_0, \epsilon)$ ball, the distributions P_θ are nearly indistinguishable up to a scale of ϵ, when measured by their KL-divergence. Corollary 6.1 implies that the local performance of an arbitrary statistical estimator cannot be better than $\underline{\epsilon}(\theta_0, u)/2$. The bound in Corollary 6.1 will be good if the ball $\pi(B(\theta_0, \epsilon))$ is relatively large. That is, there are many distributions that are statistically nearly indistinguishable (in KL-distance). Therefore the bound of Corollary 6.1 is similar to Corollary 4.2, but the localization is within a small ball which is statistically nearly indistinguishable (rather than the $R(\cdot)$ localization for the IRM estimator). From an information theoretical point of view, this difference is rather intuitive and clearly also necessary since we allow arbitrary statistical estimators (which can simply estimate the specific underlying distribution P_θ if they are distinguishable).

It follows that if we want the upper bound in Corollary 4.2 to match the lower bound in Corollary 6.1, we need to design a map $\theta \to P_\theta$ such that locally around θ_0, a ball with small $R(\cdot)$ risk is also small information theoretically in terms of the KL-distance between P_θ and P_{θ_0}. Consider the following two types of small R balls:

$$B_1(\theta, \epsilon) = \{\theta' : R_\theta(\theta') < \epsilon\},$$
$$B_2(\theta, \epsilon) = \{\theta' : R_{\theta'}(\theta) < \epsilon\}.$$

Now assume that we can find a map $\theta \to P_\theta$ such that locally around θ_0, P_θ within a small B_1-ball is small in the information theoretical sense (small KL-distance). That is, we have for some $c > 0$ that

$$\sup\{D_{KL}(P_\theta(Z_1)||P_{\theta'}(Z_1)) : \theta, \theta' \in B_1(\theta_0, c\epsilon)\} \leq \epsilon. \tag{14}$$

For problems such as density estimation and regression studied in this paper, it is easy to design such a map (under mild conditions such as the boundedness of the loss). We shall not go into the details for verifying specific examples of (14). Now, under this condition, we can take

$$\underline{\epsilon}(\theta_0, u) = \sup_{\epsilon > 0} \{\epsilon : \\ \inf_{\theta' \in B'} \ln \frac{\pi(B_1(\theta_0, cu\epsilon))}{\pi(B_2(\theta', \epsilon) \cap B_1(\theta_0, cu\epsilon))} \geq 2nu\epsilon + \ln 4 \Big\}.$$

As a comparison, according to Corollary 4.2, the IRM method at θ_0 gives an upper bound of the following form (which we simplify to focus on the main term) with some constant $u' \in (0, 1)$:

$$\bar{\epsilon}_{local} \leq \frac{2}{\rho n} + \inf \left\{ \epsilon : \rho n \epsilon \geq \sup_{\epsilon' \geq \epsilon} \ln \frac{\pi(B_1(\theta_0, \epsilon'))}{\pi(B_1(\theta_0, u'\epsilon'))} \right\}.$$

Essentially, the local IRM upper bound is achieved at $\bar{\epsilon}_{local}$ such that

$$n\bar{\epsilon}_{local} \sim \sup_{\epsilon' \geq \bar{\epsilon}_{local}} \ln \frac{\pi(B_1(\theta_0, \epsilon'))}{\pi(B_1(\theta_0, u'\epsilon'))},$$

where we use $\sim$ to denote approximately the same order, while the lower bound in Corollary 6.1 implies that (let $u' = 1/(cu)$):

$$n\underline{\epsilon} \sim \inf_{\theta' \in B'} \ln \frac{\pi(B_1(\theta_0, \underline{\epsilon}))}{\pi(B_2(\theta', u'\underline{\epsilon}) \cap B_1(\theta_0, \underline{\epsilon}))}.$$

From this, we see that our upper and lower bounds are very similar. There are two main differences which we outline below.

- In the lower bound, for technical reasons, B_2 appears in the definition of the local entropy. In order to argue that the difference does not matter, we need to assume that the prior probabilities of B_1 and B_2 are of the same order.
- In the lower bound, we use the smallest local entropy in a neighbor of θ_0, while in the upper bound, we use the largest local entropy at θ_0 across different scales. This difference is not surprising since the lower bound is with respect to the average in a small neighborhood of θ_0.

Both differences are relatively mild and somewhat expected. In fact, at the conceptual level, it is very interesting that we are able to establish lower and upper bounds that are so similar. We consider two situations which parallel Section IV-C.1 and Section IV-C.2.

A. Non-parametric type local prior

Similar to Section IV-C.1, we assume that for some sufficiently large constant v: there exist $0 < C_1 < C_2$

such that

$$C_2\epsilon^{-r} \le \inf_{\theta' \in B'} \ln \frac{1}{\pi(B_2(\theta',\epsilon) \cap B_1(\theta_0, v\epsilon))},$$
$$\ln \frac{1}{\pi(B_1(\theta_0, v\epsilon))} \le C_1\epsilon^{-r},$$

which measures the order of global entropy around a small neighborhood of θ_0. Now under condition (14) and with $u = v/c$, Corollary 6.1 implies that

$$\underline{\epsilon} \ge \sup\left\{\epsilon : 2un\epsilon + 4 \le (C_2 - C_1)\epsilon^{-r}\right\}.$$

This implies that $\underline{\epsilon}$ is of the order $n^{-1/(1+r)}$, which matches the order of the IRM upper bound $\bar{\epsilon}_{global}$ in Section IV-C.1.

B. Parametric type local prior

Similar to Section IV-C.2, we assume that for some sufficiently large constant v: there exist $0 < C_1 < C_2$ such that

$$C_2 + d\ln\frac{1}{\epsilon} \le \inf_{\theta' \in B'} \ln \frac{1}{\pi(B_2(\theta',\epsilon) \cap B_1(\theta_0, v\epsilon))},$$
$$\ln \frac{1}{\pi(B_1(\theta_0, v\epsilon))} \le C_1 + d\ln\frac{1}{\epsilon},$$

which measures the order of global entropy around a small neighborhood of θ_0. Now under the condition (14) and with $u = v/c$, Corollary 6.1 implies that

$$\underline{\epsilon} \ge \sup\left\{\epsilon : 2c^{-1}n\epsilon \le C_2 - C_1 - 4\right\}.$$

That is, we have a convergence rate of the order $1/n$, which matches the parametric upper bound $\bar{\epsilon}_{local}$ for IRM in Section IV-C.2.

