The Annals of Statistics
1998, Vol. 26, No. 5, 1800–1825

INFORMATION THEORY AND SUPEREFFICIENCY

BY ANDREW BARRON[1] AND NICOLAS HENGARTNER

Yale University

The asymptotic risk of efficient estimators with Kullback–Leibler loss in smoothly parametrized statistical models is $k/2n$, where k is the parameter dimension and n is the sample size. Under fairly general conditions, we given a simple information-theoretic proof that the set of parameter values where any arbitrary estimator is superefficient is negligible. The proof is based on a result of Rissanen that codes have asymptotic redundancy not smaller than $(k/2)\log n$, except in a set of measure 0.

1. Introduction. In this paper we weave together two stories. One is from statistics concerning the negligibility of the set of superefficient estimation of a parameter originating in the work of Le Cam (1953). The other story is from information theory concerning the negligibility of a set of superefficient data compression due to Rissanen (1984, 1986). The connection between these contexts is the use of the Kullback–Leibler informational divergence (or relative entropy) as the loss function between a parameter and its estimate. Armed with information-theoretic techniques of Rissanen's theory, we obtain a proof of the negligibility of the set of superefficiency for both parameter estimation and data compression.

The object of interest in the present work is the expected Kullback–Leibler loss $D(p\|\hat{p}) = \int p(x)\log(p(x)/\hat{p}(x))\,dx$ between a density p and its estimate $\hat{p}$. In the context of parameter estimation, we consider the Kullback–Leibler loss induced through the plug-in estimators $\hat{p}_n = p_{\hat{\theta}_n}$ of the density and denote it by $D(\theta\|\hat{\theta}_n) = D(p_\theta\|p_{\hat{\theta}_n})$. A Taylor expansion suggests that $D(\theta\|\hat{\theta}_n)$ is approximated by $(1/2)(\hat{\theta}_n - \theta)^t I(\theta)(\hat{\theta}_n - \theta)$, where $I(\theta)$ is the Fisher information matrix. Hence the anticipated behavior for maximum likelihood and Bayes estimators is

$$\lim_{n\to\infty} n\mathbb{E}_\theta\left[D\left(\theta\|\hat{\theta}_n\right)\right] = \frac{k}{2}, \tag{1}$$

where k is the parameter dimension. This motivates our definition of superefficiency.

Received December 1995; revised April 1998.
[1] Supported in part by NSF Grant ECS-9410760.
AMS 1991 *subject classifications*. Primary 62F12, 94A65; secondary 94A29, 62G20.
Key words and phrases. Superefficiency, information theory, data compression, Kullback–Leibler loss.

DEFINITION 1 (Superefficiency and efficiency under Kullback–Leibler risk). We say an estimator sequence $\{\hat{\theta}_n\}$ is *superefficient* at θ with Kullback–Leibler risk if

$$\limsup_{n\to\infty} n\mathbb{E}_\theta D\big(\theta\|\hat{\theta}_n\big) < \frac{k}{2}, \tag{2}$$

and is *efficient* if

$$\limsup_{n\to\infty} n\mathbb{E}_\theta D\big(\theta\|\hat{\theta}_n\big) \le \frac{k}{2} \quad \text{for all } \theta\in\Theta. \tag{3}$$

For more general sequences of loss functions $\{L_n(\theta, \hat{\theta}_n)\}$, Le Cam defined efficiency of an estimator $\hat{\theta}_n$ relative to the risk achieved by some choice of a standard estimator $\hat{\theta}_n^*$ (typically the maximum likelihood estimator or a Bayes estimator with a suitable prior) to mean that

$$\limsup_{n\to\infty}\Big(\mathbb{E}_\theta L_n\big(\theta,\hat{\theta}_n\big) - \mathbb{E}_\theta L_n\big(\theta,\hat{\theta}_n^*\big)\Big) \le 0 \quad \text{for all } \theta\in\Theta,$$

and he defined the set of superefficiency of an estimator to be the set of θ for which

$$\limsup_{n\to\infty}\Big(\mathbb{E}_\theta L_n\big(\theta,\hat{\theta}_n\big) - \mathbb{E}_\theta L_n\big(\theta,\hat{\theta}_n\big)\Big) < 0.$$

Le Cam restricted his considerations of superefficiency only to estimators that also were efficient. He gave regularity conditions on the family of distribution such that for bounded loss functions, the set of superefficiency has Lebesgue measures 0 in $\mathbb{R}^k$. The heart of his proof combines Fatou's lemma and the admissibility of a Bayes estimator used as the standard.

