The Annals of Statistics
1997, Vol. 25, No. 6, 2451–2492

MUTUAL INFORMATION, METRIC ENTROPY AND CUMULATIVE RELATIVE ENTROPY RISK

BY DAVID HAUSSLER[1] AND MANFRED OPPER[2]

University of California at Santa Cruz and University of Würzburg

Assume $\{P_\theta\colon \theta \in \Theta\}$ is a set of probability distributions with a common dominating measure on a complete separable metric space Y. A state $\theta^* \in \Theta$ is chosen by Nature. A statistician obtains n independent observations $Y_1, \ldots, Y_n$ from Y distributed according to P_{θ^*}. For each time t between 1 and n, based on the observations $Y_1, \ldots, Y_{t-1}$, the statistician produces an estimated distribution $\hat{P}_t$ for P_{θ^*} and suffers a loss $L(P_{\theta^*}, \hat{P}_t)$. The cumulative risk for the statistician is the average total loss up to time n. Of special interest in information theory, data compression, mathematical finance, computational learning theory and statistical mechanics is the special case when the loss $L(P_{\theta^*}, \hat{P}_t)$ is the relative entropy between the true distribution P_{θ^*} and the estimated distribution $\hat{P}_t$. Here the cumulative Bayes risk from time 1 to n is the mutual information between the random parameter Θ^* and the observations $Y_1, \ldots, Y_n$.

New bounds on this mutual information are given in terms of the Laplace transform of the Hellinger distance between pairs of distributions indexed by parameters in Θ. From these, bounds on the cumulative minimax risk are given in terms of the metric entropy of Θ with respect to the Hellinger distance. The assumptions required for these bounds are very general and do not depend on the choice of the dominating measure. They apply to both finite- and infinite-dimensional Θ. They apply in some cases where Y is infinite dimensional, in some cases where Y is not compact, in some cases where the distributions are not smooth and in some parametric cases where asymptotic normality of the posterior distribution fails.

1. Introduction. Much of classical statistics has been concerned with the estimation of probability distributions from independent and identically distributed observations drawn according to these distributions. If we let P_{θ^*} denote the true distribution generating the observations and $\hat{P}_t$ the estimated distribution obtained after seeing $t-1$ independent observations, then the success of our statistical procedure can be defined in terms of a loss function that measures the difference between the true distribution P_{θ^*} and the estimated distribution $\hat{P}_t$. One such loss function has proven to be of importance in several fields, including information theory, data compression,

Received January 1996; revised January 1997.
[1]Supported by NSF Grant IRI-9123692.
[2]Supported by a Heisenberg Fellowship of DFG.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62B10, 62C20, 94A29.
Key words and phrases. Mutual information, Hellinger distance, relative entropy, metric entropy, minimax risk, Bayes risk, density estimation, Kullback–Leibler distance.

mathematical finance, computational learning theory and statistical mechanics. This is the relative entropy function. Further, in these fields, special importance is given to the cumulative relative entropy loss suffered in a sequential estimation setting, in which there are n total observations, but these observations arrive one at a time, and at each time t a new, refined estimate $\hat{P}_t$ is made for the unknown true distribution P_{θ^*}, based on the $t-1$ previous observations. This is the setting that we study in this paper.

The average of the cumulative loss over all sequences of n observations generated according to the true distribution is the (cumulative relative entropy) *risk*. For a given family $\{P_\theta\colon\ \theta \in \Theta\}$ of distributions, two types of risk are of interest in statistics. One is the minimax risk, which is the minimum worst-case risk over possible true distributions P_{θ^*}, where $\theta^* \in \Theta$, and the minimum is over all possible sequential estimation strategies. The other is the Bayes risk, which is the minimum average-case risk over possible true distributions P_{θ^*} drawn according to a prior distribution μ on Θ, and the minimum is again over all possible sequential estimation strategies. For cumulative relative entropy loss, the Bayes risk has a fundamental information-theoretic interpretation: it is the mutual information between a random variable representing the choice of the parameter θ^* of the true distribution and the random variable given by the n observations, see [15], [23] and [35]. This provides a beautiful connection between information theory and statistics.

This connection also extends to other fields, as is discussed in [6] and [15]. In data compression, the cumulative relative entropy risk is the *redundancy*, which is the expected excess code length for the best adaptive coding method, as compared to the best coding method that has prior knowledge of the true distribution, see [15], [28] and [42]. The minimax risk is called the "*information*" *channel capacity* in [18], page 184. In mathematical finance and gambling theory, the cumulative relative entropy risk measures the expected reduction in the logarithm of compounded wealth due to lack of knowledge of the true distribution ([7] and [15]). In computational learning theory, this risk is the average additional loss suffered by an adaptive algorithm that predicts each observation before it arrives, based on the previous observations, as compared to an algorithm that makes predictions knowing the true distribution ([31] and [32]). Here we assume that the observation at time t is predicted by the "predictive" probability distribution $\hat{P}_t$, formed by the adaptive algorithm using the previous $t-1$ observations, and that when this tth observation arrives, the loss is the negative logarithm of its probability under $\hat{P}_t$. Finally, in statistical mechanics, the Bayes risk can be related to the free energy ([43] and [44]).

In this paper, we provide upper and lower bounds on the Bayes risk for cumulative relative entropy loss in the form of Laplace integrals of the Hellinger distance between pairs of distributions in $\{P_\theta\colon\ \theta \in \Theta\}$. We illustrate these bounds in a number of special cases, then use them to characterize the asymptotic growth rate of the minimax risk in terms of the metric entropy of $\{P_\theta\colon\ \theta \in \Theta\}$ under the Hellinger distance. The methods used here have the

advantage of simplicity, with proofs amounting to little more than simple applications of Jensen's inequality. The results are also quite general. The bounds apply to both finite- and infinite-dimensional Θ. They apply in some cases where the space of observations is infinite dimensional, in some cases where it is not compact, in some cases where the distributions are not smooth and in some parametric cases where asymptotic normality of the posterior distribution fails. The bounds are also fairly tight. However, in smooth parametric cases, our general bounds are too crude to give the precise estimates of the low-order additive constants that were obtained by Clarke and Barron [15, 16].

The paper is organized as follows. In Sections 2 and 3 we give precise definitions of the risks that we evaluate and discuss the conditions required for our bounds to hold. Here we also compare our bounds to those obtained previously by other authors. The bounds are given in Section 4, followed by examples in Sections 5 and 6 showing how they can be applied. Then in Section 7 we give the characterization of the minimax risk. Finally, we discuss some possible further work in Section 8.

2. Basic definitions, notation and assumptions. The following notation and assumptions will be used throughout the paper.

Let Y be a complete separable metric space. All probability distributions on Y discussed in this paper are assumed to be defined on the σ-algebra of Borel sets of Y. Let Θ be a set, and, for each $\theta \in \Theta$, let P_θ be a probability distribution on Y. We assume that, for any $\theta \neq \theta^* \in \Theta$, the distributions associated with θ and θ^* are distinct in the sense that there is a Borel set $S \subset Y$ such that $P_\theta(S) \neq P_{\theta^*}(S)$. In addition, we assume there is a fixed σ-finite measure ν on Y that dominates P_θ for all $\theta \in \Theta$ [i.e., for any Borel set $S \subseteq Y$, $\nu(S) = 0$ implies $P_\theta(S) = 0$]. We will also make (implicitly) the assumption that any other distribution Q on Y mentioned in the following results is also dominated by ν. None of our results depends on the choice of the dominating measure ν. Hence, for any distribution Q, the Radon-Nikodym derivative $dQ/d\nu$ will be abbreviated simply as dQ, following the convention in Le Cam's text [41]. Furthermore, all integrals in the following results are assumed, without specific notation, to be taken with respect to the measure ν, unless otherwise indicated. Thus, for a function f on Y and a distribution Q on Y, the expectation of f is denoted by

$$\int f\,dQ = \int_Y \frac{dQ}{d\nu}(y) f(y)\,d\nu.$$

Hence, in the special case that Y is countable and ν is the counting measure, for a probability mass function Q on Y,

$$\int f\,dQ = \sum_{y \in Y} Q(y) f(y).$$

We will also need to treat probability distributions over Θ, which we will refer to as *prior distributions*. As each $\theta \in \Theta$ is associated with a distinct distribution P_θ on a complete separable metric space, we can define prior distributions on Θ with respect to the Borel sets of the topology of weak convergence of the P_θ measures. We assume that the set $\{P_\theta : \theta \in \Theta\}$ is itself measurable w.r.t. this topology. All prior distributions μ on Θ used in this paper are assumed to be Borel distributions of this type, and suprema over priors are also assumed to be only with respect to Borel distributions of this type. Further discussion of these issues can be found in the appendix of [21].

Finally, for integer- or real-valued functions f and g, we say $f \sim g$ if $\lim_{n\to\infty} f(n)/g(n) = 1$, and $f \asymp g$ if $\liminf_{n\to\infty} f(n)/g(n) > 0$ and $\limsup_{n\to\infty} f(n)/g(n) < \infty$. All logarithms are natural logarithms unless otherwise specified. We assume throughout that $0 \log 0 = 0 \log(x/0) = 0$, where x is any nonnegative finite number. We will also employ functions taking values in the extended reals $[-\infty, +\infty]$, and, in particular, use the extended log function obtained by defining $\log 0 = -\infty$ and $\log \infty = \infty$. Expectations over extended real-valued functions are defined whenever they do not take both the value $+\infty$ with positive probability and the value $-\infty$ with positive probability. The expectation is $+\infty$ if this value has positive probability and similarly for $-\infty$.

3. Statement of the problem: the game of estimating a probability distribution. We view the problem of estimating a probability distribution from the set of distributions $\{P_\theta : \theta \in \Theta\}$ as a game in which Nature plays against the statistician. First, Nature picks $\theta^* \in \Theta$. We refer to θ^* as the (*true*) *state of Nature*. Then, for some $n \geq 1$, a sequence $Y^n = Y_1, \ldots, Y_n$ of i.i.d. random variables is observed, each variable distributed according to P_{θ^*}. The particular sequence of values observed for these random variables is denoted by $y^n = y_1, \ldots, y_n$. For each time t between 1 and n, the statistician forms an estimate $\hat{P}_t = \hat{P}_t(y_t | y^{t-1})$ for the unknown distribution P_{θ^*}, based on the values $y^{t-1} = y_1, \ldots, y_{t-1}$. In particular, for every t and every y^{t-1}, $\hat{P}_t$ is a distribution over Y called the *predictive distribution at time t*, and the set of all such predictive distributions, for all t and y^{t-1}, is called the (*predictive*) *strategy* of the statistician, denoted simply as $\hat{P}$. Note that in this formulation the statistician does not estimate the parameter θ^* itself, but rather the distribution it represents. This allows the statistician, if necessary, to use predictive distributions that are not in the set $\{P_\theta : \theta \in \Theta\}$.

The statistician suffers some loss by using a predictive distribution in place of the true distribution P_{θ^*}. We define this loss as the *KL-divergence* or *relative entropy* between the true distribution and the predictive distribution. For general distributions P and Q, the relative entropy between P and Q is defined by

$$D_{KL}(P \| Q) = \int dP \log \frac{dP}{dQ}.$$

If the statistician uses the strategy $\hat{P}$, then the *risk* (*to the statistician*) *at time* t, *when* θ^* *is the state of Nature*, is given by the average loss

$$r_{t,\hat{P}}(\theta^*) = \int_{Y^{t-1}} dP_{\theta^*}^{t-1} D_{KL}\left(P_{\theta^*} \| \hat{P}_t\right).$$

The *cumulative risk for the first n observations* is

$$R_{n,\hat{P}}(\theta^*) = \sum_{t=1}^{n} r_{t,\hat{P}}(\theta^*).$$

This paper examines only cumulative risk, which is henceforth referred to simply as *risk*, while the risk at time t is referred to as the *instantaneous risk*. The (cumulative) risk has a particularly simple interpretation. For any strategy $\hat{P}$, define the distribution $\hat{P}$ on Y^n by

$$\hat{P}(y^n) = \prod_{t=1}^{n} \hat{P}_t(y_t | y^{t-1}).$$

In this way, we can identify prediction strategies with joint distributions on $Y_1, \ldots, Y_n$. Then

$$\begin{aligned} R_{n,\hat{P}}(\theta^*) &= \sum_{t=1}^{n} \int_{Y^{t-1}} dP_{\theta^*}^{t-1}(y^{t-1}) \int_{Y} dP_{\theta^*}(y_t) \log \frac{dP_{\theta^*}(y_t)}{d\hat{P}_t(y_t | y^{t-1})} \\ &= D_{KL}\left(P_{\theta^*}^n \| \hat{P}\right) \end{aligned} \tag{1}$$

by the chain rule for relative entropy (see, e.g., [18], page 23).

Of course, the statistician seeks a strategy that minimizes risk. One approach assumes that Nature is a strategic adversary and hence selects the worst case θ^* for any particular strategy of the statistician. In this case, the best strategy for the statistician is one that minimizes the worst-case risk, and the value of the game is the *minimax risk*

$$R_n^{\text{minimax}} = \inf_{\hat{P}} \sup_{\theta^* \in \Theta} R_{n,\hat{P}}(\theta^*).$$

A strategy $\hat{P}$ that achieves this minimax value is called a *minimax strategy*.

The other approach is the Bayesian approach, where one seeks to minimize the average risk. Here we might imagine that Nature chooses θ^* at random according to a prior probability distribution μ on Θ. Then the statistician seeks to minimize the average risk (according to μ), and the value of the game is the *Bayes risk*

$$R_{n,\mu}^{\text{Bayes}} = \inf_{\hat{P}} \int_{\Theta} d\mu(\theta^*) R_{n,\hat{P}}(\theta^*).$$

A strategy $\hat{P}$ that achieves this value is called a *Bayes strategy*.

In the Bayesian approach, there are two random variables, Θ^*, giving the choice of the state of Nature, and $Y^n = Y_1, \ldots, Y_n$, giving the sequence of observations. Their joint distribution defines the behavior of Nature. The

marginal distribution of Y^n, defined by

$$M_{n,\mu}(y^n) = \int_\Theta d\mu(\theta^*)\, P^n_{\theta^*}(y^n),$$

is of particular importance here. Breaking $M_{n,\mu}$ down into a product of conditional distributions, we can write

$$M_{n,\mu}(y^n) = \prod_{t=1}^{n} P^{\text{Bayes}}_{t,\mu}\left(y_t | y^{t-1}\right),$$

where

$$P^{\text{Bayes}}_{t,\mu}\left(y_t | y^{t-1}\right) = \frac{M_{t,\mu}(y^t)}{M_{t-1,\mu}(y^{t-1})}.$$

The distributions $P^{\text{Bayes}}_{t,\mu}$ are called *predictive posterior distributions*. These form a Bayes strategy for relative entropy loss, which we call P^{Bayes}_{μ}. To see this, note that by (1) the difference between the average risk for an arbitrary strategy $\hat{P}$ and the strategy P^{Bayes}_{μ} is

$$\begin{aligned}
&\int_\Theta d\mu(\theta^*)\left(D_{KL}\left(P^n_{\theta^*}\|\hat{P}\right) - D_{KL}\left(P^n_{\theta^*}\|M_{n,\mu}\right)\right) \\
&\quad = \int_\Theta d\mu(\theta^*) \int_{Y^n} dP^n_{\theta^*}\left(\log\frac{dP^n_{\theta^*}}{d\hat{P}} - \log\frac{dP^n_{\theta^*}}{dM_{n,\mu}}\right) \\
&\quad = \int_{Y^n} dM_{n,\mu} \log\frac{dM_{n,\mu}}{d\hat{P}} \\
&\quad = D_{KL}\left(M_{n,\mu}\|\hat{P}\right) \geq 0.
\end{aligned}$$

It follows that the Bayes risk for relative entropy loss is given by

$$R^{\text{Bayes}}_{n,\mu} = \int_\Theta d\mu(\theta^*)\, D_{KL}\left(P^n_{\theta^*}\|M_{n,\mu}\right) = I(\Theta^*; Y^n),$$

the *mutual information* between the parameter Θ^* and the observations Y^n. (See [18], page 18, for a general definition and discussion of mutual information.)