VII. Proofs

In order to avoid clutter in the main text and improve the readability, we put relatively long proofs into this section. For convenience, we restate the theorems before the proofs.

A. Proofs for Section II

The key ingredient in our proof of the information exponential inequality is a well-known convex duality, which employs KL-complexity, and has already been used in some recent machine learning papers to study sample complexity bounds [2], [4], [5], [11], [12].

Proposition 7.1: Assume that $f(\theta)$ is a measurable real-valued function on $\mathcal{G}$, and $\hat{\pi}$ is a probability measure, we have

$$\mathbf{E}_{\theta\sim\hat{\pi}} f(\theta) \le D_{KL}(\hat{\pi}||\pi) + \ln \mathbf{E}_{\theta\sim\pi} \exp(f(\theta)).$$

Remark 7.1: The above convex duality also has a straight-forward information theoretical interpretation: consider $v(\theta) = \exp(f(\theta))/\mathbf{E}_{\theta\sim\pi}\exp(f(\theta))$. Since $\mathbf{E}_{\theta\sim\pi}v(\theta) = 1$, we can regard it as a density with respect to π. Now let $d\pi' = vd\pi$, it is easy to verify that the inequality in Proposition 7.1 can be rewritten as $D_{KL}(\hat{\pi}||\pi') \ge 0$, which is a well-known information theoretical inequality, and follows easily from Jensen's inequality.

Lemma 2.1 Consider a measurable real-valued function $L_\theta(Z)$ on $\mathcal{G} \times \mathcal{Z}$. Then for an arbitrary sample $\hat{Z}$-dependent randomized estimation method, with posterior $\hat{\pi}$, the following inequality holds

$$\mathbf{E}_{\hat{Z}} \exp\left[\mathbf{E}_{\theta\sim\hat{\pi}}\Delta_{\hat{Z}}(\theta) - D_{KL}(\hat{\pi}||\pi)\right] \le 1.$$

where $\Delta_{\hat{Z}}(\theta) = -L_\theta(\hat{Z}) - \ln \mathbf{E}_Z \exp(-L_\theta(Z))$.

Proof: Using the convex duality inequality in Proposition 7.1 on the parameter space $\mathcal{G}$, with prior π, posterior $\hat{\pi}$, and $f(\theta) = \Delta_{\hat{Z}}(\theta)$, we obtain

$$\exp\left[\mathbf{E}_{\theta\sim\hat{\pi}}\Delta_{\hat{Z}}(\theta) - D_{KL}(\hat{\pi}||\pi)\right] \le \mathbf{E}_{\theta\sim\pi}e^{\Delta_{\hat{Z}}(\theta)}.$$

Taking expectation with respect to $\hat{Z}$, and notice that $\mathbf{E}_{\hat{Z}}\exp(\Delta_{\hat{Z}}(\theta)) = 1$, we obtain

$$\begin{aligned}&\mathbf{E}_{\hat{Z}} \exp\left[\mathbf{E}_{\theta\sim\hat{\pi}}\Delta_{\hat{Z}}(\theta) - D_{KL}(\hat{\pi}||\pi)\right]\\ \le&\mathbf{E}_{\hat{Z}}\mathbf{E}_{\theta\sim\pi}\exp(\Delta_{\hat{Z}}(\theta)) = \mathbf{E}_{\theta\sim\pi}\mathbf{E}_{\hat{Z}}\exp(\Delta_{\hat{Z}}(\theta)) = 1.\end{aligned}$$

This proves the inequality. ■

Theorem 2.1 Consider randomized estimation, where we select posterior $\hat{\pi}$ on $\mathcal{G}$ based on $\hat{Z}$, with π a prior. Consider a real-valued function $L_\theta(Z)$ on $\mathcal{G} \times \mathcal{Z}$. Then for all t, the following event holds with probability at least $1 - \exp(-t)$:

$$-\mathbf{E}_{\theta\sim\hat{\pi}} \ln \mathbf{E}_Z\, e^{-L_\theta(Z)} \le \mathbf{E}_{\theta\sim\hat{\pi}}L_\theta(\hat{Z}) + D_{KL}(\hat{\pi}||\pi) + t.$$

Moreover, we have the following expected risk bound:

$$\begin{aligned}&-\mathbf{E}_{\hat{Z}}\mathbf{E}_{\theta\sim\hat{\pi}} \ln \mathbf{E}_Z\, e^{-L_\theta(Z)}\\ \le&\mathbf{E}_{\hat{Z}}\left[\mathbf{E}_{\theta\sim\hat{\pi}}L_\theta(\hat{Z}) + D_{KL}(\hat{\pi}||\pi)\right].\end{aligned}$$

Proof: Let

$$\hat{\delta}_{\hat{Z}} = \mathbf{E}_{\theta\sim\hat{\pi}}\Delta_{\hat{Z}}(\theta) - D_{KL}(\hat{\pi}||\pi),$$

then from Lemma 2.1, we have $\mathbf{E}_{\hat{Z}}\exp(\hat{\delta}_{\hat{Z}}) \le 1$. This implies that $\mathbf{P}(\hat{\delta}_{\hat{Z}} \ge t)\exp(t) \le 1$. Therefore with probability at most $\exp(-t)$, $\hat{\delta}_{\hat{Z}} \ge t$. Rewriting this expression gives the first inequality. In order to prove the second inequality, we note that $\mathbf{E}_{\hat{Z}}\hat{\delta}_{\hat{Z}} \le \ln \mathbf{E}_{\hat{Z}} \exp(\hat{\delta}_{\hat{Z}}) \le \ln 1 = 0$. Again, rewriting this expression gives the second inequality. ■

B. Proofs for Section IV

Theorem 4.1 Define resolvability

$$r_{\rho,S} = \inf_{\pi' \in S} \left[\mathbf{E}_{\theta\sim\pi'} \mathbf{E}_{Z_1} \ell_\theta(Z_1) + \frac{1}{\rho n} D_{KL}(\pi'||\pi) \right].$$

Then for all $\alpha \in [0,1)$, the expected generalization performance of IRM method (4) can be bounded as

$$\begin{aligned} &- \mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}} \ln \mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)} \\ \le& \frac{\rho}{1-\alpha} \left[r_{\rho,S} + \frac{1}{\rho n} \ln \mathbf{E}_{\theta\sim\pi} \mathbf{E}_{Z_1}^{\alpha n} e^{-\rho\ell_\theta(Z_1)} \right]. \end{aligned}$$