Historical motivation for work on superefficiency was the apparent conflict between the belief in the work of R. A. Fisher that for any statistic $\hat{\theta}_n$ for which $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically Normal$(0, \Sigma(\theta))$, the variance satisfies $\Sigma(\theta) \ge I^{-1}(\theta)$, and the counterexample due to Hodges of a superefficient estimator. Le Cam resolves this conflict by showing that for efficient estimators, superefficiency can only occur in a set of measure 0. An alternative proof is provided by Bahadur (1964) under Cramér-type conditions. Brown (1993) provides a proof that $\{\theta\colon\ \Sigma(\theta) < I^{-1}(\theta)\}$ has measure 0 based on a Cramér–Rao-like inequality he develops for risk with a truncated squared error loss.

Our main result is the following: If there exists a sequence of estimators $\tilde{\theta}_n$ for which $\tilde{\theta}_n - \theta$ is of order $1/\sqrt{n}$ in probability, then $k/2$ is a lower bound for the asymptotic Kullback–Leibler risk of arbitrary estimator sequences, at least for almost every parameter value. The conclusion holds for arbitrary estimator sequences, not just efficient ones as in the theory of Le Cam. The information-theoretic tools we use to derive this lower bound are a variant of results of Rissanen that bounds the measure of the set of parameters for

which $D(p_\theta^n \| q^n) < (1 - \varepsilon)(k/2)\log n$ and the chain rule of information theory that reveals the relationship between estimation and data compression.

In addition, we generalize these results in three directions. First, we want to consider parameter sets that are subsets of arbitrary metric spaces $(\mathscr{M}, d)$, possibly infinite dimensional. Second, we consider the effect of assuming tightness of the sequence $(\tilde{\theta}_n - \theta)/r_n$ for rates r_n other than $n^{-1/2}$. Third, for nonparametric estimation problems, we identify the rate of convergence such that for any estimator sequence, the set of functions that are estimated at a faster rate is negligible.

This level of generality is needed to study the following estimation problem: Suppose that the collection of distributions $\mathscr{P} = \{P_\theta\colon\ \theta \in \Theta\}$ is dominated by some measure ν. We parametrize $\mathscr{P}$ by the probability densities $\Omega = \{p = dP/d\nu\colon\ P \in \mathscr{P}\}$ and view the latter by a subset of $\mathscr{H}$, the metric space comprising all the densities of distributions dominated by ν, endowed with the Hellinger metric $h(p, q) = \{\int(\sqrt{p(x)} - \sqrt{q(x)})^2\, d\nu(x)\}^{1/2}$. If Ω has finite metric dimension, we will talk about *parametric density estimation*, and otherwise, of *nonparametric density estimation*. Parametric density estimation is more general than estimation of the parameter because we do not require the estimator $\hat{p}_n(x)$ of $p_\theta(x)$ to belong to $\mathscr{P}$. In either case, our theory quantifies the performance of estimators $\hat{p}_n(x)$ for $p_\theta(x)$. For parametric density estimation, we identify lower bounds for the risk; for nonparametric density estimation, we identify lower bounds on the rate of convergence.

The outline of the remainder of this paper is as follows: Section 2 presents the key ideas in the context of parameter estimation in smooth parametric families. Section 3 generalized Rissanen's theorem on the set of negligibility of superefficient data compression to give a result that applies to both parametric and nonparametric estimation. In Section 4, we apply the latter result to make conclusions about the negligibility of the set of superefficient estimation. Implications for nonparametric density estimation are presented in Section 5. Finally, a few useful technical lemmas are found in the Appendix.

2. Superefficient parameter estimation. The present work yields conclusions for both parametric and nonparametric estimation. However, we will first study parameter estimation in smooth parametric families of probability densities $\{p_\theta(x)\colon\ \theta \in \Theta\}$ on a sample space $\mathbf{X}$ with a k-dimensional parameter vector θ to expose the key ideas. We assume that the sample $X^n = (X_1, X_2, \ldots, X_n)$ consists of independent and identically P_θ-distributed random variables and the densities are with respect to a fixed sigma-finite measure ν.

We will give conditions for the set of superefficiency to have Lebesgue measure 0 for arbitrary estimators that are not necessarily assumed to be efficient. The information-theoretic proof technique will reveal $k/2$ as the asymptotic lower bound for almost every $\theta \in \Theta$ without recourse to analysis of a standard estimator assumed or known to be efficient. Instead we only require the existence of an estimator that converges at the right rate. This tactic permits us to obtain the bound under the following general condition.

ASSUMPTION A (Estimability in $\{p_\theta(x)\colon \theta \in \Theta\}$). For each bounded subset K of the parameter space Θ in $\mathbb{R}^k$, there exists a sequence of estimators $\tilde{\theta}_n$ based on X^n for which the sequence $\sqrt{n}(\tilde{\theta}_n - \theta)$ is tight (i.e., bounded in probability, $\lim_{c\to\infty}\sup_n P_\theta^n\{\sqrt{n}\|\tilde{\theta}_n - \theta\| > c\} = 0$) for almost every $\theta \in K$. Here $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^k$.