It turns out that there is a simple, universal relationship between the Bayes risk $R^{\text{Bayes}}_{n,\mu}$ and the minimax risk R^{minimax}_n. This result can be obtained with limited effort from the general results in an early paper of Le Cam [39]. Special cases of the result were derived by Gallager [25] and Davisson and Leon-Garcia [19], and the general result is given in [30].

THEOREM 1 (Haussler [30]).

$$R^{\text{minimax}}_n = \sup_\mu R^{\text{Bayes}}_{n,\mu},$$

where the supremum is taken over all (Borel) probability measures on the parameter space Θ. *Moreover*,

$$R^{\text{minimax}}_n = \inf_\mu \sup_{\theta^* \in \Theta} R_{n, P^{\text{Bayes}}_\mu}(\theta^*).$$

Several authors have studied the Bayes risk $R_{n,\mu}^{\text{Bayes}}$, or the equivalent mutual information $I(\Theta^*; Y^n)$, for the case of a parametric family of distributions $\{P_\theta\colon\ \theta \in \Theta\}$. Early work by Ibragimov and Hasminskii [35] showed that $I(\Theta^*; Y^n) \sim (D/2)\log n$ when Y is the real line and the conditional distributions P_θ are a smooth family of densities indexed by a real-valued parameter vector θ in a compact set Θ of dimension D, and certain other conditions apply. In this case, they were even able to estimate the lower-order additive terms in this approximation, which involve the Fisher information and the entropy of the prior. Further related results were given by Efroimovich [23] and Clarke [14]. Clarke and Barron [15] gave a detailed analysis, with applications, of the risk of the Bayes strategy as a function of the true state of Nature, discussing the relation of the Bayes risk to the notion of redundancy in information theory and giving applications to hypothesis testing and portfolio selection theory. These results were extended to the Bayes and minimax risk in [16] (see also [6]). Related lower bounds, which are often quoted, were obtained by Rissanen [49], based on certain asymptotic normality assumptions. Amari [1, 2] has developed an extensive theory that relates the risk when θ^* is the true state of Nature to certain differential–geometric properties of the parameter space Θ in the neighborhood of θ^* involving Fisher information and related quantities (see also [38], [54] and [56]).

Some authors have also looked at the value of the relative entropy risk in nonparametric cases as well, for example, [5], [9], [50], [53] and [55]. Also, the issue of consistent estimation of a general probability distribution with respect to relative entropy is addressed in [8] and [28]. However, in the nonparametric case, more extensive work has been done in bounding the risk for other loss functions (see, e.g., [20] and [36]). While this work is too extensive to summarize here, we do note that some authors have also taken the general approach that we take here in using notions of metric entropy (defined later) and specifically using the Hellinger distance in obtaining these bounds (e.g., [9]–[12], [29], [40] and [52]). The only authors we have found who have applied this methodology to the relative entropy risk are Wong and Shen [53] (see Corollary 1, page 360) and Barron and Yang [9]. This work is somewhat complementary to ours, in that it treats instantaneous risk, whereas we focus on cumulative risk. The tools that Wong and Shen employ are considerably more sophisticated, involving bracket entropy methods from empiricial processes, and it appears that the boundedness assumptions they make (e.g., in Theorem 6) are a bit stronger than ours (see the following discussion and at the end of Section 4.2). Different assumptions and different methods (using Fano's inequality) are used to obtain related general results in [9].

In this paper, we describe a new approach, employing the Hellinger metric and certain Laplace integrals, to bounding both the Bayes and the minimax risks for the cumulative relative entropy loss. For most Θ the bounds are fairly tight. We show this for Θ that satisfy some general conditions. To describe the conditions needed, for $\alpha > 1$, define the *α-affinity* between distributions P and Q by $\rho_\alpha(P, Q) = \int (dP)^\alpha (dQ)^{1-\alpha}$. To obtain useful bounds

on the minimax risk, we need to assume that Θ is such that there exist some $\alpha > 1$, some constant $C > 0$ and some distribution Q on Y such that, for all $\theta^* \in \Theta$,

$$\rho_\alpha(P_{\theta^*}, Q) < C.$$

Call this the "α-affinity boundedness condition" on Θ. By fixing α, a separate condition of this type can be defined for each $\alpha > 1$. The condition needed for the Bayes risk upper bound is similar, except that we need only that the expectation of $\rho_\alpha(P_{\theta^*}, Q)$ is at most C, when θ^* is drawn at random according to the prior.

A related boundedness condition is used in the investigation of consistent density estimation with respect to relative entropy risk in [8]. There it is assumed that there is some constant $C > 0$ and some distribution Q on Y such that, for all $\theta^* \in \Theta$,

$$D_{KL}(P_{\theta^*}, Q) < C.$$

This might be called the "relative entropy boundedness condition." It is clear that the minimax risk R_n^{minimax} is infinite for Θ, even for $n = 1$, if the relative entropy boundedness condition is not satisfied. So this condition is necessary for the analysis of the minimax risk to be nontrivial. [It should be noted that since $D_{KL}(P\|Q) = \infty$ whenever Q does not dominate P, the assumption that we mentioned in Section 2 that all the distributions P_θ for $\theta \in \Theta$ have a common dominating measure is actually weaker than the relative entropy boundedness condition. Hence, the minimax risk is trivially infinite without this assumption, showing that we obtain essentially no loss in generality by making this assumption.] In fact, it can be shown that the α-boundedness condition that we impose in our results on minimax risk is strictly stronger than the relative entropy boundedness condition for all $\alpha > 1$. However, in some sense, it is not much stronger than the relative entropy boundedness condition, as it can also be shown that a simple function of the α-affinity, the I-divergence defined later, approaches the relative entropy as α approaches 1 [47]. Furthermore, it can be shown that the α-affinity boundedness condition is weaker than the integrable envelope condition, $\int \sup_{\theta^* \in \Theta} dP_{\theta^*} < \infty$, used in the minimax analysis of relative entropy risk in [53]. Further discussion of the relationship between these conditions is given at the end of Section 4.2. Note that there we reparametrize by setting $\alpha = 1 + \lambda$, $\lambda > 0$, to avoid confusion with another usage of α.

In the following sections, we also discuss the sense in which these boundedness conditions may be viewed as assuming that the minimax risk is finite for alternate definitions of the loss function, in this case by using an α-affinity in place of the relative entropy. This viewpoint allows us to bound the minimax risk for the cumulative relative entropy loss for each $n \geq 1$ in terms of a function of n plus an additive constant that is the minimax risk for $n = 1$, but using an α-affinity loss with $\alpha > 1$ in place of the relative entropy loss (see Lemma 7).

4. Bounds on mutual information and relative entropy distance to a mixture. Since we can obtain the minimax risk as a supremum of Bayes risks, we now focus our attention on the Bayes risk. As noted previously, the Bayes risk $R_{n,\mu}^{\text{Bayes}}$ is the mutual information $I(\Theta^*, Y^n)$ between the random variable Θ^* giving the choice of θ^* according to the prior μ and the observations Y^n. We now give general bounds on this mutual information. In addition, since the risk for a particular state of Nature θ^* using the Bayes strategy P_μ^{Bayes} is

$$R_{n,P_\mu^{\text{Bayes}}}(\theta^*) = D_{KL}\left(P_{\theta^*}^n \| M_{n,\mu}\right),$$

where $M_{n,\mu} = \int P_\theta^n \, d\mu(\theta)$, we will seek bounds for this quantity as well. The latter bounds actually address the general problem of bounding the relative entropy distance from an n-fold product distribution to a mixture of such distributions.

In obtaining these bounds, we use several notions of "distance" between probability distributions based on the α-affinities. One such family of distances are the I-divergences introduced by Rényi [47]. For any real $\alpha \neq 1$ and distributions P and Q, the *I-divergence of order* α is defined by

$$I_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int (dP)^\alpha (dQ)^{1-\alpha}. \tag{2}$$

For $0 < \alpha < 1$, a related set of distances is defined by

$$\begin{aligned} D_\alpha(P,Q) &= \frac{1}{1-\alpha}\left(1 - \int (dP)^\alpha (dQ)^{1-\alpha}\right) \\ &= \frac{1}{1-\alpha}\int \left(\alpha\, dP + (1-\alpha)\, dQ - (dP)^\alpha (dQ)^{1-\alpha}\right). \end{aligned} \tag{3}$$

Since $\alpha x + (1-\alpha)y - x^\alpha y^{1-\alpha} \geq 0$ for any $x, y \geq 0$ and $0 \leq \alpha \leq 1$, the integrand is everywhere nonnegative in the rightmost definition of D_α, showing that $D_\alpha(P,Q) \geq 0$. (This is essentially Hölder's inequality.) Since $-\log x \geq 1 - x$, it follows that $I_\alpha(P\|Q) \geq D_\alpha(P,Q)$, and hence $I_\alpha(P\|Q) \geq 0$ as well. Since $-\log x \approx 1 - x$ for x near 1, these quantities are similar when the α-affinity $\int (dP)^\alpha (dQ)^{1-\alpha}$ is close to 1. Finally, for the case $\alpha = 1$, we define

$$D_1(P,Q) = I_1(P\|Q) = D_{KL}(P\|Q) = \int \left(dQ - dP - dP \log \frac{dQ}{dP}\right). \tag{4}$$

Since $\log z \leq z - 1$, it follows that $y - x - x \log(y/x) \geq 0$ for all $x, y \geq 0$. Hence the integrand in the rightmost expression is everywhere nonnegative. It can be shown that both $D_\alpha(P,Q)$ and $I_\alpha(P\|Q)$ are increasing in α for $\alpha > 0$.

One important special case of the aforementioned distances is the squared Hellinger distance

$$D_{HL}^2(P,Q) = D_{1/2}(P,Q) = \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2.$$

Unlike the other distances and divergences discussed previously, the distance $D_{HL}(P, Q)$, that is, the square root of the previously defined D_{HL}^2, is a metric, since it is symmetric and satisfies a triangle inequality. This metric has been used to give bounds on the risk of estimation procedures in statistics by many authors, including Le Cam [40], Birgé [10, 11], Hasminskii and Ibragimov [29] and van de Geer [52].

4.1. *Basic bounds.* Our main theorem gives bounds on $I(\Theta^*; Y^n)$ and $D_{KL}(P_{\theta^*}^n \| M_{n,\mu})$ in terms of the logarithms of two Laplace transforms of the I-divergence, one at the value $\alpha = 1$ (the relative entropy) and the other at some α between 0 and 1.

THEOREM 2. *Let μ be any prior measure on Θ and let $0 < \alpha < 1$. For each $\theta \in \Theta$, let Q_θ be an arbitrary conditional distribution on Y given θ and let Q_θ^n be the n-fold product of Q_θ. For every $n \geq 1$,*

1.

$$
\begin{aligned}
&-\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\bigl[-n(1-\alpha) I_\alpha\bigl(P_{\theta^*} \| P_{\tilde{\theta}}\bigr)\bigr] \\
&\quad \leq R_{n,\mu}^{\text{Bayes}} \\
&\quad = I(\Theta^*; Y^n) \\
&\quad \leq -\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\bigl[-n I_1\bigl(P_{\theta^*} \| Q_{\tilde{\theta}}\bigr)\bigr].
\end{aligned}
$$

2. *For any $\gamma > 0$, there exists a subset Θ_γ of Θ with measure at least $1 - 2e^{-\gamma}$ under the prior μ such that, for all $\theta^* \in \Theta_\gamma$,*

$$
\begin{aligned}
&-\log \int_\Theta d\mu(\tilde{\theta}) \exp\bigl[-n(1-\alpha) I_\alpha\bigl(P_{\theta^*} \| P_{\tilde{\theta}}\bigr)\bigr] - \gamma \\
&\quad \leq R_{n, P_\mu^{\text{Bayes}}}(\theta^*) \\
&\quad = D_{KL}\bigl(P_{\theta^*}^n \| M_{n,\mu}\bigr) \\
&\quad \leq -\log \int_\Theta d\mu(\tilde{\theta}) \exp\bigl[-n I_1\bigl(P_{\theta^*} \| Q_{\tilde{\theta}}\bigr)\bigr] + \gamma.
\end{aligned}
$$

The upper bound of part 1 is similar to results given in [5] and is mentioned there for the case $P = Q$. To the best of our knowledge, the lower bound and the results in part 2 are new.

The proof is given in a series of lemmas and calculations. We prove the upper bounds of both parts of the theorem first, then the lower bounds. In establishing the bounds in part 2, we will show that there is a set of μ-measure at most $e^{-\gamma}$ on which the lower bound fails and similarly for the upper bound. Hence, both bounds hold on the complement of the union of these two sets, which has μ-measure at least $1 - 2e^{-\gamma}$.

We begin with the upper bounds. This requires the following lemma which has been previously utilized in the framework of statistical physics [51].

LEMMA 1. *Let $P = P(w)$ be a measure on a set W and $Q = Q(v)$ be a measure on a set V. For any real-valued function $u(w, v)$,*

$$-\int_V dQ(v) \log \int_W dP(w) \exp[u(w,v)] \leq -\log \int_W dP(w) \exp\left[\int_V dQ(v)\, u(w,v)\right].$$

PROOF. First note that by Hölder's inequality, for any real-valued functions u_1 and u_2 and $0 \leq \alpha \leq 1$,

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

Taking logs, this shows that $\log \int_W dP(w) e^{u(w,v)}$ is convex in u. The result then follows by applying Jensen's inequality. □

We also use this simple lemma, suggested to us by Meir Feder. Let $P = P(v, w)$ be a measure on the product space $V \times W$, with conditional distribution $P(v|w)$ on V and marginal distribution $P(w)$ on W.