Proof: We obtain from (4)

$$\begin{aligned} &\mathbf{E}_{\hat{Z}} \left[\rho \mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}} \sum_{i=1}^{n} \ell_\theta(\hat{Z}_i) + D_{KL}(\hat{\pi}_{\rho,S}||\pi) \right] \\ \le& \inf_{\pi'\in S} \mathbf{E}_{\hat{Z}} \left[\rho \mathbf{E}_{\theta\sim\pi'} \sum_{i=1}^{n} \ell_\theta(\hat{Z}_i) + D_{KL}(\pi'||\pi) \right] \\ \le& \inf_{\pi'\in S} \left[\rho n \mathbf{E}_{\theta\sim\pi'} \mathbf{E}_{\hat{Z}_i} \ell_\theta(\hat{Z}_1) + D_{KL}(\pi'||\pi) \right] = \rho n r_{\rho,S}. \end{aligned}$$

Let $L_\theta(\hat{Z}) = \rho \sum_{i=1}^{n} \ell_\theta(\hat{Z}_i)$. Substituting the above bound into the right hand side of the second inequality of Corollary 3.2, and using the fact that $\mathbf{E}_Z e^{-L_\theta(Z)} = \mathbf{E}_{Z_1}^n e^{-\rho\ell_\theta(Z_1)}$, we obtain the desired inequality. ■

Theorem 4.2 Consider the IRM method in (4). Assume there exists a positive constant K such that for all θ:

$$\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \le K \mathbf{E}_{Z_1}\ell_\theta(Z_1).$$

Assume either of the following conditions:

- Bounded loss: $\exists M \ge 0$ s.t. $-\inf_{\theta, Z_1} \ell_\theta(Z_1) \le M$; let $\beta_\rho = 1 - K(e^{\rho M} - \rho M - 1)/(\rho M^2)$.
- Bernstein loss: $\exists M, b > 0$ s.t. $\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \le m! M^{m-2} K b \mathbf{E}_{Z_1}\ell_\theta(Z_1)$ for all $\theta \in \mathcal{G}$ and integer $m \ge 3$; let $\beta_\rho = 1 - K\rho(1 - \rho M + 2b\rho M)/(2 - 2\rho M)$.

Assume we choose a sufficiently small ρ such that $\beta_\rho > 0$. Let the (true) expected loss of θ be $R(\theta) = \mathbf{E}_{Z_1}\ell_\theta(Z_1)$, then the expected generalization performance of the IRM method is bounded by

$$\begin{aligned} &\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_{\rho,S}} R(\theta) \\ \le& \frac{1}{(1-\alpha)\beta_\rho} \left[r_{\rho,S} + \frac{1}{\rho n} \ln \mathbf{E}_{\theta\sim\pi} e^{-\alpha\rho\beta_\rho n R(\theta)} \right], \end{aligned}$$

where $r_{\rho,S} = \inf_{\pi'\in S} \left[\mathbf{E}_{\theta\sim\pi'} R(\theta) + \frac{1}{\rho n} D_{KL}(\pi'||\pi) \right]$.

Proof: Under the bounded-loss condition, using the last bound on logarithmic generating function in Proposition 1.2, we obtain

$$\begin{aligned} &\ln \mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)} \\ \le& -\rho \mathbf{E}_{Z_1}\ell_\theta(Z_1) + \frac{e^{\rho M} - \rho M - 1}{M^2} \mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \\ \le& -\left(\rho - \frac{K}{M^2}(e^{\rho M} - \rho M - 1) \right) \mathbf{E}_{Z_1}\ell_\theta(Z_1) \\ =& -\rho\beta_\rho \mathbf{E}_{Z_1}\ell_\theta(Z_1). \end{aligned}$$

Now substitute this bound into Theorem 4.1, and simplify to obtain the desired result.

Similarly, under the Bernstein-loss condition, the logarithmic moment generating function estimate in Proposition 1.2 implies that

$$\begin{aligned} &\ln \mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)} \\ \le& -\rho \mathbf{E}_{Z_1}\ell_\theta(Z_1) + \frac{K\rho^2}{2} \ell_\theta(Z_1) + \rho^3 K b M \frac{\ell_\theta(Z_1)}{1-\rho M} \\ =& -\rho\beta_\rho \mathbf{E}_{Z_1}\ell_\theta(Z_1). \end{aligned}$$

Now substitute this bound into Theorem 4.1, and simplify to obtain the desired result. ■

Theorem 4.3 Consider the Gibbs algorithm (5). If (10) holds with a non-negative function $R(\theta)$, then

$$\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \le \frac{\Delta(\alpha\beta_\rho, \rho n)}{(1-\alpha)\beta_\rho \rho n},$$

where

$$\begin{aligned} \Delta(a,b) =& \ln \frac{\mathbf{E}_{\theta\sim\pi} e^{-abR(\theta)}}{\mathbf{E}_{\theta\sim\pi} e^{-bR(\theta)}} \\ \le& \ln \inf_{u,v} \left[\sup_{\epsilon\le u} \frac{\max(0, \pi(\epsilon/a) - v)}{\pi(\epsilon)} \right. \\ &\left. + \inf_{\epsilon} \frac{v + (1-v)\exp(-bu)}{\pi(\epsilon)e^{-b\epsilon}} \right], \end{aligned}$$

and $\pi(\epsilon) = \pi(\{\theta : R(\theta) \le \epsilon\})$.

Proof: For the Gibbs algorithm, the resolvability is given by (7), which can be expressed as:

$$\begin{aligned} r_{\rho,S} =& -\frac{1}{\rho n} \ln \mathbf{E}_{\theta\sim\pi} e^{-\rho n R(\theta)} \\ =& -\frac{1}{\rho n} \ln \underbrace{\int \pi(\epsilon/(\rho n))\, e^{-\epsilon} d\epsilon}_{A} . \end{aligned}$$

Similarly, the second term (the localization term) on the right hand side of (10) is

$$\begin{aligned} &\frac{1}{\rho n} \ln \mathbf{E}_{\theta\sim\pi} e^{-\alpha\rho\beta_\rho n R(\theta)} \\ =&\frac{1}{\rho n} \ln \int \pi(\epsilon/(\alpha\beta_\rho \rho n))\, e^{-\epsilon} d\epsilon \\ \leq&\frac{1}{\rho n} \ln \Bigg[\underbrace{\int_0^{\rho n u} (\pi(\epsilon/(\alpha\rho\beta_\rho n)) - v)\, e^{-\epsilon} d\epsilon}_{B} \\ &\qquad + \underbrace{(v + (1-v)e^{-\rho n u})}_{C} \Bigg] . \end{aligned}$$

To finish the proof, we only need to show that $(B + C)/A \leq e^{\Delta(\alpha\beta_\rho,\rho n)}$.