From modern statistical theory, Assumption A holds for i.i.d. sampling from P_θ when there exists, for each compact set K, positive constants $\underline{c}$ and $\bar{c}$ such that the squared Hellinger distance $h^2(p_\theta, p_n) = \int(\sqrt{p_\theta(x)} - \sqrt{p_\eta(x)})^2\nu(dx)$ is bounded by

$$\underline{c}\frac{\|\theta - \eta\|^2}{1 + \|\theta - \eta\|^2} \le h^2(p_\theta, p_\eta) \le \bar{c}\|\theta - \eta\|^2 \tag{4}$$

for all $\theta, \eta \in K$. Under this condition, Ibragimov and Hasminskii (1982), page 54, show that suitable Bayes estimators converge at rate $1/\sqrt{n}$. Alternatively, satisfaction of Assumption A can be deduced from the work of Le Cam, assuming that the mapping $\theta \mapsto p_\theta$ is one to one and that (4) holds locally; that is, for $\theta \in \Theta$, there is a neighborhood $\mathcal{N}_\theta$ and constants $0 < \underline{c}_\theta < \bar{c}_\theta$ such that $\underline{c}_\theta\|\theta - \eta\|^2 \le h^2(p_\theta, p_\eta) \le \bar{c}_\theta\|\theta - \eta\|^2$ for all $\eta \in \mathcal{N}_\theta$. See Le Cam (1986), pages 580 and 608.

To avoid degenerate situations, we make explicit a technical requirement on the Kullback–Leibler divergence.

ASSUMPTION B (Finiteness). There exists an integer m and a distribution Q_0^m with density q_0^m on $\mathcal{X}^m$ such that for every $\theta \in \Theta$ there is a finite constant c_θ such that $D(p_\theta^m\|q_0^m) \le c_\theta$.

If for every θ the Kullback–Leibler divergence $D(\theta\|\eta)$ is bounded for all η in a neighborhood $\mathcal{N}_\theta$, which holds if $D(\theta\|\eta)$ is continuous in its second argument, then the mixtures $q_w(x) = \int p_\theta(x)w(\theta)\,d\theta$ for any positive probability density $w(\theta)$ on Θ satisfy Assumption B with $m = 1$. The details are spelled out in the Appendix.

The information-theoretic tool we use to establish superefficiency is a variant of the results of Rissanen (1984, 1986). The object of interest is the Kullback–Leibler distance $D(p_\theta^n\|q^n)$ between the distribution P_θ^n of the sample (depending on θ) and an arbitrary distribution Q^n (not depending on θ) on the sample space $\mathbf{X}^n$. This quantity arises as n times the per symbol redundancy of universal source codes (for arbitrarily fine quantization) of X^n using Shannon's code based on Q^n. It also arises as the cumulative risk of a sequence of estimators with Kullback–Leibler loss as discussed in Clarke and Barron (1990). As shown in Rissanen (1984), Clarke and Barron (1990) and Barron and Cover (1991), good choices for Q^n yield data compression satisfying

$$\lim_{n\to\infty}\frac{D(p_\theta^n\|q^n)}{\log n} = \frac{k}{2}. \tag{5}$$

The corresponding per symbol redundancy is of order $(k/2)(\log n)/n$.

DEFINITION 2 (Superefficient data compression). For any sequence of probability measures Q^n used to code X^n distributed according to P_θ^n, the *set of superefficient data compression* is the set of $\theta \in \Theta$ for which

$$\limsup_{n\to\infty} \frac{D(p_\theta^n \| q^n)}{\log n} < \frac{k}{2}.$$

A sequence of probability measure Q^n on $\mathbf{X}^n$ is *asymptotically efficient for data compression* if

$$\limsup_{n\to\infty} \frac{D(p_\theta^n \| q^n)}{\log n} \le \frac{k}{2} \quad \text{for all } \theta \in \Theta.$$

Rissanen (1986) gives conditions on the family $\{P_\theta^n\colon \theta \in \Theta\}$ for the set of superefficiency for data compression to have measure 0 for any choice of sequence $\{Q^n\}$. A variant of his result and the implication for superefficiency of parameter estimation are combined in the following proposition.

PROPOSITION 1. *Suppose the family* $\{P_\theta^n\colon \theta \in \Theta\}$ *satisfies Assumption* A. *Then for every sequence of probability distributions* Q^n *on* $\mathbf{X}^n$, *the set*

$$\left\{\theta \in \Theta\colon \limsup_{n\to\infty} \frac{D(p_\theta^n \| q^n)}{\log n} < \frac{k}{2}\right\}$$

has Lebesgue measure 0 *in* Θ.