LEMMA 2. *For any sets W and V, measure P on $V \times W$, and nonnegative function $f(v, w)$ such that $\int_{V \times W} dP(v, w) f(v, w) = 1$,*

1.

$$\int_{V \times W} dP(v,w) \log f(v,w) \leq 0.$$

2. *For any $\gamma > 0$,*

$$\Pr\left(w: \int_V dP(v|w) \log f(v,w) \geq \gamma\right) \leq e^{-\gamma}.$$

PROOF. For the first part, $\int_{V \times W} dP(v, w) \log f(v, w) = -\infty < 0$ if $f(v, w) = 0$ on a set of positive measure. Otherwise, note that by Jensen's inequality

$$\int_{V \times W} dP(v,w) \log f(v,w) \leq \log \int_{V \times W} dP(v,w)\, f(v,w) = 0.$$

For the second part, the case where $f(v,w)=0$ for a set of v positive measure under the conditional distribution of V given w is similarly trivial, and otherwise note that

$$\begin{aligned}
&\Pr\left(w\colon \int_V dP(v|w)\log f(v,w)\geq\gamma\right)\\
&\quad=\Pr\left(w\colon \exp\left[\int_V dP(v|w)\log f(v,w)\right]\geq\exp(\gamma)\right)\\
&\quad\leq \exp(-\gamma)\int_W dP(w)\exp\left[\int_V dP(v|w)\log f(v,w)\right]\\
&\quad\leq \exp(-\gamma)\int_W dP(w)\int_V dP(v|w)\,f(v,w)\\
&\quad=\exp(-\gamma).
\end{aligned}$$

The first inequality follows from Markov's inequality and the second from Jensen's inequality. □

In establishing the upper bounds, we use Lemma 2 with $V=Y^n$, $W=\Theta$ and $f(v,w)=\int_\Theta d\mu(\tilde\theta)\,dQ_{\tilde\theta}^n(y^n)/dM_{n,\mu}(y^n)$. Here we assume all y^n such that $dM_{n,\mu}=0$ have been removed from the domain of f, so that f is finite. The conditions of the lemma are satisfied, since this function is nonnegative and

$$\int_{\Theta\times Y^n} d\mu(\theta^*)\,dP_{\theta^*}^n(y^n)\frac{\int_\Theta d\mu(\tilde\theta)\,dQ_{\tilde\theta}^n(y^n)}{dM_{n,\mu}(y^n)}=\int_{Y^n}\int_\Theta d\mu(\tilde\theta)\,dQ_{\tilde\theta}^n(y^n)=1,$$

since $M_{n,\mu}(y^n)=\int_\Theta d\mu(\theta^*)P_{\theta^*}^n(y^n)$. Employing Lemma 2 with this choice of f, the following chain of inequalities holds for all θ^* except for a set of μ-measure at most $e^{-\gamma}$:

$$\begin{aligned}
D_{KL}\left(P_{\theta^*}^n\|M_{n,\mu}\right)&=\int_{Y^n} dP_{\theta^*}^n\log\frac{dP_{\theta^*}^n}{dM_{n,\mu}}\\
&=\int_{Y^n} dP_{\theta^*}^n\left(\log\frac{dP_{\theta^*}^n}{\int_\Theta d\mu(\tilde\theta)\,dQ_{\tilde\theta}^n}+\log\frac{\int_\Theta d\mu(\tilde\theta)\,dQ_{\tilde\theta}^n}{dM_{n,\mu}}\right)\\
&\leq\int_{Y^n} dP_{\theta^*}^n\log\frac{dP_{\theta^*}^n}{\int_\Theta d\mu(\tilde\theta)\,dQ_{\tilde\theta}^n}+\gamma\\
&=-\int_{Y^n} dP_{\theta^*}^n\log\int_\Theta d\mu(\tilde\theta)\frac{dQ_{\tilde\theta}^n}{dP_{\theta^*}^n}+\gamma\\
&=-\int_{Y^n} dP_{\theta^*}^n\log\int_\Theta d\mu(\tilde\theta)\exp\left(\log\frac{dQ_{\tilde\theta}^n}{dP_{\theta^*}^n}\right)+\gamma
\end{aligned}$$

$$\leq -\log \int_{\Theta} d\mu(\tilde{\theta}) \exp\left[\int_{Y^n} dP_{\theta^*}^n \log \frac{dQ_{\tilde{\theta}}^n}{dP_{\theta^*}^n} \right] + \gamma$$

$$= -\log \int_{\Theta} d\mu(\tilde{\theta}) \exp\left[-D_{KL}(P_{\theta^*}^n \| Q_{\tilde{\theta}}^n) \right] + \gamma$$

$$= -\log \int_{\Theta} d\mu(\tilde{\theta}) \exp\left[-nD_{KL}(P_{\theta^*} \| Q_{\tilde{\theta}}) \right] + \gamma,$$

where the first inequality follows from Lemma 2, part 2, and the second one from Lemma 1. The last equality follows from the fact that the *KL*-divergence is additive over the product of independent distributions (see, e.g., [18], page 23). Note that by our convention that $0 \log 0 = 0$, for each θ^*, the set of y^n such that $dP_{\theta^*}^n(y^n) = 0$ can simply be removed in the first equality and then reintroduced in the exponent of the second-to-the-last inequality, thus avoiding any division by 0 for these cases. Similarly, if $\int_{\Theta} d\mu(\tilde{\theta})\, dQ_{\tilde{\theta}}^n(y^n) = 0$ for a set of y^n of positive measure with respect to $P_{\theta^*}^n$, then all upper bounds from the second line on are infinite, and the result holds trivially. Otherwise, a set of y^n of measure 0 on which $\int_{\Theta} d\mu(\tilde{\theta})\, dQ_{\tilde{\theta}}^n(y^n) = 0$ can be ignored, avoiding any division by 0 in this regard. Since $D_{KL} = I_1$, this establishes the upper bound of part 2 of Theorem 2.

The upper bound of part 1 of Theorem 2 is established in a very similar manner. Here we note that

$$I(\Theta^*; Y^n) = \int_{\Theta} d\mu(\theta^*) \int_{Y^n} dP_{\theta^*}^n \log \frac{dP_{\theta^*}^n}{dM_{n,\mu}}$$

$$= \int_{\Theta} d\mu(\theta^*) \int_{Y^n} dP_{\theta^*}^n \left(\log \frac{dP_{\theta^*}^n}{\int_{\Theta} d\mu(\tilde{\theta})\, dQ_{\tilde{\theta}}^n} + \log \frac{\int_{\Theta} d\mu(\tilde{\theta})\, dQ_{\tilde{\theta}}^n}{dM_{n,\mu}} \right)$$

$$\leq \int_{\Theta} d\mu(\theta^*) \int_{Y^n} dP_{\theta^*}^n \log \frac{dP_{\theta^*}^n}{\int_{\Theta} d\mu(\tilde{\theta})\, dQ_{\tilde{\theta}}^n},$$

where the inequality follows from Lemma 2, part 1. The remainder of the proof consists of the identical chain of inequalities as in the preceding proof of the upper bound of part 2, except that we take expectation over θ^* and we do not have the term $+\gamma$.

We turn now to the lower bounds. Here we use the following lemma, which is new, as far as we can tell. Let $P = P(v, w)$ be a measure on the product space $V \times W$, with conditional distribution $P(v|w)$ on V and marginal distribution $P(w)$ on W. For any $0 < \lambda \leq 1$, define

$$I^{(\lambda)}(W; V) = -\int_{V \times W} dP(v, w^*) \log \int_W dP(w) \left(\frac{dP(v|w)}{dP(v|w^*)} \right)^{\lambda}.$$

It is easy to see that $I^{(\lambda)}(W; V)$ is well defined [34]. Note that $I^{(1)}(W; V) = I(W; V)$, the mutual information between W and V.

LEMMA 3. *Whenever* $\int_W dP(w)\, dP(v|w) > 0$ *for all* v *and* $0 < \lambda \le 1$,

1.

$$I^{(\lambda)}(W;V) - I(W;V) \le 0.$$

2.

$$\Pr\left\{ w^*\colon dP(v|w^*) > 0 \text{ and } \int_V dP(v|w^*) \times \left(\log \int_W dP(w)\, \frac{dP(v|w)}{dP(v|w^*)} - \log \int_W dP(w) \left(\frac{dP(v|w)}{dP(v|w^*)} \right)^{\lambda} \right) \ge \gamma \right\} \le e^{-\gamma}.$$

PROOF. This follows from Lemma 2 using the function

$$f(v, w^*) = \frac{\int_W dP(w)(dP(v|w)/dP(v|w^*))}{\int_W dP(w)(dP(v|w)/dP(v|w^*))^{\lambda}} = \frac{dP^{\lambda-1}(v|w^*) \int_W dP(w)\, dP(v|w)}{\int_W dP(w)\, dP^{\lambda}(v|w)}.$$

The conditions of the lemma are satisfied, since f is nonnegative and

$$\begin{aligned} &\int_{V\times W} dP(v, w^*) f(v, w^*) \\ &\quad = \int_{V\times W} dP(v, w^*) \frac{dP^{\lambda-1}(v|w^*) \int_W dP(w)\, dP(v|w)}{\int_W dP(w)\, dP^{\lambda}(v|w)} \\ &\quad = \int_V \int_W dP(w^*)\, dP(v|w^*) \frac{dP^{\lambda-1}(v|w^*) \int_W dP(w)\, dP(v|w)}{\int_W dP(w)\, dP^{\lambda}(v|w)} \qquad (5) \\ &\quad = \int_V \frac{\int_W dP(w^*)\, dP^{\lambda}(v|w^*) \int_W dP(w)\, dP(v|w)}{\int_W dP(w)\, dP^{\lambda}(v|w)} \\ &\quad = \int_V \int_W dP(w)\, dP(v|w) \\ &\quad = 1. \end{aligned}$$

□

Now note that if $\{y^n\colon \int_\Theta d\mu(\tilde{\theta})\, dP_{\tilde{\theta}}^n(y^n) = 0\}$ has positive measure under the distribution $P_{\theta^*}^n$, then $D_{KL}(P_{\theta^*}^n \| M_{n,\mu}) = \infty$. Hence, the lower bound holds

trivially. Otherwise, a set of such y^n of measure 0 can be ignored, and, using part 2 of Lemma 3 with $W = \Theta$ and $V = Y^n$, we can show that the following inequalities hold except on a set of θ^* with μ-measure at most $e^{-\gamma}$:

$$\begin{aligned}
D_{KL}\big(P_{\theta^*}^n \| M_{n,\mu}\big) &= -\int_{Y^n} dP_{\theta^*}^n \log \int_\Theta d\mu(\tilde{\theta}) \frac{dP_{\tilde{\theta}}^n}{dP_{\theta^*}^n} \\
&\geq -\int_{Y^n} dP_{\theta^*}^n \log \int_\Theta d\mu(\tilde{\theta}) \left(\frac{dP_{\tilde{\theta}}^n}{dP_{\theta^*}^n}\right)^\lambda - \gamma \\
&\geq -\log \int_\Theta d\mu(\tilde{\theta}) \int_{Y^n} dP_{\theta^*}^n \left(\frac{dP_{\tilde{\theta}}^n}{dP_{\theta^*}^n}\right)^\lambda - \gamma \\
&= -\log \int_\Theta d\mu(\tilde{\theta}) \int_{Y^n} \big(dP_{\theta^*}^n\big)^{1-\lambda} \big(dP_{\tilde{\theta}}^n\big)^\lambda - \gamma \\
&= -\log \int_\Theta d\mu(\tilde{\theta}) \left[\int_Y \big(dP_{\theta^*}\big)^{1-\lambda} \big(dP_{\tilde{\theta}}\big)^\lambda\right]^n - \gamma \\
&= -\log \int_\Theta d\mu(\tilde{\theta}) \exp\left[n \log \int_Y \big(dP_{\theta^*}\big)^{1-\lambda} \big(dP_{\tilde{\theta}}\big)^\lambda\right] - \gamma \\
&= -\log \int_\Theta d\mu(\tilde{\theta}) \exp\big[-n\lambda I_{1-\lambda}\big(P_{\tilde{\theta}} \| P_{\theta^*}\big)\big] - \gamma.
\end{aligned}$$

As in the proof of the upper bound, to avoid division by 0 and to apply Lemma 3, we can remove the set of y^n such that $dP_{\theta^*}^n(y^n) = 0$ from the first line and reintroduce it in the fourth line. Setting $\alpha = 1 - \lambda$, this establishes the lower bound of part 2.

As with the upper bound, the lower bound of part 1 is established easily by removing the $-\gamma$ terms and taking expectation over θ^* in the previous chain of inequalities, using part 1 of Lemma 3 in line 2. This establishes the lower bounds and completes the proof of Theorem 2. □

A few brief comments about Theorem 2 are in order. First, note that if in part 2 we let γ grow with n in a suitable way, we obtain bounds which asymptotically hold for almost all $\theta^* \in \Theta$. An even stronger result is obtained when we choose $\gamma(n)$ such that $\sum_{n=1}^\infty e^{-\gamma(n)} < \infty$. This holds, for example, if we let $\gamma(n)$ grow faster than $\log n$. Then the first Borel–Cantelli lemma shows that, for μ almost all $\theta^* \in \Theta$, the bounds will be violated only a finite number of times as $n \to \infty$.

It should also be noted that, in the important special case when $P = Q$, the upper bound of part 2 of the theorem holds with $\gamma = 0$, since we can omit the first few steps of its derivation in this case, where γ is introduced. Thus, both this strengthened upper bound and the given lower bound hold on a set of measure $1 - e^{-\gamma}$ in this case.

Finally, we note that part 2 is related to part 1 in the same way that the strong redundancy–capacity theorem of universal coding in [42] is related to the usual theorems concerning average redundancy.

It is possible to state a variant of Theorem 2 using the D_α-distances. Here we also make use of a particular choice for the family of distributions Q_θ that appear in Theorem 2. Another possible choice is explored in Theorem 3. We will need the following definition.

For each $0 < \alpha < 1$ and $x > 0$, define

$$b_\alpha(x) = \frac{(1-\alpha)(x - \log x - 1)}{\alpha + (1-\alpha)x - x^{1-\alpha}}. \tag{6}$$

Define $b_\alpha(0) = \infty$. It is easily verified that $b_\alpha(x)$ is strictly decreasing in x, approaches 1 as $x \to \infty$ and approaches ∞ as $x \to 0$. Let

$$B_\alpha(\Theta) = \sup_{y \in Y,\ \theta^*,\ \theta \in \Theta} b_\alpha\left(\frac{dP_{\theta^*}(y)}{dP_\theta(y)}\right).$$

Clearly, this constant does not depend on the choice of the dominating measure ν.

COROLLARY 1. *For every* $0 < \alpha < 1$ *and* $n \geq 1$,

1.

$$\begin{aligned} -\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-n(1-\alpha) D_\alpha(P_{\theta^*}, P_{\tilde{\theta}})\right] \\ \leq R^{\text{Bayes}}_{n,\mu} \\ = I(\Theta^*; Y^n) \\ \leq -\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-nB_\alpha(\Theta) D_\alpha(P_{\theta^*}, P_{\tilde{\theta}})\right]. \end{aligned}$$

2. *For any* $\gamma > 0$, *there exists a subset* Θ_γ *of* Θ *with measure at least* $1 - 2e^{-\gamma}$ *under the prior* μ *such that, for all* $\theta^* \in \Theta_\gamma$,

$$\begin{aligned} -\log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-n(1-\alpha) D_\alpha(P_{\theta^*}, P_{\tilde{\theta}})\right] - \gamma \\ \leq R_{n, P^{\text{Bayes}}_\mu}(\theta^*) \\ = D_{KL}\left(P^n_{\theta^*} \| M_{n,\mu}\right) \\ \leq -\log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-nB_\alpha(\Theta) D_\alpha(P_{\theta^*}, P_{\tilde{\theta}})\right] + \gamma. \end{aligned}$$

PROOF. Since $I_\alpha(P\|Q) \geq D_\alpha(P, Q)$, the lower bounds follow directly from the lower bounds of Theorem 2. For the upper bounds, we will need the following lemma, which is a simple extension of Lemma 4.4 of [10].