Consider arbitrary real numbers u and v. From the expressions, it is easy to see that $B/A \leq \sup_{\epsilon\leq u} \frac{\max(0,\pi(\epsilon/(\alpha\beta_\rho))-v)}{\pi(\epsilon)}$. Moreover, since $A \geq \sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})$, we have $C/A \leq C/\sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})$. Combining these inequalities, we have $(B + C)/A \leq e^{\Delta(\alpha\beta_\rho,\rho n)}$. The desired bound is now a direct consequence of (10). ■

Corollary 4.1 Consider the Gibbs algorithm (5). If (10) holds, then

$$\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \leq \frac{\inf_\epsilon[\rho n\epsilon - \ln\pi(\epsilon)]}{\beta_\rho \rho n} \leq \frac{2\bar{\epsilon}_{global}}{\beta_\rho},$$

where $\pi(\epsilon) = \pi(\{\theta : R(\theta) \leq \epsilon\})$ and $\bar{\epsilon}_{global} = \inf\left\{\epsilon : \epsilon \geq \frac{1}{\rho n}\ln\frac{1}{\pi(\epsilon)}\right\}$.

Proof: For the first inequality, we take $v = 1$ in Theorem 4.3, and let $\alpha \to 0$. For the second inequality, we simply note from the definition of $\bar{\epsilon}_{global}$ that $\inf_\epsilon[\rho n\epsilon - \ln\pi(\epsilon)] \leq 2\rho n\bar{\epsilon}_{global}$. ■

Corollary 4.2 Consider the Gibbs algorithm (5). If (10) holds, then

$$\begin{aligned} &\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \leq \frac{1}{(1-\alpha)\beta_\rho \rho n} \inf_{u_1\leq u_2} \\ &\quad \ln\left[\sup_{\epsilon\in[u_1,u_2]} \frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)} + \exp(\frac{\rho n u_1}{\alpha\beta_\rho}) + \frac{e^{-\rho n u_2/2}}{\pi(u_2/2)} \right] \\ \leq&\frac{\bar{\epsilon}_{local}}{(1-\alpha)\beta_\rho}, \end{aligned}$$

where $\pi(\epsilon) = \pi(\{\theta : R(\theta) \leq \epsilon\})$, and

$$\begin{aligned} \bar{\epsilon}_{local} =& \frac{2}{\rho n} + \inf\Bigg\{ \frac{\epsilon}{\alpha\beta_\rho} : \epsilon \geq \\ &\qquad \sup_{\epsilon'\in[\epsilon,2u]} \frac{\alpha\beta_\rho}{\rho n} \ln\left[\frac{\pi(\epsilon'/(\alpha\beta_\rho))}{\pi(\epsilon')} + \frac{e^{-\rho n u}}{\pi(u)} \right] \Bigg\} . \end{aligned}$$

Proof: For the first inequality, we simply take $u = u_2$ and $v = \pi(u_1/(\alpha\beta_\rho))$ in Theorem 4.3, and use the following bounds

$$\begin{aligned} &\sup_{\epsilon\leq u} \frac{\max(0, \pi(\epsilon/(\alpha\beta_\rho)) - v)}{\pi(\epsilon)} \leq \sup_{\epsilon\in[u_1,u_2]} \frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)}, \\ &\frac{v}{\sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})} \leq \frac{v}{v e^{-\rho n u_1/(\alpha\beta_\rho)}} = \exp(\frac{\rho n u_1}{\alpha\beta_\rho}), \\ &\frac{(1-v)\exp(-\rho n u)}{\sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})} \leq \frac{e^{-\rho n u_2}}{\pi(u_2/2)e^{-\rho n u_2/2}} = \frac{e^{-\rho n u_2/2}}{\pi(u_2/2)}. \end{aligned}$$

For the second inequality, we let $u_2 = 2u$ and $u_1/(\alpha\beta_\rho) = \bar{\epsilon}_{local} - \ln 2/(\rho n)$. Then by the definition of $\bar{\epsilon}_{local}$, we have

$$\begin{aligned} &\sup_{\epsilon\in[u_1,u_2]} \frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)} + \exp(\frac{\rho n u_1}{\alpha\beta_\rho}) + \frac{e^{-\rho n u_2/2}}{\pi(u_2/2)} \\ \leq& 2\exp(\frac{\rho n u_1}{\alpha\beta_\rho}) = \exp(\rho n\bar{\epsilon}_{local}). \end{aligned}$$

This gives the second inequality. ■

C. Proofs for Section V

Proposition 5.1 If there exists a constant $M_{\mathcal{G}} \geq 2/3$ such that $\mathbf{E}_{Z_1}\ell_\theta(Z_1)^3 \leq M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2$. Then $\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \leq \frac{8M_{\mathcal{G}}}{3}\mathbf{E}_{Z_1}\ell_\theta(Z_1)$.

Proof: Consider an arbitrary $\theta \in \mathcal{G}$. Note that $\forall\beta \to 0^+$:

$$\mathbf{E}_{Z_1}\ell_{\theta_{\mathcal{G}}+\beta(\theta-\theta_{\mathcal{G}})}(Z_1) = -\beta\mathbf{E}_{Z_1}\frac{\theta(Z_1) - \theta_{\mathcal{G}}(Z_1)}{\theta_{\mathcal{G}}(Z_1)} + O(\beta^2).$$

Let $\beta \to 0$, then from the convexity of $\mathcal{G}$ and the optimality of $\theta_{\mathcal{G}}$, we have

$$\mathbf{E}_{Z_1}\frac{\theta(Z_1)-\theta_{\mathcal{G}}(Z_1)}{\theta_{\mathcal{G}}(Z_1)} \le 0.$$