Moreover, *if in addition to Assumption* A, *Assumption* B *holds*, *then for any sequence of density estimators* $\hat{p}_n(x)$ *based on* X^n, *the set of parameters* θ *for which* $\limsup n\mathbb{E}_\theta D(p_\theta \| \hat{p}_n) < k/2$ *has Lebesgue measure* 0. *In particular*, *for sequences of estimators* $\hat{\theta}_n$ *based on* X^n, *the set*

$$\left\{\theta \in \Theta\colon \limsup_{n\to\infty} nD\big(\theta \| \hat{\theta}_n\big) < \frac{k}{2}\right\}$$

has Lebesgue measure 0.

PROOF. To establish the negligibility of the set of parameter values for which $\limsup_{n\to\infty} D(p_\theta^n \| q^n)/\log n < k/2$, we take $E \subset \mathbf{X}$ and bound the total Kullback–Leibler loss between P_θ^n and any probability distribution Q^n on $\mathbf{X}$ by

$$D(p_\theta^n \| q^n) \ge P_\theta^n(E)\log\frac{P_\theta^n(E)}{Q^n(E)} + P_\theta^n(E^c)\log\frac{P_\theta^n(E^c)}{Q^n(E^c)} \tag{6}$$

$$= P_\theta^n(E)\log\frac{1}{Q^n(E)} + P_\theta^n(E^c)\log\frac{1}{Q^n(E^c)} + P_\theta^n(E)\log P_\theta^n(E) + P_\theta^n(E^c)\log P_\theta^n(E^c) \tag{7}$$

$$\ge P_\theta^n(E)\log\frac{1}{Q^n(E)} - \log 2. \tag{8}$$

Here (6) is a standard inequality between a total Kullback–Leibler loss and its restriction to a partition [this is a consequence of Theorem 4 in Kullback and Leibler (1951)] and inequality (8) is from the bound $\log 2$ on the binary entropy function [see Cover and Thomas (1991), page 15]. With the sequence of estimators $\tilde{\theta}_n$ of Assumption A and $a_n = \log n/\sqrt{n}$, define the sets

$$\mathscr{E}_{\theta,n} := \{x^n \in \mathbf{X}^n : \|\tilde{\theta}_n(x^n) - \theta\| \le a_n\},$$

which we use in inequality (8). Tightness of $\sqrt{n}(\tilde{\theta}_n - \theta)$ together with the dominated convergence theorem imply that, for each subset K of Θ with finite Lebesgue measure, the sequence of sets

$$\mathscr{B}_n = \left\{\theta \in K : P_\theta^n\left[\sqrt{n}\|\tilde{\theta}_n - \theta\| \le \log n\right] < 1 - \varepsilon\right\}$$

has measure tending to 0 with n. To upper bound $Q^n(\mathscr{E}_{\theta,n})$, we show that the set

$$\mathscr{A}_n := \{\theta \in K : Q^n(\mathscr{E}_{\theta,n}) \ge a_n^k \log n\}$$

has Lebesgue measure tending to 0 with n.

Let M_n be the maximal number of disjoint balls of radius a_n having center in $\mathscr{A}_n$, and denote by $\mathscr{C}_n$ the associated collection of centers. The triangle inequality implies

$$\mathscr{A}_n \subset \bigcup_{\theta \in \mathscr{C}_n} \{\eta \in \Theta : \|\eta - \theta\| \le 2a_n\},$$

and therefore

$$\text{Volume}(\mathscr{A}_n) \le V_2(k) a_n^k M_n, \tag{9}$$

where $V_2(k)$ denotes the volume of a ball of radius 2 in $\mathbb{R}^k$, and volume refers to Lebesgue measure. An upper bound for M_n is obtained by noting that the sets $\{\mathscr{E}_{\theta,n} : \theta \in \mathscr{C}_n\}$ are disjoint, and thus

$$1 \ge \sum_{\theta \in \mathscr{C}_n} Q^n(\mathscr{E}_{\theta,n}) \ge M_n a_n^k \log n. \tag{10}$$

Combining (9) and (10) shows that

$$\text{Volume}(\mathscr{A}_n) \le V_2(k) \frac{1}{\log n},$$

which tends to 0 with increasing n. For $\theta \in K - (\mathscr{A}_n \cup \mathscr{B}_n)$, the Kullback–Leibler loss is lower bounded by

$$\begin{aligned} D(P_\theta^n \| Q^n) &\ge P_\theta^n[\mathscr{E}_{\theta,n}] \log\left(\frac{1}{Q^n[\mathscr{E}_{\theta,n}]}\right) - \log 2 \\ &\ge (1 - \varepsilon)\left(k \log \frac{\sqrt{n}}{\log n} - \log\log n\right) - \log 2 \\ &\ge (1 - 2\varepsilon)\frac{k}{2}\log n. \end{aligned}$$

Whence for every $\varepsilon > 0$, the sets

$$\left\{\theta \in K\colon D\big(P_\theta^n \| Q^n\big) \le (1 - 2\varepsilon)\frac{k}{2}\log n \text{ for all } n \ge k\right\}$$

have Lebesgue measure 0 since their volume is bounded by the volume of $\mathscr{A}_n \cup \mathscr{B}_n$ for all $n > k$. The set

$$\left\{\theta \in K\colon \limsup_{n\to\infty} \frac{D\big(P_\theta^n \| Q^n\big)}{\log n} < \frac{k}{2}\right\}$$

has measure 0 since it is contained in the set

$$\bigcup_{j>1}\bigcap_{k>1}\left\{\theta\colon \frac{D\big(P_\theta^n \| Q^n\big)}{\log n} \le \big(1 - j^{-1}\big)\frac{k}{2} \text{ for all } n \ge k\right\}.$$

This holds for all finite measure sets K in Θ, and so the first conclusion follows.