LEMMA 4. *For any distributions P and Q on Y and any $0 < \alpha < 1$,*

$$D_{KL}(P\|Q) \le \left(\sup_{y \in Y} b_\alpha\left(\frac{dQ(y)}{dP(y)} \right) \right) D_\alpha(P, Q).$$

PROOF. If $dP = dQ$ except on a set of measure 0 (w.r.t. the dominating measure ν), then $D_{KL}(P\|Q) = 0$ and hence the result holds. So it suffices to consider the case where $D_\alpha(P, Q) > 0$. Let $S = \{y \in Y: dP(y) = 0\}$. Separating Y into S and $Y - S$ and factoring a dP out of the integrands in (3) and (4) in the latter case, we have

$$\frac{D_{KL}(P\|Q)}{D_\alpha(P, Q)} = \frac{(1-\alpha)\int_{Y-S} dP\left(\frac{dQ}{dP} - \log\frac{dQ}{dP} - 1 \right) + \int_S dQ}{\int_{Y-S} dP\left(\alpha + (1-\alpha)\frac{dQ}{dP} - \left(\frac{dQ}{dP} \right)^{1-\alpha} \right) + \int_S dQ}$$

$$\le \sup_{y \in Y} b_\alpha\left(\frac{dQ(y)}{dP(y)} \right),$$

since $b_\alpha \ge 1$. □

The upper bounds of Corollary 1 follow from Theorem 2 and this lemma by setting $Q_\theta = P_\theta$. □

Whenever dP_θ is uniformly bounded above 0 and below ∞ for all y and θ for some choice of the dominating measure, $B_\alpha(\Theta)$ is finite, and this corollary can be applied. However, in some other cases, $B_\alpha(\Theta) = \infty$ for all $0 < \alpha < 1$, making the upper bound in the Corollary 1 useless. One case where this occurs is when there are θ and θ^* in Θ such that P_θ is not dominated by P_{θ^*}. For example, if $Y = \{0, 1\}$ and there is a θ^* such that $P_{\theta^*}(Y = 1)$ is 0 (or 1) and there is also a θ where $P_\theta(Y = 1)$ is not 0 (or not 1), then P_θ is not dominated by P_{θ^*}. We can also have $B_\alpha(\Theta) = \infty$ in cases where such lack of domination does not occur. For example, if $Y = \{0, 1\}$, Θ is the open interval $(0, 1)$ and $P_\theta(Y = 1) = \theta$, then $B_\alpha(\Theta) = \infty$ not because there are two distributions that fail to mutually dominate each other, but because $\inf_{y \in Y,\ \theta^*,\ \theta \in \Theta} dP_{\theta^*}(y)/dP_\theta(y) = 0$. Such cases can be handled by the results in the following section.

4.2. *The $(1 + \lambda)$-affinity boundedness condition.* Here we prove a version of Corollary 1 that can be used in cases when $B_\alpha(\Theta) = \infty$ for all $0 < \alpha < 1$. This new theorem requires only the Bayes version of the weaker $(1 + \lambda)$-affinity boundedness condition described in Section 3, for some $\lambda > 0$. For a fixed prior μ, define

$$R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = \inf_{\hat{P}} \int_\Theta d\mu(\theta^*) \int (dP_{\theta^*})^{1+\lambda} (d\hat{P})^{-\lambda}.$$

This is the Bayes risk for a game much like the one we are studying, except that the relative entropy loss is replaced by the $(1 + \lambda)$-affinity loss, and we have fixed the number n of observations to 1. Using Jensen's inequality, it can be verified that when $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$, the minimizing $\hat{P}$, that is, the Bayes strategy, is the distribution $U = U_\mu$ defined by

$$dU = \frac{\left(\int_\Theta d\mu(\theta^*)\,(dP_{\theta^*})^{1+\lambda}\right)^{1/(1+\lambda)}}{C_{\lambda,\mu}},$$

where

$$C_{\lambda,\mu} = \int_Y \left(\int_\Theta d\mu(\theta^*)\,(dP_{\theta^*})^{1+\lambda}\right)^{1/(1+\lambda)};$$

see Zhu and Rohwer [57]. Hence, for each individual θ^*, the risk of the Bayes strategy is

$$\begin{aligned} R_{1,U_\mu,\rho_{1+\lambda}}(\theta^*) &= \int_Y (dP_{\theta^*})^{1+\lambda}(dU)^{-\lambda} \\ &= C^\lambda_{\lambda,\mu}\int_Y (dP_{\theta^*})^{1+\lambda}\left(\int_\Theta d\mu(\theta^*)\,(dP_{\theta^*})^{1+\lambda}\right)^{-\lambda/(1+\lambda)} \end{aligned}$$

and the Bayes risk is

$$\begin{aligned} R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} &= \int_\Theta d\mu(\theta^*)\int_Y (dP_{\theta^*})^{1+\lambda}(dU)^{-\lambda} \\ &= C^\lambda_{\lambda,\mu}\int_\Theta d\mu(\theta^*)\int_Y (dP_{\theta^*})^{1+\lambda}\left(\int_\Theta d\mu(\theta^*)\,(dP_{\theta^*})^{1+\lambda}\right)^{-\lambda/(1+\lambda)} \\ &= C^\lambda_{\lambda,\mu}\int_Y \left(\int_\Theta d\mu(\theta^*)\,(dP_{\theta^*})^{1+\lambda}\right)^{1/(1+\lambda)} \\ &= C^{1+\lambda}_{\lambda,\mu}. \end{aligned}$$

We have the following theorem.

THEOREM 3. *Let* $0 < \alpha < 1$ *and* $0 < \lambda \le 1$. *Assume* $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$. *Then, for every* $n \ge 1$,

1.

$$\begin{aligned} &-\int_\Theta d\mu(\theta^*)\log\int_\Theta d\mu(\tilde{\theta})\exp\left[-n(1-\alpha)D_\alpha(P_{\theta^*},P_{\tilde{\theta}})\right] \\ &\quad\le R^{\text{Bayes}}_{n,\mu} \\ &\quad= I(\Theta^*;Y^n) \end{aligned}$$

$$\leq -\int_\Theta d\mu(\theta^*)\log\int_\Theta d\mu(\tilde\theta) \times\exp\left[-(n\log n)\frac{(1+o(1))4(1-\alpha)}{\alpha\lambda}D_\alpha(P_{\theta^*},P_{\tilde\theta})\right] + R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} + o(1).$$

2. *For any $\gamma>0$ there exists a subset Θ_γ of Θ with measure at least $1-2e^{-\gamma}$ under the prior μ such that, for all $\theta^*\in\Theta_\gamma$,*

$$\begin{aligned}
&-\log\int_\Theta d\mu(\tilde\theta)\exp\left[-n(1-\alpha)D_\alpha(P_{\theta^*},P_{\tilde\theta})\right]-\gamma\\
&\quad\leq R_{n,P_\mu^{\text{Bayes}}}(\theta^*)\\
&\quad= D_{KL}\left(P_{\theta^*}^n\|M_{n,\mu}\right)\\
&\quad\leq -\log\int_\Theta d\mu(\tilde\theta)\exp\left[-(n\log n)\frac{(1+o(1))4(1-\alpha)}{\alpha\lambda}D_\alpha(P_{\theta^*},P_{\tilde\theta})\right]\\
&\qquad + R_{1,U_\mu,\rho_{1+\lambda}}(\theta^*)+\gamma+o(1),
\end{aligned}$$

where, in each case for fixed α and λ, $o(1)$ is a function $f(n)$ such that $f(n)\to 0$ as $n\to\infty$. Furthermore, the same results also hold replacing the quantity $D_\alpha(P_{\theta^*},P_{\tilde\theta})$ with $I_\alpha(P_{\theta^*}\|P_{\tilde\theta})$.

PROOF. That D can be replaced by I follows from the fact that $I_\alpha(P\|Q)\geq D_\alpha(P,Q)$ for all α, as pointed out in the proof of Corollary 1. To prove the result for D, we will need a lemma. (We recently noticed that a related lemma is given in [53], Theorem 5, although no explicit relationship with the α-affinities is given in the latter result.)

LEMMA 5. *Assume $0<\alpha<1$ and $\lambda>0$. Let P, R and U be any distributions on Y. Let $c_\lambda=\int(dP)^{1+\lambda}(dU)^{-\lambda}$. Let $Q=(1-\varepsilon)R+\varepsilon U$ for some $\varepsilon>0$ such that $\log\log(1/\varepsilon)/\log(1/\varepsilon)\leq\lambda/2$ and $\varepsilon\leq\exp[-\alpha/(2(1-\alpha))]$. Then*

$$D_{KL}(P\|Q)\leq\frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)}D_\alpha(P,R)+\frac{2\varepsilon\log(1/\varepsilon)}{(1-\alpha)f_\alpha(\varepsilon^2)}+\varepsilon^{\lambda/2}c_\lambda,$$

where

$$f_\alpha(x)=\frac{\alpha+(1-\alpha)x-x^{1-\alpha}}{1-\alpha}.$$

The proof of this lemma is given in the Appendix.

Now let U_μ be the Bayes strategy as defined previously. Since $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}<\infty$, U_μ is well defined. For each $\theta\in\Theta$, let

$$Q_\theta=(1-\varepsilon)P_\theta+\varepsilon U_\mu,$$

with $\varepsilon = n^{-2/\lambda}$. It is clear that $f_\alpha(\varepsilon^2) \to \alpha/(1-\alpha)$ as $\varepsilon \to 0$. Hence, by Lemma 5, for sufficiently large n, for all θ,

$$
\begin{aligned}
&D_{KL}(P_{\theta^*} \| Q_{\tilde{\theta}}) \\
&\quad \le \frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)} D_\alpha(P_{\theta^*}, P_{\tilde{\theta}}) + \frac{2\varepsilon\log(1/\varepsilon)}{(1-\alpha) f_\alpha(\varepsilon^2)} + \varepsilon^{\lambda/2} R_{1, U_\mu, \rho_{1+\lambda}}(\theta^*) \\
&\quad = \frac{4\log n}{\lambda f_\alpha(n^{-4/\lambda})} D_\alpha(P_{\theta^*}, P_{\tilde{\theta}}) + \frac{R_{1, U_\mu, \rho_{1+\lambda}}(\theta^*) + o(1)}{n} \quad \text{since } \lambda \le 1 \\
&\quad = \log n \frac{(1+o(1))4(1-\alpha)}{\alpha\lambda} D_\alpha(P_{\theta^*}, P_{\tilde{\theta}}) + \frac{R_{1, U_\mu, \rho_{1+\lambda}}(\theta^*) + o(1)}{n}.
\end{aligned}
$$

Since $\int_\Theta d\mu(\theta^*) R_{1, U_\mu, \rho_{1+\lambda}}(\theta^*) = R^{\text{Bayes}}_{1, \mu, \rho_{1+\lambda}}$, the result then follows from Theorem 2. □

Note that no attempt has been made to optimize the constants in this theorem.

Now let

$$S(\Theta) = \int_Y \sup_{\theta \in \Theta} dP_\theta.$$

[If $\sup_{\theta \in \Theta} dP_\theta$ is not measurable, then any measurable function that majorizes it can be used instead in the definition of $S(\Theta)$.] We call $\sup_{\theta \in \Theta} dP_\theta$ the *envelope function* for Θ. Note that $S(\Theta)$ is independent of the choice of the dominating measure. Since, for all $\lambda \ge 0$,

$$\left(\int_\Theta d\mu(\theta^*) (dP_{\theta^*})^{1+\lambda} \right)^{1/(1+\lambda)} \le \sup_{\theta^* \in \Theta} dP_{\theta^*}.$$

It follows that

$$R^{\text{Bayes}}_{1, \mu, \rho_{1+\lambda}} = C^{1+\lambda}_{\lambda, \mu} = \left(\int_Y \left(\int_\Theta d\mu(\theta^*) (dP_{\theta^*})^{1+\lambda} \right)^{1/(1+\lambda)} \right)^{1+\lambda} \le S^{1+\lambda}(\Theta)$$

for all $\lambda > 0$. Hence, whenever Θ has an integrable envelope function, that is, whenever $S(\Theta) < \infty$, then $R^{\text{Bayes}}_{1, \mu, \rho_{1+\lambda}} < \infty$, and the bounds in part 1 of Theorem 3 hold with $\lambda = 1$ and $R^{\text{Bayes}}_{1, \mu, \rho_{1+\lambda}}$ replaced with $S^2(\Theta)$. It is clear that $S(\Theta) < \infty$ whenever Y is finite and whenever Y is a bounded set in R^k for some $k \ge 1$ and the densities in $\{P_\theta\colon \theta \in \Theta\}$ are uniformly upper bounded. Hence, Theorem 3 always applies in these cases.

Theorem 3 also applies in many cases where $S(\Theta)$ is infinite; an example of such a case is given in the following section. To characterize the types of Θ and priors μ not covered by Theorem 3, let us define the function $f_{\Theta, \mu}(\lambda)$ for $\lambda \ge 0$ by

$$f_{\Theta, \mu}(\lambda) = \frac{1}{\lambda} \log R^{\text{Bayes}}_{1, \mu, \rho_{1+\lambda}}$$

for $\lambda > 0$ and

$$f_{\Theta,\mu}(0) = R^{\text{Bayes}}_{1,\mu},$$

that is, the risk for one observation ($n = 1$) for the relative entropy loss. It can be shown that, for any Θ and μ, $f_{\Theta,\mu}(\lambda)$ is a nondecreasing function on $[0,\infty)$ taking values in $[0,\infty]$, and if $f_{\Theta,\mu}(\lambda)$ is finite for any $\lambda > 0$, then

$$\lim_{\lambda\to 0} f_{\Theta,\mu}(\lambda) = f_{\theta,\mu}(0).$$

To verify this last property, note that $\lim_{\lambda\to 0} R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = 1$. Hence, by l'Hospital's rule

$$\lim_{\lambda\to 0} f_{\Theta,\mu}(\lambda) = \frac{d}{d\lambda}\left(R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}\right)\Big|_{\lambda=0}.$$

It can be verified by direct calculation that the latter quantity is the mutual information $I(\Theta^*, Y)$, which is the same as $R^{\text{Bayes}}_{1,\mu}$.

It is clear that whenever $R^{\text{Bayes}}_{1,\mu}$ is infinite, then $R^{\text{Bayes}}_{n,\mu}$ is infinite for all $n \geq 1$. Thus, there are only three possible cases for the pair (Θ, μ):

1. $f_{\Theta,\mu}(\lambda) < \infty$ for some $\lambda > 0$. In this case, $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} < \infty$ and hence Theorem 3 applies and may be used to get bounds on $R^{\text{Bayes}}_{n,\mu}$ for all n.
2. $f_{\Theta,\mu}(0) = \infty$. In this case, $R^{\text{Bayes}}_{n,\mu} = \infty$ for all n and hence the problem of bounding this quantity is trivial.
3. $f_{\Theta,\mu}(0) < \infty$ but $f_{\Theta,\mu}(\lambda) = \infty$ for all $\lambda > 0$. In this case, we say that the pair (Θ, μ) is *irregular*. These are the only nontrivial cases where Theorem 3 does not apply.

While it would not be expected that irregular (Θ, μ) would show up much in practice, it is possible to construct one.

EXAMPLE 1. Let $Y = \{1, 2, 3, \ldots\}$, $\Theta = \{3, 4, 5, \ldots\}$ and, for each $\theta \in \Theta$ and $y \in Y$, define $P_\theta(Y = y)$ to be $1 - (1/\log\theta)$ if $y = 1$, $1/\log\theta$ if $y = \theta$ and 0 otherwise. Let $\mu(\theta) = c/(\theta \log^2\theta)$, where $c = \sum_{i=3}^{\infty} 1/(i \log^2 i) < \infty$. Then it can be shown that (Θ, μ) is irregular.