This implies that $\forall \rho \in (0,1)$:

$$\mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)} \le \mathbf{E}^{\rho}_{Z_1} e^{-\ell_\theta(Z_1)} = \mathbf{E}^{\rho}_{Z_1}\frac{\theta(Z_1)}{\theta_{\mathcal{G}}(Z_1)} \le 1.$$

We obtain from $e^{-x} \ge 1 - x + x^2/2 - x^3/6$:

$$\begin{aligned}1 \ge &\mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)} \\ \ge &1 - \rho\mathbf{E}_{Z_1}\ell_\theta(Z_1) + \rho^2\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2/2 \\ &- \rho^3\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 M_{\mathcal{G}}/6.\end{aligned}$$

Now let $\rho = 3/(2M_{\mathcal{G}})$ and rearrange, we obtain the desired bound. ■

Proposition 5.2 Let

$$M_{\mathcal{G}} = \sup_{X_1,\theta\in\mathcal{G}} \mathbf{E}_{Y_1|X_1}(\theta(X_1)-Y_1)^2.$$

Then $\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \le 4M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1)$.

Proof: Note that $\forall \beta \in (0,1)$:

$$\begin{aligned}\mathbf{E}_{Z_1}\ell_{\theta_{\mathcal{G}}+\beta(\theta-\theta_{\mathcal{G}})}(Z_1) = &\beta^2\mathbf{E}_{X_1}(\theta(X_1)-\theta_{\mathcal{G}}(X_1))^2 \\ &+ 2\beta\mathbf{E}_{Z_1}(\theta(X_1)-\theta_{\mathcal{G}}(X_1))(\theta_{\mathcal{G}}(X_1)-Y_1).\end{aligned}$$

Let $\beta \to 0$, then from the convexity of $\mathcal{G}$ and the optimality of $\theta_{\mathcal{G}}$, we have

$$\mathbf{E}_{Z_1}(\theta(X_1)-\theta_{\mathcal{G}}(X_1))(\theta_{\mathcal{G}}(X_1)-Y_1) \ge 0.$$

Now let $\beta = 1$, we obtain from the above inequality that

$$\mathbf{E}_{Z_1}\ell_\theta(Z_1) \ge \mathbf{E}_{X_1}(\theta(X_1)-\theta_{\mathcal{G}}(X_1))^2.$$

Therefore we have

$$\begin{aligned}&\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 = \mathbf{E}_{Z_1}(\theta(X_1)-\theta_{\mathcal{G}}(X_1))^2 \\ &\qquad\qquad \cdot [(\theta(X_1)-Y_1)+(\theta_{\mathcal{G}}(X_1)-Y_1)]^2 \\ \le &4M\mathbf{E}_{Z_1}\ell_\theta(Z_1).\end{aligned}$$

■

Proposition 5.3 Let $A_{\mathcal{G}} = \sup_{X_1,\theta\in\mathcal{G}}|\theta(X_1)-\theta_{\mathcal{G}}(X_1)|$ and $\sup_{X_1,\theta\in\mathcal{G}}\mathbf{E}_{Y_1|X_1}|\theta(X_1)-Y_1|^m \le m!B_{\mathcal{G}}^{m-2}M_{\mathcal{G}}$ for $m \ge 2$. Then we have:

$$\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \le m!(2A_{\mathcal{G}}B_{\mathcal{G}})^{m-2}4M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1).$$

Proof: From the proof of Proposition 5.2, we know

$$\mathbf{E}_{Z_1}\ell_\theta(Z_1) \ge \mathbf{E}_{X_1}(\theta(X_1)-\theta_{\mathcal{G}}(X_1))^2.$$

Therefore $\forall m \ge 2$, we have

$$\begin{aligned}&\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \\ \le &\mathbf{E}_{X_1}[|\theta(X_1)-\theta_{\mathcal{G}}(X_1)|^m \\ &\qquad \mathbf{E}_{Y_1|X_1}|(\theta(X_1)-Y_1)+(\theta_{\mathcal{G}}(X_1)-Y_1)|^m] \\ \le &m!(2A_{\mathcal{G}}B_{\mathcal{G}})^{m-2}4M_{\mathcal{G}}\mathbf{E}_{X_1}|\theta(X_1)-\theta_{\mathcal{G}}(X_1)|^2 \\ \le &m!(2A_{\mathcal{G}}B_{\mathcal{G}})^{m-2}4M_{\mathcal{G}}\mathbf{E}_{Z_1}\ell_\theta(Z_1).\end{aligned}$$

This gives the desired bound. ■

Corollary 5.1 Assume that there exists a function $y_0(X)$ such that

- For all X_1, the random variable $|Y_1 - y_0(X_1)|$, conditioned on X_1, is dominated by the absolute value of a zero-mean Gaussian random variable with standard deviation σ.
- $\exists$ constant $b > 0$ such that $\sup_{X_1}|y_0(X) - \theta(X_1)| \le b$.

If we also choose A such that $A \ge \sup_{X_1,\theta\in\mathcal{G}}|\theta(X_1) - \theta_{\mathcal{G}}(X_1)|$, then (10) holds with $\beta_\rho = 1 - 4\rho(b+\sigma)^2/(1 - 2A(b+\sigma)\rho) > 0$.

Proof: Let ξ be the zero-mean Gaussian random variable with unit variance. Using the simple recursion

$$\begin{aligned}&\int_0^\infty \xi^m e^{-\xi^2/2}d\xi \\ = &\xi^{m-1}e^{-\xi^2/2}|_\infty^0 + (m-1)\int_0^\infty \xi^{m-2}e^{-\xi^2/2}d\xi,\end{aligned}$$

it is not difficult to check that $\forall m \ge 2$: $\mathbf{E}|\xi|^m \le m!$. We thus have

$$\begin{aligned}&(\mathbf{E}_{Y_1|X_1}|Y_1-\theta(X_1)|^m)^{1/m} \\ \le &|y_0(X_1)-\theta(X_1)| + (\mathbf{E}_{Y_1|X_1}|Y_1-y_0(X_1)|^m)^{1/m} \\ \le &b + \sigma(m!)^{1/m} \le (b+\sigma)(m!)^{1/m}.\end{aligned}$$

Therefore Proposition 5.3 holds with $A_{\mathcal{G}} = A$, $B_{\mathcal{G}} = b + \sigma$ and $M_{\mathcal{G}} = (b+\sigma)^2$. Now we can just apply Theorem 5.2. ■