The relationship between the superefficiency for estimation and coding is obtained by the chain rule of information theory. For any sequence of density estimators $\{\hat{p}_l\}$ that depend only on X^l, $l = 1, \ldots, n-1$, consider the probability measure Q^n on $\mathbf{X}^n$ having probability density

$$q^n(x^n) = q_0^m(x^m)\prod_{l=m}^{n-1}\hat{p}_l(x_{l+1}),$$

where $q_0^m(x^m)$ satisfies Assumption B. The chain rule gives

$$\begin{aligned} D\big(P_\theta^n \| q^n\big) &= \mathbb{E}_{P_\theta^n}\left[\log\frac{p_\theta^n(X_1, \ldots, X_n)}{q^n(X_1, \ldots, {}_n)}\right] \\ &= \mathbb{E}_{P_\theta^n}\left[\log\left(\frac{p(X^m|\theta)}{q_0^m(X^m)}\right)\right] + \sum_{l=m}^{n-1}\mathbb{E}_{P_\theta^n}\left[\log\frac{p(X_{m+1}|\theta)}{\hat{p}_l(X_{l+1})}\right] \\ &= \mathbb{E}_{P_\theta^m}\left[\log\left(\frac{p(X^m|\theta)}{q_0^m(X^m)}\right)\right] + \sum_{l=m}^{n-1}\mathbb{E}_\theta D\big(p_\theta \| \hat{p}_l\big). \end{aligned}$$

An individual risk $\mathbb{E}_\theta D(p_\theta \| \hat{p}_l)$ of $c/l + o(1/l)$ corresponds to a cumulative risk $D(p_\theta^n \| q^n)$ of $c \log n + o(\log n)$. Using this choice for the sequence $\{q^n\}$, we see that the set of superefficiency, taken to be the set with $\limsup_n n\mathbb{E}_\theta D(p_\theta \| \hat{p}_n) < k/2$ is contained in the set of θ with $\limsup_n D(p_\theta^n \| q^n)/\log n < k/2$, the latter having asymptotically Lebesgue measure 0. Finally, the conclusion for $D(\theta \| \hat{\theta})$ follows by setting $\hat{p}_l(x) = p_{\hat{\theta}_l}(x)$. □

The original proof of Rissanen (1986) for superefficient data compression required that $\mathbb{P}_\theta[\sqrt{n}\,\|\tilde{\theta}_n - \theta\| > \log n]$ be summable in n uniformly in Θ for an application of the Borel–Cantelli lemma. Inspection of our proof reveals that we only need for each θ the tightness of the sequence $\sqrt{n}\,\|\tilde{\theta}_n - \theta\|/\gamma_n$ for

some sequence $\gamma_n \geq 1$ of smaller order than n^ε for all $\varepsilon > 0$, that is, $\log(\gamma_n) = o(\log n)$. For example, $\lim_n \mathbb{P}_\theta[\sqrt{n}\,\|\tilde{\theta}_n - \theta\| > \log n] = 0$ for each θ suffices to obtain the stated conclusions. Our proof avoids use of the Borel–Cantelli lemma by considering the limit superior.

Other choices of loss functions may be natural for investigation of superefficiency. We find that the Kullback–Leibler loss is most natural for a proof that relates distance between estimates to total distance between joint distributions. Among a broad class of measure of divergence between probability distributions, discussed in Csiszár (1967) or Ali and Silvey (1966), Kullback–Leibler divergence is the unique choice satisfying the chain rule.

As we now discuss, the argument of Le Cam (1953), originally developed for bounded loss functions, also works for unbounded loss functions such as squared error loss or Kullback–Leibler loss to show that the set of parameter values at which a sequence of efficient estimators is superefficient has measure 0. Recall that a sequence of estimators $\{\hat{\theta}_n\}$ is said to be efficient under the Kullback–Leibler loss if it satisfies $\limsup_{n\to\infty} n\mathbb{E}_\theta D(\theta\|\hat{\theta}_n) \leq k/2$.