5. Examples. We now illustrate Theorems 2 and 3 by applying them to a few simple problems. We begin with a classical case in which each point $\theta \in \Theta$ is a vector of D real numbers, Θ is a compact set and the prior μ is specified as a density $d\mu(\theta)$. To apply Theorem 2, fix $\theta^* \in \Theta_\gamma$, where θ^* is in the interior of Θ. We assume that the prior $d\mu$ is continuous and positive at θ^*. We also assume that $\{P_\theta\}$ is a smooth family of probabilities such that the *Fisher information* matrix at θ^*, defined by $J(\theta^*)$, where

$$J_{ij}(\theta^*) = \int_Y dP_{\theta^*}\left[\frac{\partial}{\partial\theta_i}\log dP_\theta(y)\,\frac{\partial}{\partial\theta_j}\log dP_\theta(y)\right]\Big|_{\theta=\theta^*}$$

exists and is positive definite. In this case, we will focus on the bounds on the risk for individual θ^*, rather than bounds on the mutual information. Even the simplest choice $Q = P$ will be sufficient to obtain a useful bound in the smooth case. For large n, obviously the main contributions to the inner expectations in Theorem 2 come from small neighborhoods of θ^*. Hence, under certain regularity conditions, Laplace's method can be used to evaluate these expectations asymptotically. We perform a Taylor expansion of the exponents in Theorem 2 to second order in the difference between $\tilde{\theta}$ and θ^* using the partial derivatives

$$\frac{\partial}{\partial \theta_i} I_\alpha(P_{\theta^*} \| P_\theta)\Big|_{\theta = \theta^*} = 0$$

and

$$\frac{\partial^2}{\partial \theta_i \, \partial \theta_j} I_\alpha(P_{\theta^*} \| P_\theta)\Big|_{\theta = \theta^*} = \alpha J_{ij}(\theta^*). \tag{7}$$

Note that these results are also valid for $\alpha = 1$. Hence, Laplace's method would yield for the lower bound

$$\begin{aligned}\int_\Theta d\mu(\tilde{\theta}) &\exp\left[-n(1-\alpha) I_\alpha(P_{\theta^*} \| P_{\tilde{\theta}})\right] \\ &= d\mu(\theta^*) \int_{R^D} d\theta \exp\left[-\frac{n}{2}\alpha(1-\alpha) \sum_{ij} (\theta_i - \theta_i^*) J_{ij}(\theta^*)(\theta_j - \theta_j^*)\right] \\ &\quad \times (1 + o(1)).\end{aligned}$$

A similar expression is obtained for the upper bound. By evaluating the Gaussian integrals, we get

$$\begin{aligned}&\frac{D}{2}\log\frac{n}{2\pi} - \log d\mu(\theta^*) + \frac{1}{2}\log\det J(\theta^*) - \frac{D}{2}\log\frac{1}{\alpha(1-\alpha)} - \gamma + o(1) \\ &\quad \le R_{n, P_\mu^{\text{Bayes}}}(\theta^*) \\ &\quad \le \frac{D}{2}\log\frac{n}{2\pi} - \log d\mu(\theta^*) + \frac{1}{2}\log\det J(\theta^*) + o(1).\end{aligned}$$

(Here we can set $\gamma = 0$ in the upper bounds, as per the comments following Theorem 2.) Note that asymptotically the lower bound is optimized by setting $\alpha = \frac{1}{2}$. In this case, for large n, both bounds differ by a constant approximately equal to $(D \log 4)/2$ for small γ. In this classical case, Clarke and Barron [15] have determined the exact answer to within $o(1)$, and it is

$$R_{n, P_\mu^{\text{Bayes}}}(\theta^*) = \frac{D}{2}\log\frac{n}{2\pi} - \log d\mu(\theta^*) + \frac{1}{2}\log\det J(\theta^*) - \frac{D}{2} + o(1).$$

Thus, our simpler methods do not give the best known additive constants in the bounds for this classical case, but they do provide good bounds for large n.

As pointed out by Clarke and Barron [15], the scaling $\sim (D/2)\log n$ of the Bayes risk for the smooth parametric families is strongly related to the asymptotic normality of the properly normalized posterior distribution. It is interesting to look at nonregular families of probabilities for which the posterior fails to converge to a nontrivial limit. (For conditions that are necessary for convergence, see [26]). As an example for such nonsmooth densities, we study the following simple family on $\mathbb{R}$:

$$dP_\theta(y) = e^{-(y-\theta)} I_{\{y>\theta\}}, \qquad \theta \in \mathbb{R}. \tag{8}$$

Obviously, $D_{KL}(P_{\theta^*} \| P_\theta) = \infty$, whenever $\theta > \theta^*$ and the Fisher information does not exist for any θ. Hence, the previous analysis is not applicable and we have to resort to the more sophisticated upper bounds. Specializing to $\alpha = \frac{1}{2}$, we easily find

$$D_{1/2}(P_{\theta^*}, P_\theta) = 2(1 - e^{-|\theta-\theta^*|}),$$
$$I_{1/2}(P_{\theta^*} \| P_\theta) = |\theta - \theta^*|.$$

This result clearly shows the difference from the smooth families. The distances $D_{1/2}$ and $I_{1/2}$ do not behave locally like a quadratic function for θ close to θ^*, but have a linear scaling. Hence, a different scaling of the risk at θ^* and the mutual information is also expected.

An explicit result using Theorem 3 is easily obtained for the prior $d\mu(\theta) = \frac{1}{2}e^{-|\theta|}$. Note that the envelope of Θ is not integrable, so we must obtain direct bounds on $R_{1,\mu,\rho_{1+\lambda}}$ rather than using $S(\Theta)$. To upper bound $R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} = \inf_{\hat{P}} \int_\Theta d\mu(\theta^*) \int (dP_{\theta^*})^{1+\lambda} (d\hat{P})^{-\lambda}$, it suffices to choose any distribution U and bound the expectation of $c_\lambda(\theta^*) = \int (dP_{\theta^*})^{1+\lambda}(dU)^{-\lambda}$. Here we can set $dU(y) = \frac{1}{2}e^{-|y|}$. In this case, we have $c_\lambda(\theta^*) < e^{\lambda|\theta^*|}$ and $\int_\Theta d\mu(\theta^*) c_\lambda(\theta^*) < \infty$ for all $\lambda < 1$. To evaluate the bounds, we use the fact that, for $a > 1$,

$$\frac{1}{2}\int_{-\infty}^{\infty} d\theta\, e^{-|\theta|-a|\theta-\theta^*|} = \frac{e^{-|\theta^*|} - e^{-a|\theta^*|}}{2(a-1)} + \frac{e^{-|\theta^*|} + e^{-a|\theta^*|}}{2(a+1)}.$$

Hence, for $\alpha = \frac{1}{2}$, we get

$$\begin{aligned} &\log\left(\frac{n}{2}\right) + |\theta^*| - \gamma + o(1) \\ &\quad \le R_{n, P_\mu^{\text{Bayes}}}(\theta^*) \\ &\quad \le \log\left(\frac{4n\log n(1+o(1))}{\lambda}\right) + e^{\lambda|\theta^*|} + |\theta^*| + \gamma + o(1). \end{aligned}$$

Hence, an asymptotic scaling $\sim \log n$ for the risk is observed. This gives a factor of 2 difference compared to the risk of a smooth one-dimensional family of densities.

Finally, we will consider an example where both the parameter space and the space of observations are infinite dimensional. We assume that an unknown real continuous function $\theta(x)$ with $0 \le x \le 1$ is corrupted by a

Gaussian white-noise process. The statistician observes n random functions Y_t, $t = 1, \ldots, n$, which, conditioned on θ, are independent realizations of the process

$$Y(x) = \int_0^x \theta(z)\, dz + \sigma W(x). \tag{9}$$

Here $W(x)$ is a standard Wiener process with $W(0) = 0$ and covariance $\mathbb{E}[W(x_1)W(x_2)] = \min(x_1, x_2)$. In this case, it is easy to calculate the I-divergences explicitly for all α. Let P_θ be the measure corresponding to the random process $Y(x)$ and let the dominating measure ν be the Wiener measure. Then, from the Cameron–Martin formula [13], the Radon-Nikodym derivative is found to be

$$\frac{dP_\theta}{d\nu} = \exp\left[\frac{1}{\sigma}\int_0^1 \theta(x)\, dW(x) - \frac{1}{2\sigma^2}\int_0^1 \theta^2(x)\, dx\right]. \tag{10}$$

Inserting this into the definition of the I-divergences, we obtain

$$I_\alpha(P_{\theta^*} \| P_\theta) = \frac{\alpha}{2\sigma^2}\int_0^1 (\theta(x) - \theta^*(x))^2\, dx. \tag{11}$$

For the case where the prior over the space of functions $\theta(x)$ is a Gaussian measure [such that $\theta(x)$ is a realization of a Gaussian random process], our bounds can be evaluated in closed form. We will restrict ourselves to the case of the mutual information $I(\Theta^*; Y^n)$ and use the fact that, for Gaussian processes and $c > 0$,

$$\begin{aligned} &-\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-\frac{c}{2}\int_0^1 (\tilde{\theta}(x) - \theta^*(x))^2\, dx\right] \\ &\quad = \frac{1}{2}\sum_k \left[\log(1 + c\lambda_k) + \frac{c\lambda_k}{1 + c\lambda_k}\right]. \end{aligned} \tag{12}$$

Here λ_k, $k = 1, 2, \ldots, \infty$, are the eigenvalues of the process on the interval $[0, 1]$. Specializing to the Wiener process, we get $\lambda_k = 1/\pi^2(k - \frac{1}{2})^2$ for $k = 1, 2, 3 \ldots$. Using

$$\frac{1}{2}\sum_{k=1}^{\infty} \log\left(1 + \frac{c}{\pi^2\left(k - \frac{1}{2}\right)^2}\right) = \frac{1}{2}\log\cosh\left(\sqrt{c}\right)$$

and

$$\frac{1}{2}\sum_{k=1}^{\infty} \frac{c}{c + \pi^2\left(k - \frac{1}{2}\right)^2} = \frac{\sqrt{c}}{4}\tanh\sqrt{c}$$

and setting $\alpha = \frac{1}{2}$ in the lower bound and $\alpha = 1$ in the upper bound, we get

$$\begin{aligned} \frac{1}{2}\log\cosh\left(\frac{\sqrt{n}}{2\sigma}\right) + \frac{\sqrt{n}}{8\sigma}\tanh\left(\frac{\sqrt{n}}{2\sigma}\right) &\le I(\Theta^*; Y^n) \\ &\le \frac{1}{2}\log\cosh\left(\frac{\sqrt{n}}{\sigma}\right) + \frac{\sqrt{n}}{4\sigma}\tanh\left(\frac{\sqrt{n}}{\sigma}\right). \end{aligned}$$

Hence, asymptotically,

$$\frac{3\sqrt{n}}{8\sigma}(1+o(1)) \le I(\Theta^*;Y^n) \le \frac{3\sqrt{n}}{4\sigma}(1+o(1)).$$

Notice that, in the preceding examples, it was always the case that, asymptotically, the best bounds were obtained with the value $\alpha = \frac{1}{2}$. In general, for large n, the value of the Laplace transform

$$\int_\Theta d\mu(\tilde{\theta}) \exp\bigl[-n(1-\alpha)I_\alpha(P_{\theta^*}\|P_{\tilde{\theta}})\bigr]$$

in the lower bound of Theorem 2 is largely determined by those $\tilde{\theta}$ such that $I_\alpha(P_{\theta^*}\|P_{\tilde{\theta}})$ is near 0, that is, such that P_{θ^*} is close to $P_{\tilde{\theta}}$. The same also holds for the corresponding Laplace transform

$$\int_\Theta d\mu(\tilde{\theta}) \exp\bigl[-nI_1(P_{\theta^*}\|P_{\tilde{\theta}})\bigr]$$

in the upper bound. However, it can be shown that, as the distributions P and Q become close, in the sense that $dP/dQ \to 1$ uniformly, then

$$\frac{I_1(P\|Q)}{(1-\alpha)I_\alpha(P\|Q)} \to \frac{1}{\alpha(1-\alpha)}.$$

Hence, we might expect to very often get the best asymptotic lower bound in Theorem 2 by choosing $\alpha = \frac{1}{2}$, so as to minimize $1/\alpha(1-\alpha)$. This choice also has another desirable property, since, as mentioned previously, for $\alpha = \frac{1}{2}$, the distance D_α used in Corollary 1 and Theorem 3 is then the squared Hellinger distance, which has some nice metric properties that we will exploit later in applications of the bounds. For these reasons, in what follows, we will for simplicity restrict ourselves to the case $\alpha = \frac{1}{2}$, using the notation

$$D_{1/2}(P,Q) = D_{HL}^2(P,Q).$$

6. Bounds on the cumulative risk for countable Θ. Recall that we have assumed that, for all distinct $\theta, \theta^* \in \Theta$, the conditional densities dP_θ and dP_{θ^*} differ on a set of positive measure and hence $D_{HL}(P_\theta, P_{\theta^*}) > 0$. We can make this assumption without essential loss of generality, since, otherwise, we can replace Θ by a set of equivalence classes with the property that $\theta \equiv \theta^*$ iff $dP_\theta = dP_{\theta^*}$ (except on a set of measure 0) in a natural way, without changing the risks we are interested in calculating.

Suppose Θ is countable, say $\Theta = \{\theta_i\}$. Let $H(\Theta^*) = -\sum_i \mu(\theta_i)\log\,\mu(\theta_i)$ denote the entropy of the random variable Θ^*, distributed according to the prior measure μ. The entropy of Θ^* may be infinite. Then

COROLLARY 2. *For all n, $R_{n,\mu}^{\text{Bayes}} = I(\Theta;Y^n) \le H(\Theta^*)$ and*

$$\lim_{n\to\infty} R_{n,\mu}^{\text{Bayes}} = H(\Theta^*).$$

PROOF. Recall that $R^{\text{Bayes}}_{n,\mu} = I(\Theta^*; Y^n)$. If $H(\Theta^*)$ is infinite, then clearly

$$\limsup_{n\to\infty} I(\Theta; Y^n) \le H(\Theta^*).$$

Assume $H(\Theta^*)$ is finite. Let

$$H(\Theta^*|Y^n) = -\int_{Y^n} dM_{n,\mu}(y^n) \sum_i \mu(\theta_i|y^n) \log\, \mu(\theta_i|y^n),$$

the conditional entropy of Θ given Y^n. Note that this quantity is nonnegative. When $H(\Theta)$ is finite, it is easily verified that

$$I(\Theta^*; Y^n) = H(\Theta^*) - H(\Theta^*|Y^n)$$

(see, e.g., [18], page 20) and thus $\limsup_{n\to\infty} I(\Theta^*; Y^n) \le H(\Theta^*)$ in this case as well.

For the lower bound, using Theorem 2 with $\alpha = \frac{1}{2}$ and Fatou's lemma,

$$\begin{aligned}
\liminf_{n\to\infty} I(\Theta^*; Y^n) &\ge \liminf_{n\to\infty} -\sum_i \mu(\theta_i) \log \sum_j \mu(\theta_j) \exp\left[-\frac{n}{2} D^2_{HL}(P_{\theta_i}, P_{\theta_j})\right] \\
&\ge -\sum_i \mu(\theta_i) \liminf_{n\to\infty} \log \sum_j \mu(\theta_j) \exp\left[-\frac{n}{2} D^2_{HL}(P_{\theta_i}, P_{\theta_j})\right] \\
&= -\sum_i \mu(\theta_i) \log\, \mu(\theta_i) \\
&= H(\Theta^*). \qquad \square
\end{aligned}$$

This result generalizes a similar result in [16] (Corollary 1) by removing the additional conditions assumed there. More general results, including the preceding corollary, follow from results in Pinsker's book [45] (see also [4]). Applying Theorem 1 and taking the supremum over μ in Corollary 2, it follows that if Θ is finite, then, for all n, $R^{\text{minimax}}_n \le \log|\Theta|$ and $\lim_{n\to\infty} R^{\text{minimax}}_n = \log|\Theta|$. It also follows that if Θ is infinite, then $\lim_{n\to\infty} R^{\text{minimax}}_n = \infty$.

In the case that Θ is finite, the results of Rényi [48] show further that the difference $I(\Theta^*; Y^n) - H(\Theta^*)$ converges to 0 exponentially fast in n. We also obtain this result as follows.