D. Proofs for Section VI

Theorem 6.1 Consider an arbitrary randomized estimator $\hat{\pi}(Z)$ on $B' \subset \mathcal{G}$, where Z consists of n independent samples $Z = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ from some underlying distribution P_θ, then for all non-negative function $R_\theta(\theta')$, we have for all B that contains θ:

$$\begin{aligned}&\mathbf{E}_{\theta\sim\pi_B}\mathbf{E}_{Z\sim P_\theta^n}\mathbf{E}_{\hat{\theta}\sim\hat{\pi}(Z)}R_\theta(\hat{\theta}) \ge 0.5 \\ &\sup\left\{\epsilon : \inf_{\theta'\in B'}\ln\frac{1}{\pi_B(\{\theta : R_\theta(\theta') < \epsilon\})} \ge 2n\Delta_B + \ln 4\right\},\end{aligned}$$

where $\Delta_B = \mathbf{E}_{\theta\sim\pi_B}\mathbf{E}_{\theta'\sim\pi_B} D_{KL}(P_\theta(Z_1)||P_{\theta'}(Z_1))$.

Proof: The joint distribution of (θ, Z) is given by $\prod_{i=1}^n dP_\theta(Z_i)d\pi_B(\theta)$. Denote by $I(\theta, Z)$ the mutual information between θ and Z. Now let Z' be a random variable independent of θ and with the same marginal of Z, then by definition, the mutual information can be regarded as the KL-divergence between the joint distributions of (θ, Z) and (θ, Z'), which we write (with a slight abuse of notation) as:

$$I(\theta, Z) = D_{KL}((\theta, Z)||(\theta, Z')).$$

Now consider an arbitrary randomized estimator $\hat{\pi}(Z)$, and let $\hat{\theta} \in B'$ be a random variable that is drawn from the distribution $\hat{\pi}(Z)$. By the data processing theorem for KL-divergence (that is, processing does not increase KL-divergence), with input $(\theta, Z) \in \mathcal{G}\times\mathcal{Z}$ and (random) binary output $\mathbf{1}(R_\theta(\hat{\theta}(Z)) \le \epsilon)$, we obtain

$$\begin{aligned}
&D_{KL}(\mathbf{1}(R_\theta(\hat{\theta}(Z)) \le \epsilon)||\mathbf{1}(R_\theta(\hat{\theta}(Z')) \le \epsilon))\\
\le&D_{KL}((\theta, Z)||(\theta, Z')) = I(\theta, Z)\\
=&\mathbf{E}_{\theta_1\sim\pi_B}\mathbf{E}_{Z\sim P^n_{\theta_1}} \ln \frac{dP^n_{\theta_1}(Z)}{\mathbf{E}_{\theta_2\sim\pi_B} dP^n_{\theta_2}(Z)}\\
\le&\mathbf{E}_{\theta_1\sim\pi_B}\mathbf{E}_{Z\sim dP^n_{\theta_1}}\mathbf{E}_{\theta_2\sim\pi_B} \ln \frac{dP^n_{\theta_1}(Z)}{dP^n_{\theta_2}(Z)}\\
=&n\mathbf{E}_{\theta_1\sim\pi_B}\mathbf{E}_{\theta_2\sim\pi_B} D_{KL}(P_{\theta_1}||P_{\theta_2}) = n\Delta_B.
\end{aligned}$$

The second inequality is a consequence of Jensen's inequality and the concavity of logarithm.

Now let $p_1 = \mathbf{P}(R_\theta(\hat{\theta}(Z)) \le \epsilon)$ and $p_2 = \mathbf{P}(R_\theta(\hat{\theta}(Z')) \le \epsilon)$, then the above inequality can be rewritten as:

$$D_{KL}(p_1||p_2) = p_1 \ln \frac{p_1}{p_2} + (1 - p_1) \ln \frac{1 - p_1}{1 - p_2} \le n\Delta_B.$$

Since $\hat{\theta}(Z')$ is independent of θ, we have

$$p_2 \le \sup_{\theta'\in B'} \pi_B(\{\theta : R_\theta(\theta') \le \epsilon\}).$$

Now we consider any ϵ such that $\sup_{\theta'\in B'} \pi_B(\{\theta : R_\theta(\theta') \le \epsilon\}) \le 0.25e^{-2n\Delta_B}$. This implies that $p_2 \le 0.25e^{-2n\Delta_B} \le 0.25$.

We now show that in this case, $p_1 \le 1/2$. Since $D_{KL}(p_1||p_2)$ is increasing in $[p_2, 1]$, we only need to show that $D_{KL}(0.5||p_2) \ge n\Delta_B$. This easily follows from the inequality

$$D_{KL}(0.5||p_2) \ge 0.5 \ln \frac{0.5}{p_2} + 0.5 \ln \frac{0.5}{1} \ge n\Delta_B.$$

Now, we have shown that $p_1 \le 0.5$, which implies that $\mathbf{P}(R_\theta(\hat{\theta}(Z)) \ge \epsilon) \ge 0.5$. Therefore we have $\mathbf{E}_{\theta\sim\pi_B}\mathbf{E}_{Z\sim P^n_\theta}\mathbf{E}_{\hat{\theta}\sim\hat{\pi}(Z)} R_\theta(\hat{\theta}) \ge 0.5\epsilon$. ■

Remark 7.2: The data-processing theorem (for divergence) states that for random variables A and B, and a (possibly random) processing function h, the inequality $D_{KL}(A||B) \ge D_{KL}(h(A)||h(B))$ holds. The proof is not difficult. For a random variable C, we use P_C to denote the corresponding probability measure. Then it is easy to check

$$\begin{aligned}
&D_{KL}((A_1, A_2)||(B_1, B_2))\\
=&D_{KL}(P_{A_1}||P_{A_2}) + \mathbf{E}_{\xi\sim P_{A_1}} D_{KL}(P_{A_2}(\cdot|\xi)||P_{B_2}(\cdot|\xi)).
\end{aligned}$$

Now let $A_1 = A$, $A_2 = h(A)$, $B_1 = B$, and $B_2 = h(B)$, then $P_{A_2}(\cdot|\xi) = P_{B_2}(\cdot|\xi)$, which implies that $K(A||B) = K((A, h(A))||(B, h(B)))$. Similarly, we can let $A_2 = A$, $A_1 = h(A)$, $B_2 = B$, and $B_1 = h(B)$, then we obtain $K((h(A), A)||(h(B), B)) \ge D_{KL}((h(A))||(h(B)))$.