Consider the truncated Kullback–Leibler loss $L_{n,c}(\theta\|\eta) = \min\{c, nD(\theta\|\eta)\}$, and denote by $\hat{\theta}^\pi_{n,c}$ the Bayes estimate with respect to the prior distribution π and loss function $L_{n,c}(\theta, \eta)$. Under classical regularity conditions on the parametric family p_θ and the prior distribution π, the limit $\lim_{n\to\infty} \mathbb{E} L_{n,c}(\theta, \hat{\theta}^*_n)$ exists and equals $\mathbb{E}\min\{c, \|Z\|^2/2\}$ for $Z \sim \text{Normal}(0, I)$. The latter approaches $k/2$ for large c, and hence, for any $\varepsilon > 0$, there exists a $c < \infty$ such that $\mathbb{E}\min\{c, \|Z\|^2/2\} \geq k/2 - \varepsilon$. With this choice of c and ε, it follows for any sequence of efficient estimators $\hat{\theta}_n$ that

$$\begin{aligned}
0 &\leq \int \left\{ \frac{k}{2} - \limsup_{n\to\infty} \mathbb{E}_\theta D\big(\theta\|\hat{\theta}_n\big) \right\} \pi(d\theta) \\
&\leq \int \left\{ \lim_{n\to\infty} \mathbb{E}_\theta L_{n,c}\big(\theta, \hat{\theta}^\pi_{n,c}\big) + \varepsilon - \limsup_{n\to\infty} \mathbb{E}_\theta L_{n,c}\big(\theta, \hat{\theta}_n\big) \right\} \pi(d\theta) \\
&\leq \limsup_{n\to\infty} \int \mathbb{E}_\theta \left\{ L_{n,c}\big(\theta, \hat{\theta}^\pi_{n,c}\big) - L_{n,c}\big(\theta, \hat{\theta}_n\big) \right\} \pi(d\theta) + \varepsilon,
\end{aligned}$$

where the last inequality is an application of Fatou's lemma since the function $L_{n,c}(\theta, \hat{\theta}^\pi_{n,c}) - L_{n,c}(\theta, \hat{\theta}_n)$ is bounded from below by $-c$. The estimator $\hat{\theta}^\pi_{n,c}$ being Bayes implies that, for each n,

$$\int \mathbb{E}_\theta \left\{ L_{n,c}\big(\theta, \hat{\theta}^\pi_{n,c}\big) - L_{n,c}\big(\theta, \hat{\theta}_n\big) \right\} \pi(d\theta) \leq 0.$$

Whence we have shown that, for arbitrary $\varepsilon > 0$,

$$0 \leq \int \left\{ \frac{k}{2} - \limsup_{n\to\infty} \mathbb{E}_\theta D\big(\theta\|\hat{\theta}_n\big) \right\} \pi(d\theta) \leq \varepsilon.$$

Letting ε tend to 0, we conclude that

$$\int \left\{ \frac{k}{2} - \limsup_{n\to\infty} \mathbb{E}_\theta D\big(\theta\|\hat{\theta}_n\big) \right\} \pi(d\theta) = 0$$

and hence that the set of superefficiency has π-measure 0. It is important to realize that the above argument only works for sequences of estimators assumed to be efficient whereas our Proposition 1 yields the same conclusion, but for arbitrary sequences of estimators.

The constant $k/2$ is sharp. Cencov (1982) shows that, for efficient estimators in smooth parametric families, $n\mathbb{E}D(\hat{\theta}_n\|\theta)$ converges to $k/2$ (this is the reverse order of estimator $\hat{\theta}_n$ and parameter θ to the one we consider). Komaki (1994) shows that, in curved exponential families, Bayes predictive densities have Kullback–Leibler risk $\mathbb{E}_\theta^n D(p_\theta\|\hat{p}_n) = k/2n + O(1/n^2)$. Hartigan (1998) reveals the asymptotics beyond the leading $k/2n$ term. He gives conditions such that for maximum likelihood and Bayes estimates, the truncated Kullback–Leibler risk satisfies $\mathbb{E}_\theta D(p_\theta\|p_{\hat{\theta}}) \wedge \varepsilon = k/2n + C_\theta/n^2 + o(1/n^2)$ for all $\varepsilon > 0$. Furthermore, he identifies the constant C_θ and how it depends on the choice of the prior.

Efficiency may also be addressed in the context of the exponential convergence rate (large deviation principle) for $\mathbb{P}_\theta^n[\|\hat{\theta}_n - \theta\| \geq \varepsilon]$ as $n \to \infty$. Bahadur (1967, 1971) identifies this rate (also related to Kullback–Leibler divergence and to Fisher information in the limit of small ε) and shows that superefficiency is not possible in his setting. For an algorithmic information-theoretic perspective on statistical efficiency, see Vovk (1991).