COROLLARY 3. *For all n,*

$$H(\Theta^*) - I(\Theta^*; Y^n) \le (|\Theta| - 1)\left(\max_{1\le i<j\le|\Theta|} \int_Y \sqrt{dP_{\theta_i}\, dP_{\theta_j}}\right)^n.$$

PROOF. From Theorem 2,

$$
\begin{aligned}
I(\Theta^*; Y^n) &\geq -\sum_i \mu(\theta_i) \log \sum_j \mu(\theta_j) \left(\int_Y \sqrt{dP_{\theta_i}\, dP_{\theta_j}} \right)^n \\
&= -\sum_i \mu(\theta_i) \log\, \mu(\theta_i) \\
&\quad - \sum_i \mu(\theta_i) \log \left[1 + \sum_{j \neq i} \frac{\mu(\theta_j)}{\mu(\theta_i)} \left(\int_Y \sqrt{dP_{\theta_i}\, dP_{\theta_j}} \right)^n \right] \\
&\geq H(\Theta^*) - \sum_i \sum_{j \neq i} \mu(\theta_j) \left(\int_Y \sqrt{dP_{\theta_i}\, dP_{\theta_j}} \right)^n \\
&\geq H(\Theta^*) - (|\Theta| - 1) \left(\max_{1 \leq i < j \leq |\Theta|} \int_Y \sqrt{dP_{\theta_i}\, dP_{\theta_j}} \right)^n,
\end{aligned}
$$

where the second inequality follows from $-\log(1 + x) \geq -x$. □

Assuming as before that the densities dP_{θ_i} and dP_{θ_j} are different for $j \neq i$, an application of the Cauchy inequality yields $\int_Y \sqrt{dP_{\theta_i}\, dP_{\theta_j}} < 1$ for $j \neq i$. Hence, the corollary shows exponential convergence.

Finally, let us note that Theorem 2 and Corollary 1 can also be used to characterize the mutual information between Θ^* and Y^n (Bayes risk) in the general case when Θ is uncountably infinite but finite dimensional. This was demonstrated in [33]. Here, in the sequel, we focus instead on the minimax risk.

7. Bounds on minimax risk using covering and packing numbers, and metric entropy. For each $\theta^*, \theta \in \Theta$, let

$$h(\theta^*, \theta) = D_{HL}(P_{\theta^*}, P_\theta).$$

As mentioned previously, we assume that, for distinct states of Nature $\theta, \theta^* \in \Theta$, the conditional distributions P_θ and P_{θ^*} differ on a set of positive measure. Under this assumption, (Θ, h) is a metric space. We show how bounds on the minimax risk can be obtained by looking at properties of this metric space. These are the packing and covering numbers, and the associated metric entropy, introduced by Kolmogorov and Tikhomirov in [37] and commonly used in the theory of empirical processes (see, e.g., [12], [22], [27] and [46]).

For the following definitions, let (S, ρ) be any complete separable metric space.

DEFINITION 1 (Metric entropy, also called Kolmogorov ε-entropy [37]). A partition Π of S is a collection $\{\pi_i\}$ of Borel subsets of S that are pairwise disjoint and whose union is S. The diameter of a set $A \subseteq S$ is given by $\operatorname{diam}(A) = \sup_{x, y \in A} \rho(x, y)$. The diameter of a partition is the supremum of the diameters of the sets in the partition. For $\varepsilon > 0$, we denote by $\mathscr{D}_\varepsilon(S, \rho)$

the cardinality of the smallest finite partition of S of diameter at most ε, or we use ∞ if no such finite partition exists. The metric entropy of (S, ρ) is defined by

$$\mathscr{H}_\varepsilon(S, \rho) = \log \mathscr{D}_\varepsilon(S, \rho).$$

We say S is *totally bounded* if $\mathscr{D}_\varepsilon(S, \rho) < \infty$ for all $\varepsilon > 0$.

DEFINITION 2 (Packing and covering numbers). For $\varepsilon > 0$, an ε-cover of S is a subset $A \subseteq S$ such that for all $x \in S$ there exists a $y \in A$ with $\rho(x, y) \leq \varepsilon$. We denote by $\mathscr{N}_\varepsilon(S, \rho)$ the cardinality of the smallest finite ε-cover of S, or we use ∞ if no such finite cover exists. For $\varepsilon > 0$, an *ε-separated subset* of S is a subset $A \subseteq S$ such that, for all distinct $x, y \in A$, $\rho(x, y) > \varepsilon$. We denote by $\mathscr{M}_\varepsilon(S, \rho)$ the cardinality of the largest finite ε-separated subset of S. This quantity is infinity if arbitrarily large such sets exist.

The following lemma is easily verified [37].

LEMMA 6. *For any* $\varepsilon > 0$,

$$\mathscr{M}_{2\varepsilon}(S, \rho) \leq \mathscr{D}_{2\varepsilon}(S, \rho) \leq \mathscr{N}_\varepsilon(S, \rho) \leq \mathscr{M}_\varepsilon(S, \rho).$$

It follows that the metric entropy $\mathscr{H}_\varepsilon$ (and the condition defining total boundedness) can also be defined using either the packing or the covering numbers in place of $\mathscr{D}_\varepsilon$, to within a constant factor in ε.

Kolmogorov and Tikhomirov also introduced an abstract notion of the dimension of a metric space in their seminal paper [37]. In the following, the metric ρ is omitted from the notation, being understood from the context.

DEFINITION 3. The *upper* and *lower metric dimensions* [37] of S are defined by

$$\overline{\dim}(S) = \limsup_{\varepsilon \to 0} \frac{\mathscr{H}_\varepsilon(S)}{\log(1/\varepsilon)}$$

and

$$\underline{\dim}(S) = \liminf_{\varepsilon \to 0} \frac{\mathscr{H}_\varepsilon(S)}{\log(1/\varepsilon)},$$

respectively. When $\overline{\dim}(S) = \underline{\dim}(S)$, then this value is denoted $\dim(S)$ and called the *metric dimension* of S. Thus,

$$\dim(S) = \lim_{\varepsilon \to 0} \frac{\mathscr{H}_\varepsilon(S)}{\log(1/\varepsilon)}.$$

Using the results given in the theorems from Section 4, with $\alpha = \frac{1}{2}$, we can obtain bounds on the minimax risk R_n^{minimax} in terms of the metric entropy of

the space (Θ, h). For every $\varepsilon > 0$, let

$$b(\varepsilon) = \sup\left\{ \frac{D_{KL}(P_{\tilde{\theta}} \| P_{\theta^*})}{D_{HL}^2(P_{\tilde{\theta}}, P_{\theta^*})} : \tilde{\theta}, \theta^* \in \Theta \text{ and } D_{HL}^2(P_{\tilde{\theta}}, P_{\theta^*}) \le \varepsilon \right\}.$$

Let

$$R_{1,\rho_{1+\lambda}}^{\text{minimax}} = \inf_{\hat{P}} \sup_{\theta^* \in \Theta} \int (dP_{\theta^*})^{1+\lambda} (d\hat{P})^{-\lambda}.$$

This is the minimax analog of $R_{1,\mu,\rho_{1+\lambda}}^{\text{Bayes}}$, used in Section 4.2 to obtain general bounds on the Bayes risk. It is the minimax risk for a game much like the one we are studying, except that the relative entropy loss is replaced by the $(1 + \lambda)$-affinity loss, and we have fixed the number n of observations to 1.

LEMMA 7. *Assume* (Θ, h) *is totally bounded. Then, for all* $n \ge 1$,

1.

$$R_n^{\text{minimax}} \ge \sup_{\varepsilon \ge 0} \left\{ -\log\left(\frac{1}{\mathscr{M}_\varepsilon(\Theta, h)} + \exp\left(-\frac{n\varepsilon^2}{2} \right) \right) \right\}$$

$$\ge \sup_{\varepsilon \ge 0} \min\left\{ \mathscr{K}_\varepsilon(\Theta, h), \frac{n\varepsilon^2}{8} \right\} - \log 2$$

and

2.

$$R_n^{\text{minimax}} \le \inf_{\varepsilon \ge 0} \left\{ \mathscr{K}_\varepsilon(\Theta, h) + b(\varepsilon) n\varepsilon^2 \right\} \le \inf_{\varepsilon \ge 0} \left\{ \mathscr{K}_\varepsilon(\Theta, h) + b_{1/2}(\Theta) n\varepsilon^2 \right\}.$$

Furthermore, for any $\lambda > 0$ *such that* $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$,

$$R_n^{\text{minimax}} \le \inf_{\varepsilon \ge 0} \left\{ \mathscr{K}_\varepsilon(\Theta, h) + \frac{(1 + o(1)) 4\varepsilon^2 n \log n}{\lambda} \right\} + R_{1,\rho_{1+\lambda}}^{\text{minimax}} + o(1),$$

where in each case $o(1)$ *is a function* $f(n)$ *such that* $f(n) \to 0$ *as* $n \to \infty$.

PROOF. To establish the first inequality of part 1, let $A = \{\theta_1, \ldots, \theta_M\}$ be an ε-separated subset of Θ of maximal size and let μ be the discrete prior distribution on Θ that is uniform over the elements of A. Using Theorem 1 and Corollary 1, we have

$$R_n^{\text{minimax}} \ge R_{n,\mu}^{\text{Bayes}}$$

$$\ge -\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-\frac{nh^2(\theta^*, \tilde{\theta})}{2} \right]$$

$$= -\frac{1}{M} \sum_{i=1}^M \log \frac{1}{M} \sum_{j=1}^M \exp\left[-\frac{nh^2(\theta_i, \theta_j)}{2} \right]$$

$$\geq \log M - \log\left(1 + (M-1)\exp\left(-\frac{n\varepsilon^2}{2}\right)\right)$$

$$\geq -\log\left(\frac{1}{M} + \exp\left(-\frac{n\varepsilon^2}{2}\right)\right).$$

Since this holds for all ε, it follows that

$$R_n^{\text{minimax}} \geq \sup_{\varepsilon \geq 0}\left\{-\log\left(\frac{1}{\mathscr{M}_\varepsilon(\Theta, h)} + \exp\left(-\frac{n\varepsilon^2}{2}\right)\right)\right\}.$$

To complete the proof of part 1, simply note that

$$-\log(x+y) \geq -\log(2\max(x,y)) = -\log 2 + \min\{-\log x, -\log y\}.$$

It follows that

$$R_n^{\text{minimax}} \geq \sup_{\varepsilon \geq 0} \min\left\{\log \mathscr{M}_\varepsilon(\Theta, h), \frac{n\varepsilon^2}{2}\right\} - \log 2.$$

Since $\mathscr{K}_{2\varepsilon} = \log \mathscr{D}_{2\varepsilon} \leq \log \mathscr{M}_\varepsilon$, replacing ε with $\varepsilon/2$, the second inequality follows.

We now turn to the upper bounds in part 2. Let $\Pi = \{\pi_1, \ldots, \pi_M\}$ be any partition of Θ of diameter at most ε. For any prior measure μ on Θ, let $\mu_i = \mu(\pi_i)$. Then we use Theorem 1 and the upper bound given in Theorem 2 as follows:

$$\begin{aligned}
R_n^{\text{minimax}} &= \sup_\mu R_{n,\mu}^{\text{Bayes}} \\
&\leq \sup_\mu \left\{-\int_\Theta d\mu(\theta^*) \log \int_\Theta d\mu(\tilde{\theta}) \exp\left[-nD_{KL}(P_{\theta^*} \| P_{\tilde{\theta}})\right]\right\} \\
&= \sup_\mu \left\{-\sum_i \mu_i \int_{\pi_i} \frac{d\mu(\theta^*)}{\mu_i} \log \sum_j \mu_j \int_{\pi_j} \frac{d\mu(\tilde{\theta})}{\mu_j} \exp\left[-nD_{KL}(P_{\theta^*} \| P_{\tilde{\theta}})\right]\right\} \\
&\leq \sup_\mu \left\{-\sum_i \mu_i \log\left(\mu_i \exp\left[-b(\varepsilon) n\varepsilon^2\right]\right)\right\} \\
&= \sup_\mu \left\{-\sum_i \mu_i \log \mu_i\right\} + b(\varepsilon) n\varepsilon^2 \\
&= \log M + b(\varepsilon) n\varepsilon^2.
\end{aligned}$$

The second inequality follows by ignoring all but the ith term in the inner sum whenever the index on the outer sum is i and noting that, because the diameter of π_i is at most ε,

$$D_{KL}(P_{\theta^*} \| P_{\tilde{\theta}}) \leq b(\varepsilon) h^2(\theta^*, \tilde{\theta}) \leq b(\varepsilon)\varepsilon^2$$

for all $\theta^*, \tilde{\theta} \in \pi_i$. The last equality follows from the fact that the entropy of a finite distribution is maximal for the uniform distribution. Since the particular partition of diameter ε can be chosen arbitrarily in the preceding chain of inequalities, it follows that $R_n^{\text{minimax}} \le \mathscr{H}_\varepsilon(\Theta, h) + b(\varepsilon) n\varepsilon^2$ for any ε. This establishes the first inequality of part 2. The second inequality follows since $b(\varepsilon) \le b_{1/2}(\Theta)$ for all ε. The third inequality, but with $\sup_\mu R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}}$ in place of $R^{\text{minimax}}_{1,\rho_{1+\lambda}}$, follows by an argument similar to that used for the first inequality, using Theorem 3. Since maximin $\le$ minimax always, we have $\sup_\mu R^{\text{Bayes}}_{1,\mu,\rho_{1+\lambda}} \le R^{\text{minimax}}_{1,\rho_{1+\lambda}}$, and from this we obtain the result stated in the theorem. □

The method used in obtaining the upper bound in the preceding result is a familiar one (see, e.g., [4] and [31]). The method for obtaining the lower bound by choosing a discrete prior on a well-separated set of θ is also similar in many respects to standard lower-bound methods, such as those that use Fano's inequality or Assouad's lemma (see, e.g., [9], [11] and [55]), but the method is particularly clean in the present framework, giving a fairly good match to the upper bound.

In some cases, $\mathscr{H}_\varepsilon$ may not be a continuous function of ε, and even so it may not be obvious what kinds of asymptotic bounds on the risk R_n^{minimax} are implied by Lemma 7. For such cases, we make the following definitions.

Fix a totally bounded Θ and let $f_l(x)$ and $f_u(x)$ be any continuous, nondecreasing, unbounded functions on $(0, \infty)$ such that

$$\liminf_{\varepsilon \to 0} \frac{\mathscr{H}_\varepsilon(\Theta, h)}{f_l(1/\varepsilon)} \ge 1 \quad \text{and} \quad \limsup_{\varepsilon \to 0} \frac{\mathscr{H}_\varepsilon(\Theta, h)}{f_u(1/\varepsilon)} \le 1. \tag{13}$$

For every positive real n, let $\varepsilon_l(n)$ be the unique solution to the equation $f_l(1/\varepsilon) = n\varepsilon^2$, and let $\varepsilon_u(n)$ be the unique solution to the equation $f_u(1/\varepsilon) = n\varepsilon^2$. Let

$$\begin{aligned} F_l(n) &= f_l\left(\frac{1}{\varepsilon_l(n)}\right) = n\varepsilon_l^2(n), \\ F_u(n) &= f_u\left(\frac{1}{\varepsilon_u(n)}\right) = n\varepsilon_u^2(n). \end{aligned} \tag{14}$$

Then we have the following lemma.