Remark 7.3: The author learned the idea of deriving lower bounds using data-processing theorem from [13], where a generalization of Fano's inequality was obtained. The authors in that paper attributed this idea back to Blahut [14]. In Theorem 6.1, this technique is used to directly obtain a general lower bound without going through the much more specialized Fano's inequality, as has been typically done in the minimax literature (see [9] and discussions therein). This technique also enables us to obtain lower bounds with respect to an arbitrary prior (which is impossible using the Fano's inequality approach).

VIII. Discussions

In this paper, we established upper and lower bounds for some statistical estimation problems. Our upper bound analysis is based on a simple information theoretical inequality which we call the information exponential inequality. Consequences of this new inequality were explored. In particular, we showed that the inequality naturally leads to a (randomized) statistical estimation method which we call information complexity minimization, which depends on a pre-specified prior. The resulting upper bounds rely on the local prior decay rate in some small ball around the truth. Examples are provided for some important statistical estimation problems such as conditional density estimation and regression.

Moreover, based on novel applications of some well known information theoretical inequalities, we are able to obtain lower bounds that have similar forms to the upper bounds. For some problems (such as density estimation and regression), the upper and lower bounds match under mild conditions. This suggests that our upper bound and lower bound analyses are relatively tight.

Our work can be regarded as an extension of the standard minimax framework since we allow the performance of the estimator to vary for different underlying distributions, according to the pre-defined prior. Our analysis can also be applied in the minimax framework by employing a uniform prior on an ϵ-net of the underlying space. The framework we study here is closely related to the concept of adaption in the statistical literature. At the conceptual level, both seek to find locally near optimal estimators around any possible true underlying distribution within the class. However, we believe our treatment is cleaner and more general since for many problems, the simple IRM estimator with appropriately chosen prior is automatically locally optimal around any possible truth. The traditional approach to adaptation usually requires us to design specialized estimators that are much less general.

Our results also indicate that if we pick a reasonably well-behaved prior (locally relatively smooth around the truth), then local statistical complexity is fully determined by the local prior structure. Specifically, the rate of convergence is determined by the decay rate of the local entropy (local prior probability of a ball at some scale relative to a ball at a scale larger by a constant factor). It is possible to obtain a near optimal estimator using the IRM method so that the performance is no more than a constant time worse than the best local estimator around every possible underlying truth.

Related to the optimality of IRM with respect to a general prior, it was pointed out in Section III-A that there is no need for model selection in our framework. Similar to the Bayesian method which does not require prior selection (at least in principle), in our framework, a prior mixture instead of prior selection can be used for the IRM method. We believe that the ability of incorporating model selection into prior knowledge is an important conceptual advantage of our analysis.

This paper also implies that for some problems, the IRM method can be better behaved than (possibly penalized) empirical risk minimization that picks an estimator to minimize the (penalized) empirical risk. In particular, for density estimation and regression, the IRM method can achieve the best possible convergence rate under relatively mild assumptions on the prior structure. However, it is known that for some non-parametric problems, empirical risk minimization can lead to suboptimal convergence rate if the covering number grows too rapidly (or in our case, prior decays too rapidly) when ϵ (the size of the covering ball) decreases [15].

APPENDIX I
LOGARITHMIC MOMENT GENERATING FUNCTION

Let W be a random variable. In the following, we study properties of its logarithmic generating function $\Lambda_W(\rho) = \ln \mathbf{E} \exp(\rho W)$ as a function of $\rho > 0$, so that we can gain additional insights into the behavior of the left-hand side of Theorem 4.1. Although the properties presented below are not necessarily new, it is useful to list them for reference. In the following, we also use $\mathbf{Var}W = \mathbf{E}(W - \mathbf{E}W)^2$ to denote the variance of W.

Proposition 1.1: The following properties of $\Lambda_W(\rho)$ hold for $\rho > 0$:

1) $\Lambda_W(\rho) \geq \rho \mathbf{E}W$.
2) $\frac{1}{\rho}\Lambda_W(\rho)$ is an increasing function of ρ.
3) $\Lambda'_W(\rho) = \mathbf{E}W e^{\rho W - \Lambda_W(\rho)}$ is non-decreasing.
4) $\Lambda_W(\rho) \leq \rho \Lambda'_W(\rho)$.
5) $\Lambda''_W(\rho) = \mathbf{E}(W - \Lambda'_W(\rho))^2 e^{\rho W - \Lambda_W(\rho)} \leq \mathbf{E}(W - \mathbf{E}W)^2 e^{\rho W - \Lambda_W(\rho)} \leq \mathbf{E}(W - \mathbf{E}W)^2 e^{\rho(W - \mathbf{E}W)}$.
6) $\Lambda_W(\rho) \leq \rho \mathbf{E}W + \frac{\rho^2}{2}\mathbf{Var}W + \frac{\rho^3}{2}\mathbf{E}_{W \geq \mathbf{E}W}(W - \mathbf{E}W)^3 e^{\rho(W - \mathbf{E}W)}$.
7) $\Lambda_{aW+b}(\rho) \rightarrow \Lambda_W(a\rho) + b$.

Proof:

1) Using Jensen's inequality, we have $\mathbf{E}e^{\rho W} \geq \exp(\rho \mathbf{E}W)$.
2) Again using Jensen's inequality, for $0 < \rho' < \rho$, we have $\mathbf{E}e^{\rho W} \geq (\mathbf{E}\exp(\rho' W))^{\rho/\rho'}$.
3) This is the same as $\Lambda_W(\rho)$ is convex, or $\Lambda''_W(\rho)$ is non-negative (see below).
4) Using Taylor expansion, we have $\Lambda_W(\rho) = \rho \Lambda'_W(\rho')$ where $\rho' \in (0, \rho)$. Since $\Lambda'_W(\rho') \leq \Lambda_W(\rho)$, we obtain the inequality.
5) The first equality is by direct calculation. The second inequality follows from the calculation that $\mathbf{E}(W - \Lambda'_W(\rho))^2 e^{\rho W - \Lambda_W(\rho)} = \mathbf{E}(W - \mathbf{E}W)^2 e^{\rho W - \Lambda_W(\rho)} - (\mathbf{E}W - \Lambda'_W(\rho))^2$. The third inequality uses $-\Lambda_W(\rho) \leq -\rho \mathbf{E}W$.
6) Using Taylor expansion, we have $\Lambda_W(\rho) = \Lambda_W(0) + \rho \Lambda'_W(0) + \frac{\rho^2}{2}\Lambda''_W(\rho')$ where $\rho' \in (0, \rho)$, $\Lambda_W(0) = 0$, and $\Lambda'_W(0) = \mathbf{E}W$. Now, we obtain the desired bound by using the following estimate: $\Lambda''_W(\rho') \leq \mathbf{E}(W - \mathbf{E}W)^2 e^{\rho'(W - \mathbf{E}W)} \leq \mathbf{E}(W - \mathbf{E}W)^2 \max(e^{\rho(W - \mathbf{E}W)}, 1)$.
7) Direct calculation.