Merhav and Feder (1995) provide additional insight into the information-theoretic nature of the lower bounds for the total Kullback–Leibler divergence (source coding redundancy). The fundamental quantity is the minimax value $C_n = \min_{q^n} D(p_\theta^n\|q^n)$ which is in agreement with the maximin value $\max_W \min_{q^n} \int D(p_\theta\|q^n)W(d\theta)$, known in information theory as the information capacity. They show that for any (code) distribution Q^n the set of parameter values for which $D(p_\theta^n\|q^n) \leq (1 - \varepsilon)C_n$ has probability, under an asymptotic least favorable (capacity-achieving) probability measure, that vanishes to 0 as n tends to ∞. In particular, as they point out, in smooth parametric families, Jeffreys prior [proportional to $\sqrt{\det(I(\theta))}$] is asymptotically least favorable [cf. Clarke and Barron (1990)] and $C_n \asymp (k/2)\log n$, thereby establishing in a different manner the negligibility of superefficient data compression.

Brown, Low and Zhao (1997) study the superefficiency phenomena in the context of nonparametric regression under squared error loss. They exhibit sequences of estimators whose mean squared error, when divided by the minimax rate of convergence, converges to 0 for each regression function in the considered class. That pointwise convergence to 0 cannot be uniformly at a faster rate. In a related nonparametric problem, we show in Section 5 that the set of functions, when reconvergence occurs at a faster than the minimax rate, is negligible in a sense that will be made precise.

In this paper we focus on the case that $X_1, \ldots, X_n$ are modeled as *independent* and identically distributed. Nevertheless, the techniques permit general forms of dependence for $P_{X_1,\ldots,X_n|\theta}$. The cumulative Kullback–Leibler divergence $D(p_{X_1,\ldots,X_n|\theta}\|q_{X_1,\ldots,X_n})$ is made to be of order $(k/2)\log n + O(1)$

by suitable mixtures $Q_{X_1,\ldots,X_n}$, as long as $n^{-1}D(p_{X_1,\ldots,X_n|\theta}\|p_{X_1,\ldots,X_n|\eta}) \le c_\theta\|\theta-\eta\|^2$ for η in a neighborhood $\mathcal{N}_\theta$ of θ (cf. Lemma 12 in the Appendix). If Assumption A holds, then for any $Q_{X_1,\ldots,X_n}$, the set of θ for which $\limsup_n D(p_{X_1,\ldots,X_n|\theta}\|q_{X_1,\ldots,X_n})/\log n < k/2$ has Lebesgue measure 0.

When $X_1,\ldots,X_n$ are allowed to be dependent, the chain rule becomes

$$D\big(p_{X_1,\ldots,X_n|\theta}\|q_{X_1,\ldots,X_n}\big) = \sum_{l=0}^{n-1} \mathbb{E}_{P_\theta^l} D\big(P_{X_{l+1}|X^l,\theta}\|Q_{X_{l+1}|X^l}\big).$$

Correspondingly, the set of θ for which

$$\limsup_{n\to\infty} n\mathbb{E}_{P_\theta^n} D\big(p_{X_{n+1}|X^n,\theta}\|q_{X_{n+1}|X^n}\big) < k/2$$

has measure 0. That is, the predictive distributions $P_{X_{n+1}|X^n,\theta}$ are estimated with Kullback–Leibler loss not less than $k/2n$ except for a negligible set of parameters.

3. Superefficiency in a metric space context. In this section we study the asymptotics for the Kullback–Leibler divergence between members of the family P_θ^n and any distribution Q^n on $\mathbf{X}^n$. For the lower bound, we examine the negligibility of the set of θ with small total Kullback–Leibler loss $D(p_\theta^n\|q^n)$ in the sense of having a small cardinality cover as we shall make precise. For the upper bound, we explicitly construct distributions Q^n that, under additional assumptions, have total Kullback–Leibler loss of the same order as the lower bound. Both these theorems are stated for general metric spaces.

Let K be a compact set in a metric space $(\mathcal{M}, d)$. For subsets $A \subset K \subset \mathcal{M}$, the packing number $M(A,\varepsilon)$ is the largest number of points in A that are at least distance ε apart and the covering number of $N(A,\varepsilon)$ is the smallest number of balls of radius ε with centers in A needed to cover the set A. The related metric entropy is

$$H(A,\varepsilon) = H(\varepsilon) = \log N(A,\varepsilon).$$

Both the packing and covering numbers (and associated metric entropy) provide an intrinsic measure of dimension as ε tends to 0. For example, the covering number of open balls B on $\mathbb{R}^k$ is of order $N(B,\epsilon) \asymp (1/\varepsilon)^k$, where we use $a_n \asymp b_n$ to indicate that $a_n/b_n \to 1$. By analogy, sets with covering numbers of this order are said to have metric dimension k. If the covering number grows faster than $(1/\varepsilon)^k$ for every k, the set A is said to be infinite dimensional. Implicitly, we will assume that the covering number $N(\Theta,\varepsilon)$ tends to ∞ as ε goes to 0.