LEMMA 8. *For every integer* $n \ge 1$,

1.

$$\liminf_{n \to \infty} \frac{R_n^{\text{minimax}}}{F_l(n/8)} \ge 1.$$

2. *If* $\lim_{\varepsilon \to 0} b(\varepsilon) < \infty$, *then, for any function* $g(n)$ *such that* $g(n) \to \infty$ *as* $n \to \infty$,

$$\limsup_{n \to \infty} \frac{R_n^{\text{minimax}}}{F_u(ng(n))} \leq 1,$$

and if there exists $\lambda > 0$ *such that* $R^{\text{minimax}}_{1,\, \rho_{1+\lambda}} < \infty$, *then*

$$\limsup_{n \to \infty} \frac{R_n^{\text{minimax}}}{F_u(ng(n)\log n)} \leq 1.$$

PROOF. Using Lemma 7 and the definitions of f_l and F_l, we have

$$\begin{aligned}
\liminf_{n \to \infty} \frac{R_n^{\text{minimax}}}{F_l(n/8)} &\geq \liminf_{n \to \infty} \frac{\min\left(\mathscr{H}_{\varepsilon_l(n/8)}, (n/8)\varepsilon_l^2(n/8)\right)}{F_l(n/8)} \\
&\geq \min\left(\liminf_{n \to \infty} \frac{\mathscr{H}_{\varepsilon_l(n/8)}}{F_l(n/8)}, \liminf_{n \to \infty} \frac{(n/8)\varepsilon_l^2(n/8)}{F_l(n/8)} \right) \\
&\geq \min\left(\liminf_{n \to \infty} \frac{f_l(1/\varepsilon_l(n/8))}{F_l(n/8)}, 1 \right) \\
&= 1.
\end{aligned}$$

Now let $N = N(n) = ng(n)$. Let $\lim_{\varepsilon \to 0} b(\varepsilon) = b < \infty$. Then we also have

$$\begin{aligned}
\limsup_{n \to \infty} \frac{R_n^{\text{minimax}}}{F_u(ng(n))} &\leq \limsup_{n \to \infty} \frac{\mathscr{H}_{\varepsilon_u(N)} + bn\varepsilon_u^2(N)}{F_u(N)} \\
&\leq \limsup_{n \to \infty} \left(\frac{f_u(1/\varepsilon_u(N))}{F_u(N)} + \frac{b}{g(n)} \right) \\
&= 1 + \limsup_{n \to \infty} \frac{b}{g(n)} \\
&= 1.
\end{aligned}$$

The last inequality follows similarly, using the last inequality of Lemma 7. □

Essentially, when $F_l(n)$ and $F_u(n \log n)$ are close asymptotically, as can often be arranged, this lemma shows that asymptotic growth rates for R_n^{minimax} can be obtained by "solving" the equation $\mathscr{H}_\varepsilon(\Theta, h) = n\varepsilon^2$. This general approach was developed by Le Cam [40] and Birgé [10, 11] in the context of other loss functions. We illustrate it now by applying Lemma 7 to establish a simple relationship between the metric dimension of (Θ, h) and the asymptotic growth rate of the minimax risk R_n^{minimax}.

THEOREM 4. *Assume there exists* $\lambda > 0$ *such that* $R^{\text{minimax}}_{1,\, \rho_{1+\lambda}} < \infty$.

1. *If* Θ *is finite, then*

$$R_n^{\text{minimax}} \to \log|\Theta| \quad \text{as } n \to \infty.$$

2. *If* $\dim(\Theta, h) = 0$, *then*

$$R_n^{\text{minimax}} \in o(\log n).$$

3. *If* $\dim(\Theta, h) = D$ *where* $0 < D < \infty$, *then*

$$R_n^{\text{minimax}} \sim \frac{D}{2}\log n.$$

4. *If* $\dim(\Theta, h) = \infty$ *but* (Θ, h) *is totally bounded, then*

$$R_n^{\text{minimax}} \in O(n) \quad \textit{but} \quad R_n^{\text{minimax}} \notin O(\log n).$$

5. *If* (Θ, h) *is not totally bounded, then*

$$\textit{if } R_1^{\text{minimax}} < \infty \textit{ then } R_n^{\text{minimax}} \asymp n \textit{ else } R_n^{\text{minimax}} = \infty \quad \textit{for all } n.$$

Actually, only the upper bounds in parts 2 and 3 of the theorem require the assumption that there exists $\lambda > 0$ such that $R_{1,\,\rho_{1+\lambda}}^{\text{minimax}} < \infty$.

PROOF. As mentioned after Corollary 2, part 1 follows from that corollary and Theorem 1. Parts 2 and 3 and the second half of part 4 follow easily from Lemma 8 by plugging in the appropriate rates for f_l and f_u and solving for F_l and F_u. We illustrate this for part 3; the others are similar. Since $\dim(\Theta, h) = D$ where $0 < D < \infty$, we may choose

$$f_l(x) = f_u(x) = D \log x.$$

Solving $D\log(1/\varepsilon) = n\varepsilon^2$, we find that

$$\varepsilon_l(n) = \varepsilon_u(n) \sim \sqrt{\frac{D}{2n}\log n}\,,$$

and hence by (14)

$$F_l(n) = F_u(n) \sim \frac{D}{2}\log n.$$

From the lower bound of Lemma 8, it follows that

$$\liminf_{n\to\infty} \frac{R_n^{\text{minimax}}}{(D/2)\log(n/8)} \geq 1.$$

Let $g(n) = \log n$. From the second upper bound of Lemma 8, it follows that

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{(D/2)\log(n\log^2 n)} \leq 1.$$

The result in part 3 follows.

To verify the first half of part 4 and part 5, first note that the minimax risk R_n^{minimax} is nondecreasing in n. Furthermore, if R_n^{minimax} is finite, then it can

grow at most linearly, as is seen in the following series of inequalities:

$$\begin{aligned} R_n^{\text{minimax}} &= \inf_{\text{dist. } R \text{ on } Y^n} \sup_{\theta \in \Theta} D_{KL}(P_\theta^n \| R) \\ &\le \inf_{\text{dist. } Q \text{ on } Y} \sup_{\theta \in \Theta} D_{KL}(P_\theta^n \| Q^n) \\ &= n \inf_{\text{dist. } Q \text{ on } Y} \sup_{\theta \in \Theta} D_{KL}(P_\theta \| Q) \\ &= n R_1^{\text{minimax}}. \end{aligned}$$

Hence, for any Θ, either $R_n^{\text{minimax}} = \infty$ for all n or R_n^{minimax} is finite and bounded by nR_1^{minimax} for all n.

If (Θ, h) is totally bounded, then we must have $R_1^{\text{minimax}} < \infty$ by part 2 of Lemma 7. Hence, $R_n^{\text{minimax}} \in O(n)$ in this case.

If (Θ, h) is not totally bounded, then $\mathscr{M}_{\varepsilon_0}(\Theta, h)$ is infinite for some $\varepsilon_0 > 0$. In this case, the first lower bound from part 1 of Lemma 7 shows that

$$R_n^{\text{minimax}} \ge \frac{n\varepsilon_0^2}{2}$$

for all $n \ge 1$. Hence, $R_n^{\text{minimax}} \asymp n$ in this case, if it is not infinite. □

The preceding theorem generalizes the standard results for the case when Θ is a finite-dimensional vector-valued parameter space, but does not give much information about the infinite-dimensional case. As mentioned previously, several authors have studied the minimax risk in infinite-dimensional (i.e., nonparametric) density estimation under other loss functions and related it to the metric entropy of Θ under Hellinger distance. Using Lemma 7, we can give a general characterization of the asymptotic growth rate of the minimax risk R_n^{minimax} in terms of the metric entropy of Θ in most infinite-dimensional cases as well.

In infinite-dimensional cases, instead of growing like $D\log(1/\varepsilon)$, the metric entropy $\mathscr{H}_\varepsilon(\Theta, h)$ usually grows like $(1/\varepsilon)^\alpha \log(1/\varepsilon)^\beta$ for some $\alpha > 0$ and/or $\beta > 1$. A classical example is the following. Let Θ be the Lipschitz class $F_{p,\delta}(C, L)$ of densities on $Y = [0, 1]$ satisfying $\sup_{y \in [0,1]} |dP_\theta(y)| \le C$ and having derivatives $dP_\theta^{(k)}(y)$ of order $k \le p$ with the Lipschitz condition on the pth derivative $|dP_\theta^{(p)}(y) - dP_\theta^{(p)}(y')| \le L|y - y'|^\delta$ for $y, y' \in [0, 1]$. Since the functions in $F_{p,\delta}(C, L)$ are uniformly bounded, they have an integrable envelope function and hence $R_{1,\rho_{1+\lambda}}^{\text{minimax}} < \infty$ for all $\lambda > 0$. As shown by Barron and Yang [9], a further restriction to uniformly lower-bounded densities makes the Hellinger distance equivalent to the L_2-distance on this class of functions and does not change the metric entropy of this class asymptotically. By a result of Clements [17], the metric entropy of Θ under the L_2-distance is given by $\mathscr{H}_\varepsilon(\Theta, L_2) \asymp \varepsilon^{-1/(p+\delta)}$. Hence, $\mathscr{H}_\varepsilon(\Theta, h) \asymp \varepsilon^{-1/(p+\delta)}$.

The asymptotic growth rate of the minimax risk R_n^{minimax} for the preceding example, and many others, can be determined using the following consequence of Lemma 8.

THEOREM 5. *Assume there exists* $\lambda > 0$ *such that* $R^{\text{minimax}}_{1,\,\rho_{1+\lambda}} < \infty$. *Let* $l(x)$ *be a continuous, nondecreasing function defined on the positive reals such that, for all* $\gamma \geq 0$ *and* $C > 0$,

1.

$$\lim_{x\to\infty} \frac{l\big(Cx(l(x))^{\gamma}\big)}{l(x)} = 1$$

and

2.

$$\lim_{x\to\infty} \frac{l\big(Cx(\log(x))^{\gamma}\big)}{l(x)} = 1.$$

Then

1.

$$\text{If } \mathscr{K}_{\varepsilon}(\Theta, h) \sim l\left(\frac{1}{\varepsilon}\right), \text{ then } R_n^{\text{minimax}} \sim l\big(\sqrt{n}\,\big).$$

2. *If, for some* $\alpha > 0$,

$$\mathscr{K}_{\varepsilon}(\Theta, h) \asymp \left(\frac{1}{\varepsilon}\right)^{\alpha} l\left(\frac{1}{\varepsilon}\right),$$

then

(a)

$$\text{If } \lim_{\varepsilon\to 0} b(\varepsilon) < \infty, \text{ then } R_n^{\text{minimax}} \asymp n^{\alpha/(\alpha+2)}\big[l\big(n^{1/(\alpha+2)}\big)\big]^{2/(\alpha+2)}$$

else

(b)

$$\liminf_{n\to\infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)}\big[l\big(n^{1/(\alpha+2)}\big)\big]^{2/(\alpha+2)}} > 0$$

and

$$\limsup_{n\to\infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)}\big[l\big(n^{1/(\alpha+2)}\big)\big]^{2/(\alpha+2)}(\log n)^{\alpha/(\alpha+2)}} < \infty.$$

PROOF. Consider part 2 first. Since $\mathscr{K}_{\varepsilon}(\Theta, h) \asymp (1/\varepsilon)^{\alpha} l(1/\varepsilon)$, we may choose

$$f_l(x) = ax^{\alpha} l(x) \quad \text{and} \quad f_u(x) = bx^{\alpha} l(x)$$

for suitable constants $0 < a \leq b$. Solving $f_l(x) = n/x^2$, we find that

$$x \sim \left(\frac{N}{l\big(N^{1/(\alpha+2)}\big)}\right)^{1/(\alpha+2)},$$

where $N = n/a$. Here we use property 1 of $l(x)$. Hence,

$$\varepsilon_l(n) \sim \left(\frac{l(N^{1/(\alpha+2)})}{N} \right)^{1/(\alpha+2)},$$

and thus, by (14) and again using property 1 of $l(x)$,

$$F_l(n) \asymp n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)}) \right]^{2/(\alpha+2)}.$$

By similar reasoning,

$$F_u(n) \asymp n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)}) \right]^{2/(\alpha+2)}.$$

From the lower bound of Lemma 8 and property 1, it follows that

$$\liminf_{n \to \infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l(n^{1/(\alpha+2)}) \right]^{2/(\alpha+2)}} > 0.$$

From the second upper bound of Lemma 8, it follows that, for any unbounded, increasing function $g(n)$,

$$\limsup_{n \to \infty} \frac{R_n^{\text{minimax}}}{n^{\alpha/(\alpha+2)} \left[l\left((ng(n)\log n)^{1/(\alpha+2)} \right) \right]^{2/(\alpha+2)} (g(n)\log n)^{\alpha/(\alpha+2)}} < \infty.$$

Part 2(b) follows easily from this, using property 2 of the function $l(x)$. For part 2(a), note that, from the first upper bound of Lemma 8, if $\lim_{\varepsilon \to 0} b(\varepsilon) < \infty$, then the $\log n$ factors can be removed from the preceding lim sup, yielding the desired result. Part 1 follows by a similar argument, essentially by setting $\alpha = 0$ and $a = b = 1$, so that most terms in the denominators of the previous expressions go away, and tracking the lim inf and lim sup more precisely. □

Note that, in finite-dimensional cases, we have $\mathscr{H}_\varepsilon \sim D \log(1/\varepsilon) = l(1/\varepsilon)$, and part 1 of Theorem 5 gives $R_n^{\text{minimax}} \sim l(\sqrt{n}) = (D/2)\log n$, as obtained in the previous theorem. Part 1 generalizes this to infinite-dimensional cases in which, for example, $\mathscr{H}_\varepsilon \sim C(\log(1/\varepsilon))^\beta$ for $\beta > 1$. To illustrate part 2, note that, in the case when Θ is the Lipschitz class described previously with a uniform lower bound on the densities, the condition $\lim_{\varepsilon \to 0} b(\varepsilon) < \infty$ holds, and hence, using Theorem 5 and the fact that $\mathscr{H}_\varepsilon(\Theta, h) \asymp \varepsilon^{-1/(p+\delta)}$, we get

$$R_n^{\text{minimax}} \asymp n^{1/[2(p+\delta)+1]}.$$

Finally, note that, as $\alpha \to \infty$, the lower bounds in Theorem 5 show that R_n^{minimax} approaches a linear growth rate, the fastest possible for finite minimax risk. So this theorem covers all the interesting growth rates.

Theorem 5 is not applicable in all cases. In particular, it can be shown that the condition that $R_{1,\,\rho_{1+\lambda}}^{\text{minimax}} < \infty$ in Theorem 5 and the preceding results of this section cannot be removed. For example, this condition is violated by the Θ defined in Example 1. In this case, (Θ, h) is totally bounded and $R_n^{\text{minimax}} \sim n$, yet $\mathscr{H}_\varepsilon \sim (1/\varepsilon)^2$, which would yield via Theorem 5 an estimated rate of $\sqrt{n}$ for R_n^{minimax}. This is off by a factor of $\sqrt{n}$. Of course, the lower bounds in

Theorem 5 and the preceding results are valid in this and any other case without any special assumptions, but, in this case, we see that they are not tight.

8. Discussion, open problems, further work. We have shown that, under relatively weak assumptions [in particular, whenever there exist a distribution U and a $\lambda > 0$ such that the $(1 + \lambda)$-affinity between P_θ and U is uniformly bounded for all $\theta \in \Theta$], one can obtain explicit bounds on the mutual information $I(\Theta^*; Y^n)$ between the true parameter and the observations in terms of a Laplace transform of the Hellinger distance in Θ, and from these one can obtain bounds on the cumulative minimax risk in estimating a distribution in Θ under relative entropy loss in terms of the metric entropy of Θ with respect to Hellinger distance. In fact, in each case, only the upper bounds depend on the assumptions; the lower bounds hold for any Θ. We also show by example that some assumptions are needed to get the type of general characterizations of the mutual information and minimax risk in terms of the Hellinger distance that we obtain. It remains open to get a useful characterization of these quantities for the cases where our assumptions do not hold, and to get more precise bounds when they do.

In [34] we also show how general bounds on instantaneous risk in estimating a distribution for various other loss functions can be derived in a very simple manner from the bounds on cumulative relative entropy risk. While the resulting bounds are not usually as tight as those obtained by more direct methods for specific Θ, this approach does have the advantage of giving a simple, unified and general treatment to this problem, moreover, one in which no more sophisticated mathematical methods than Jensen's inequality are needed to derive the results. In the future, we hope to further explore the applications of these results to specific estimation problems, such as the "concept learning" or "pattern classification" problems examined in current machine learning and neural network research. Some initial results along these lines can be found in [33], [44] and [57] (see also [24] and [41]).