■

Remark 1.1: The bound in 6 indicates that if $\Lambda_W(\rho_0) < \infty$ for some $\rho_0 > 0$, then when $\rho \rightarrow 0^+$, $\Lambda_W(\rho) = \rho \mathbf{E}W + \frac{\rho^2}{2}\mathbf{Var}W + O(\rho^3)$. In general, we have the property that $\Lambda_W(\rho) = \rho \mathbf{E}W + O(\rho^2)$, which is useful in our analysis.

Proposition 1.1 gives the behavior of $\Lambda_W(\rho)$ for general random variables. In the following, for reference purposes, we also list some examples of $\Lambda_W(\rho)$ for more concrete random variables.

Proposition 1.2: Given a random variable W, we have the following bounds for its logarithmic moment generating function $\Lambda_W(\rho)$.

1) If W is a Gaussian random variable, then $\Lambda_W(\rho) = \rho\mathbf{E}W + \frac{\rho^2}{2}\mathbf{Var}W$.
2) If W is a bounded variable in $[0,1]$, then $\Lambda_W(\rho) \leq \ln(1+(e^\rho-1)\mathbf{E}W) \leq \rho\mathbf{E}W + \frac{\rho^2}{8}$.
3) If W satisfies the moment condition: $\mathbf{E}W^m \leq m!M^{m-2}b$ for integer $m \geq 3$, then $\forall\rho \in [0,1/M)$, $\Lambda_W(\rho) \leq \rho\mathbf{E}W + \frac{\rho^2}{2}\mathbf{E}W^2 + \rho^3 bM(1-\rho M)^{-1}$.
4) If $W \leq 1$, then $\forall\rho \geq 0$: $\Lambda_W(\rho) \leq 1 + \rho\mathbf{E}W + (e^\rho - \rho - 1)\mathbf{E}W^2$.

Proof:

1) Direct calculation.
2) The first inequality follows from Jensen's inequality $e^{\rho W} \leq (1-W)e^0 + We^\rho$, and is the basis of Hoeffding's inequality in [16]. The second inequality can be easily checked using a Taylor expansion (for example, see [16]). The equality holds for the first inequality when $W \in \{0,1\}$ is Bernoulli.
3) This simple bound is the basis of a slightly more refined Bernstein's inequality under moment conditions. We have $\mathbf{E}e^{\rho W} = 1 + \rho\mathbf{E}W + \frac{\rho^2}{2}\mathbf{E}W^2 + \sum_{m=3}^\infty \frac{\rho^m}{m!}\mathbf{E}W^m \leq 1 + \rho\mathbf{E}W + \frac{\rho^2}{2}\mathbf{E}W^2 + \rho^3 bM(1-\rho M)^{-1}$. Now, use the fact that $\ln x \leq x - 1$.
4) Using $\ln x \leq x - 1$, we have $\Lambda_W(\rho) \leq \rho\mathbf{E}W + \rho^2\mathbf{E}\frac{e^{\rho W}-\rho W-1}{(\rho W)^2}W^2 \leq \rho\mathbf{E}W + \rho^2\mathbf{E}\frac{e^\rho-\rho-1}{\rho^2}W^2$, where the second inequality follows from the fact that the function $(e^x - x - 1)/x^2$ is non-decreasing and $\rho W \leq \rho$.

■

References

[1] T. Zhang, "Learning bounds for a generalized family of Bayesian posterior distributions," in *NIPS 03*, 2004.

[2] D. McAllester, "PAC-Bayesian model averaging," in *COLT'99*, 1999, pp. 164–170.

[3] O. Catoni, *Statistical Learning Theory and Stochastic Optimization*, ser. Lecture Notes in Mathematics. Ecole d'Ete de Probabilites de Saint-Flour XXXI, 2001, vol. 1851.

[4] ——, "A PAC-Bayesian approach to adaptive classification," Laboratoire de Probabilites et Modeles Aleatoires, Tech. Rep., 2003, http://www.proba.jussieu.fr/users/catoni/homepage/classif.pdf.

[5] J.-Y. Audibert, "PAC-Bayesian statistical learning theory," Ph.D. dissertation, Universit Paris VI, 2004.

[6] T. Zhang, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, 2006, to appear.

[7] S. van de Geer, *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

[8] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes*, ser. Springer Series in Statistics. New York: Springer-Verlag, 1996.

[9] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *The Annals of Statistics*, vol. 27, pp. 1564–1599, 1999.

[10] A. Barron and T. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. 37, pp. 1034–1054, 1991.

[11] R. Meir and T. Zhang, "Generalization error bounds for Bayesian mixture algorithms," *Journal of Machine Learning Research*, vol. 4, pp. 839–860, 2003.

[12] M. Seeger, "PAC-Bayesian generalization error bounds for Gaussian process classification," *JMLR*, vol. 3, pp. 233–269, 2002.

[13] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Transactions on Information Theory*, vol. 40, pp. 1247–1251, 1994.

[14] R. Blahut, "Information bounds of the Fano-Kullback type," *IEEE Transactions on Information Theory*, vol. 22, pp. 410–421, 1976.

[15] L. Birgé and P. Massart, "Rates of convergence for minimum contrast estimators," *Probab. Theory Related Fields*, vol. 97, no. 1-2, pp. 113–150, 1993.

[16] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963.

PLACE
PHOTO
HERE

Tong Zhang Tong Zhang received a B.A. in mathematics and computer science from Cornell University in 1994 and a Ph.D. in computer Science from Stanford University in 1998. After being a research staff member of IBM T.J. Watson Research Center in Yorktown Heights, New York, he joined Yahoo in 2005. His research interests include machine learning, numerical algorithms, and their applications in text processing.