There are also several other directions for further research one might pursue. Apart from general tightening of the bounds, these include treating the case of nonindependent observations, extending the results giving bounds for individual θ^* in Theorems 2 and 3 to the case where P_{θ^*} is not a distribution in Θ but is "close to" a distribution Θ and giving a more complete characterization of the mutual information $I(\Theta^*; Y^n)$ in terms of the metric entropy properties of Θ for the infinite-dimensional case, as was done for the finite-dimensional case in [33].

APPENDIX

Here we give the proof of Lemma 5.

LEMMA 9. *Assume* $0 < \alpha < 1$ *and* $\lambda > 0$. *Let* P, R *and* U *be any distributions on* Y. *Let* $c_\lambda = \int dP^{1+\lambda}\, dU^{-\lambda}$. *Let* $Q = (1 - \varepsilon)R + \varepsilon U$ *for some* $\varepsilon > 0$

such that $\log\log(1/\varepsilon)/\log(1/\varepsilon) \le \lambda/2$ *and* $\varepsilon \le \exp[-\alpha/(2(1-\alpha))]$. *Then*

$$D_{KL}(P\|Q) \le \frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)} D_\alpha(P,R) + \frac{2\varepsilon\log(1/\varepsilon)}{(1-\alpha)f_\alpha(\varepsilon^2)} + \varepsilon^{\lambda/2} c_\lambda,$$

where

$$f_\alpha(x) = \frac{\alpha + (1-\alpha)x - x^{1-\alpha}}{1-\alpha}.$$

PROOF. We use the easily verified fact that, for $0 < \alpha < 1$ and $0 < x < 1$, $f_\alpha(x)$ is positive and decreasing in x. Let $Y_0 = \{y\colon dP(y) = 0\}$. For $y \in Y - Y_0$, let $S(y) = dQ(y)/dP(y)$ and $T(y) = dU(y)/dP(y)$. Then, using (3) and (4) and the definition of b_α, we have

$$D_{KL}(P\|Q) = \int_{Y-Y_0} dP\, b_\alpha(S) f_\alpha(S) + \int_{Y_0} dQ. \tag{15}$$

Consider two cases for $y \in Y - Y_0$.

1. $S(y) > \varepsilon^2$ or $T(y) > \varepsilon$. Here we note that, since

$$S(y) = \frac{(1-\varepsilon)\, dR(y) + \varepsilon\, dU(y)}{dP(y)} \ge \varepsilon \frac{dU(y)}{dP(y)} = \varepsilon T(y),$$

in either case $S(y) > \varepsilon^2$. Hence,

$$b_\alpha(S(y)) \le b_\alpha(\varepsilon^2) = \frac{\varepsilon^2 + 2\log(1/\varepsilon) - 1}{f_\alpha(\varepsilon^2)} \le \frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)}, \tag{16}$$

since b_α is decreasing.

2. $S(y) \le \varepsilon^2$ and $T(y) \le \varepsilon$. In this case,

$$\begin{aligned} b_\alpha(S(y)) &= \frac{S(y) + \log(1/S(y)) - 1}{f_\alpha(S(y))} \le \frac{\log(1/S(y))}{f_\alpha(S(y))} \\ &\le \frac{\log(1/\varepsilon) + \log(1/T(y))}{f_\alpha(S(y))} \\ &\le \frac{\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)} + \frac{\log(1/T(y))}{f_\alpha(S(y))}, \end{aligned} \tag{17}$$

where in the last inequality we use the fact that $S(y) \le \varepsilon^2$ and $f_\alpha(x)$ is decreasing in x for $0 < x < 1$, and in the previous inequality we use the fact that $S(y) \ge \varepsilon T(y)$ and that $\log(x)$ is increasing.

Let

$$W(\varepsilon) = \int_{y\colon S(y) \le \varepsilon^2 \text{ and } T(y) \le \varepsilon} dP \log\frac{1}{T}.$$

From (15), (16) and (17) it follows that

$$
\begin{aligned}
D_{KL}(P\|Q) &\le \frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)}\int_{Y-Y_0} dP f_\alpha(S) + \int_{Y_0} dQ + W(\varepsilon) \\
&\le \frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)} D_\alpha(P,Q) + W(\varepsilon),
\end{aligned} \tag{18}
$$

since $D_\alpha(P,Q) = \int_{Y-Y_0} dP f_\alpha(S) + \int_{Y_0} dQ$ and $2\log(1/\varepsilon)/f_\alpha(\varepsilon^2) \ge 1$.

Note now that

$$
\begin{aligned}
D_\alpha(P,Q) &= \frac{1}{1-\alpha}\left(1 - \int (dP)^\alpha ((1-\varepsilon)\,dR + \varepsilon\,dU)^{1-\alpha}\right) \\
&\le \frac{1}{1-\alpha}\left(1 - \int (dP)^\alpha ((1-\varepsilon)\,dR)^{1-\alpha}\right) \\
&\le \frac{1}{1-\alpha}\left(1 - \int (dP)^\alpha (dR)^{1-\alpha}\right) + \frac{1}{1-\alpha}\left(1 - (1-\varepsilon)^{1-\alpha}\right) \\
&= D_\alpha(P,R) + \frac{1}{1-\alpha}\left(1 - (1-\varepsilon)^{1-\alpha}\right) \\
&\le D_\alpha(P,R) + \frac{\varepsilon}{1-\alpha}.
\end{aligned}
$$

Hence,

$$
D_{KL}(P\|Q) \le \frac{2\log(1/\varepsilon)}{f_\alpha(\varepsilon^2)} D_\alpha(P,R) + \frac{2\varepsilon\log(1/\varepsilon)}{(1-\alpha) f_\alpha(\varepsilon^2)} + W(\varepsilon). \tag{19}
$$

Finally, note that $T(y) \le \varepsilon$ implies that $(\varepsilon/T(y))^{\lambda/2} \ge 1$ and $\log\log(1/\varepsilon)/\log(1/\varepsilon) \le \lambda/2$ implies that $\log(1/\gamma) \le (1/\gamma)^{\lambda/2}$ for all $\gamma \le \varepsilon$. Hence, when $\log\log(1/\varepsilon)/\log(1/\varepsilon) \le \lambda/2$,

$$
\begin{aligned}
W(\varepsilon) &= \int_{y:\, S(y)\le\varepsilon^2 \text{ and } T(y)\le\varepsilon} dP \log\frac{1}{T} \\
&\le \int_{y:\, S(y)\le\varepsilon^2 \text{ and } T(y)\le\varepsilon} dP \left(\frac{\varepsilon}{T}\right)^{\lambda/2} \log\frac{1}{T} \\
&\le \varepsilon^{\lambda/2} \int_{y:\, S(y)\le\varepsilon^2 \text{ and } T(y)\le\varepsilon} dP \left(\frac{1}{T}\right)^{\lambda} \\
&= \varepsilon^{\lambda/2} \int_{y:\, S(y)\le\varepsilon^2 \text{ and } T(y)\le\varepsilon} (dP)^{1+\lambda} (dU)^{-\lambda} \\
&\le \varepsilon^{\lambda/2} \int (dP)^{1+\lambda} (dU)^{-\lambda} \\
&= \varepsilon^{\lambda/2} c_\lambda.
\end{aligned}
$$

The result follows then from inequality (19). □

Acknowledgments. The authors would like to thank Andrew Barron for inspiring them to work on these problems and for many helpful discussions of these ideas. We also thank Shun-ichi Amari, Meir Feder, Yoav Freund, Michael Kearns, Sebastian Seung, Tom Cover, Bin Yu and Lucien Le Cam for helpful conversations, and Laszlo Györfi and an anonymous referee for their comments on an earlier version of this paper.

REFERENCES

[1] AMARI, S. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–385.

[2] AMARI, S. and MURATA, N. (1993). Statistical theory of learning curves under entropic loss. *Neural Comput.* **5** 140–153.

[3] BARRON, A. (1985). The strong ergodic theorem for densities: generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303.

[4] BARRON, A. (1987). Are Bayes rules consistent in information? In *Open Problems in Communication and Computation* (T. M. Cover and B. Gopinath, eds.) 85–91. Springer-Verlag, New York.

[5] BARRON, A. (1987). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Dept. Statistics, Univ. Illinois Urbana-Champaign.

[6] BARRON, A., CLARKE, B. and HAUSSLER, D. (1993). Information bounds for the risk of Bayesian predictions and the redundancy of universal codes. *Proceedings of the International Symposium on Information Theory*. IEEE Press, New York.

[7] BARRON, A. and COVER, T. (1988). A bound on the financial value of information. *IEEE Trans. Inform. Theory* **34** 1097–1100.

[8] BARRON, A., GYÖRFI, L. and VAN DER MEULEN, E. (1992). Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Trans. Inform. Theory* **38** 1437–1454.

[9] BARRON, A. and YANG, Y. (1995). Information theoretic lower bounds on convergence rates of nonparametric estimators. Unpublished manuscript.

[10] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

[11] BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291.

[12] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

[13] CAMERON, R. H. and MARTIN, W. T. (1944). Transformation of Wiener integrals under translations. *Ann. Math.* **45** 386–396.

[14] CLARKE, B. (1989). Asymptotic cumulative risk and Bayes risk under entropy loss with applications. Ph.D. thesis, Dept. Statistics, Univ. Illinois.

[15] CLARKE, B. and BARRON, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory* **36** 453–471.

[16] CLARKE, B. and BARRON, A. (1994). Jefferys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference* **41** 37–60.

[17] CLEMENTS, G. F. (1963). Entropy of several sets of real-valued functions. *Pacific J. Math.* **13** 1085–1095.

[18] COVER, T. and THOMAS, J. (1991). *Elements of Information Theory*. Wiley, New York.

[19] DAVISSON, L. and LEON-GARCIA, A. (1980). A source matching approach to finding minimax codes. *IEEE Trans. Inform. Theory* **26** 166–174.

[20] DEVROYE, L. and GYÖRFI, L. (1986). *Nonparametric Density Estimation, the L_1 View*. Wiley, New York.

[21] DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–26.

[22] DUDLEY, R. M. (1984). A course on empirical processes. *Lecture Notes in Math*. **1097** 2–142. Springer, New York.
[23] EFROIMOVICH, S. Y. (1980). Information contained in a sequence of observations. *Problems Inform. Transmission* **15** 178–189.
[24] FEDER, M., FREUND, Y. and MANSOUR, Y. (1995). Optimal universal learning and prediction of probabilistic concepts. In *Proceedings of the IEEE Information Theory Conference* 233. IEEE, New York.
[25] GALLAGER, R. (1979). Source coding with side information and universal coding. Technical Report LIDS-P-937, Laboratory for Information and Decision Systems, MIT.
[26] GHOSH, J., GHOSAL, S. and SAMANTA, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related Topics. V* (S. Gupta and J. O. Berger, eds.). Springer, New York.
[27] GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab*. **12** 929–989.
[28] GYÖRFI, L., PÁLI, I. and VAN DER MEULEN, E. (1994). There is no universal source code for an infinite alphabet. *IEEE Trans. Inform. Theory* **40** 267–271.
[29] HASMINSKII, R. and IBRAGIMOV, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist*. **18** 999–1010.
[30] HAUSSLER, D. (1997). A general minimax result for relative entropy. *IEEE Trans. Inform. Theory* **40** 1276–1280.
[31] HAUSSLER, D. and BARRON, A. (1992). How well do Bayes methods work for on-line prediction of $\{+1, -1\}$ values? In *Proceedings of the Third NEC Symposium on Computation and Cognition* 74–100. SIAM, Philadelphia.
[32] HAUSSLER, D., KEARNS, M. and SCHAPIRE, R. E. (1994). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning* **14** 83–113.
[33] HAUSSLER, D. and OPPER, M. (1995). General bounds on the mutual information between a parameter and n conditionally independent observations. In *Proceedings of the Seventh Annual ACM Workshop on Computational Learning Theory* 402–411. ACM Press, New York.
[34] HAUSSLER, D. and OPPER, M. (1996). Mutual information, metric entropy, and risk in estimation of probability distributions. Technical Report UCSC-CRL-96-27, Comput. Res. Lab., Univ. California, Santa Cruz.
[35] IBRAGIMOV, I. and HASMINSKII, R. (1972). On the information in a sample about a parameter. In *Second International Symposium on Information Theory* 295–309. IEEE, New York.
[36] IZENMAN, A. J. (1991). Recent developments in nonparametric density estimation. *J. Amer. Statist. Assoc*. **86** 205–224.
[37] KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1961). ε-entropy and ε-capacity of sets in functional spaces. *Amer. Math. Soc. Trans. Ser. 2* **17** 277–364.
[38] KOMAKI, F. (1994). On asymptotic properties of predictive distributions. Technical Report METR 94-21, Dept. Math. Engrg. Phys., Univ. Tokyo.
[39] LE CAM, L. (1955). An extension of Wald's theory of statistical decision functions. *Ann. Math. Statist*. **26** 69–81.
[40] LECAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.
[41] MEIR, R. and MERHAV, N. (1995). On the stochastic complexity of learning realizable and unrealizable rules. *Machine Learning* **19** 241–261.
[42] MERHAV, N. and FEDER, M. (1995). A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Inform. Theory* **41** 714–722.
[43] OPPER, M. and HAUSSLER, D. (1991). Calculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory* 75–87. Morgan Kaufmann, San Mateo, CA.
[44] OPPER, M. and HAUSSLER, D. (1995). Bounds for predictive errors in the statistical mechanics of in supervised learning. *Phys. Rev. Lett*. **75** 3772–3775.

[45] PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day, Oakland, CA.
[46] POLLARD, D. (1990). *Empirical Processes*: *Theory and Applications*. IMS, Hayward, CA.
[47] RENYI, A. (1960). On measures of entropy and information. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 547–561. Univ. California Press, Berkeley.
[48] RENYI, A. (1964). On the amount of information concerning an unknown parameter in a sequence of observations. *Publ. Math. Inst. Hungar. Acad. Sci.* **9** 617–625.
[49] RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.
[50] RISSANEN, J., SPEED, T. and YU, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Inform. Theory* **38** 315–323.
[51] SYMANZIK, K. (1965). Proof and refinements of an inequality of Feynman. *J. Math. Phys.* **6** 1155–1165.
[52] VAN DEGEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44.
[53] WONG, W. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates for sieve MLE's. *Ann. Statist.* **23** 339–362.
[54] YAMANISHI, K. (1995). A loss bound model for on-line stochastic prediction algorithms. *Inform. Comput.* **119** 39–54.
[55] YU, B. (1996). Lower bounds on expected redundancy for nonparametric classes. *IEEE Trans. Inform. Theory* **42** 272–275.
[56] ZHU, H. and ROHWER, R. (1995). Information geometric measurements of generalization. Technical Report NCRG 4350, Neural Computing Research Group, Aston Univ., England.
[57] HAUSSLER, D. and OPPER, M. (1997). Metric entropy and minimax risk in classification. In *Lecture Notes in Comp. Sci.*: *Studies in Logic and Comp. Sci.* (J. Mycielski, G. Rozenberg and A. Salomaa, eds.) **1261** 212–235. Springer-Verlag, New York.

COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF CALIFORNIA
SANTA CRUZ, CALIFORNIA 95064
E-MAIL: haussler@cse.ucsc.edu

UNIVERSITY OF WÜRZBURG
WÜRZBURG
GERMANY
E-MAIL: opper@physik.uni-wüerzburg.de