For $A \subset B \subset K$, the relative size of A to the size of B can be measured by the ratio of the number of small balls needed to cover A and B, respectively.

DEFINITION 3. For sets $A \subset B \subset K$, the relative measure of A to B is

$$v_B^*(A) = \limsup_{\varepsilon \to 0^+} \frac{N(A, \varepsilon)}{N(B, \varepsilon)}.$$

Moreover, the set A is said to be negligible in B if $v_B^*(A) = 0$.

In the Appendix we show for parameter sets Θ that have finite metric dimension, that if there exists a locally invariant measure on Θ assigning the same mass to small radius balls, then negligibility is equivalent to having measure 0. When d is the Euclidean distance, Lebesgue is the invariant measure; when d is the Hellinger metric, Jeffreys prior provides the local invariant measure [cf. Jeffreys (1946) and Hartigan (1983), page 49]. To handle more general contexts of interest, we reformulate the assumption and conclusion in terms of negligibility.

ASSUMPTION C. Let $\Theta \subset \mathscr{M}$, a metric space with metric $d(\cdot, \cdot)$. For a sequence $\{r_n\}$ which tends to 0 as $n \to \infty$ and for each compact subset $K \subset \Theta$, there exists a sequence of estimator $\tilde{\theta}_n$ based on $X_1, \ldots, X_n$ (possibly depending on K) and finite integer n_o, such that the sets

$$\mathscr{B}_{l,\eta} = \left\{ \theta \in K \colon \sup_{n \geq n_o} P_\theta^n \left[d\left(\tilde{\theta}_n, \theta\right) > l r_n \right] > \eta \right\}$$

have covering numbers satisfying

$$\lim_{l \to \infty} \limsup_{\varepsilon \to 0^+} \frac{N\left(\mathscr{B}_{l,\eta}, \varepsilon\right)}{N(K, \varepsilon)} = 0$$

for every $\eta > 0$.

In the setting of Assumption A, tightness of the sequence of $d(\tilde{\theta}_n, \theta)/r_n$ for almost every θ implies Assumption C.

We first exhibit the lower bound for the total Kullback–Leibler divergence.

THEOREM 2. *Suppose the sequence of families of distributions* $\{P_\theta^n \colon \theta \in \Theta\}$ *on* $\mathbf{X}^n$ *has a parameter set* Θ *contained in a compact set* K *in the metric space* $(\mathscr{M}, d)$. *Assume that the* $H(\varepsilon) = \log N(\Theta, \varepsilon)$ *increases to* ∞ *while* ε *decreases to* 0. *If Assumption* C *holds, then for every sequence of probability distributions* $\{Q^n\}_{n=1}^\infty$ *on* $\mathbf{X}^n$ *and divergent sequence* $\{\alpha_n\}$ *for which* $\alpha_n r_n = o(1)$, *the set*

$$\left\{ \theta \in \Theta \colon \limsup_{n \to \infty} \frac{D(p_\theta^n \| q^n)}{H(\alpha_n r_n)} < 1 \right\}$$

is negligible in Θ *with respect to the metric* d.

REMARK. The conclusion of the theorem is easily extended to σ-compact parameter spaces. Given a countable collection of compact sets K_m whose

union contains Θ, we apply Theorem 2 to each of the sets $\Theta \cap K_m$. Take H to be the logarithm of the minimum of the covering numbers required to cover each individual intersection $\Theta \cap K_m$, $m = 1, 2, \ldots$, where $\Theta \subset \bigcup_m K_m$ produces a lower bound for the total Kullback–Leibler divergence.

PROOF OF THEOREM 2. As in Proposition 1, we consider the inequality

$$D(p_\theta^n \| q^n) \geq P_\theta^n(E) \log\left(\frac{1}{Q^n(E)}\right) - \log 2,$$

with sets of the form

$$\mathscr{E}_{\theta,n} \equiv \left\{x^n \in \mathbf{X}^n : d\big(\tilde{\theta}_n(x^n), \theta\big) \leq \frac{\delta_n}{2}\right\},$$

where $\delta_n = 2\alpha_n r_n$. We bound $Q^n(\mathscr{E}_{\theta,n})$ by showing that the set

$$\mathscr{A}_n \equiv \left\{\theta : Q^n(\mathscr{E}_{\theta,n}) \geq \frac{\log N(\Theta, \delta_n)}{N(\Theta, \delta_n)}\right\}$$

is asymptotically negligible. Let M_n be the maximal number of disjoint balls of radius $\delta_n/2$ having center in $\mathscr{A}_n$ and denote by $\mathscr{C}_n$ the collection of their centers. The triangle inequality implies $N(\mathscr{A}_n, \delta_n) < M_n$. Since the sets $\{\mathscr{E}_{\theta,n} : \theta \in \mathscr{C}_n\}$ are disjoint,

$$1 \geq \sum_{\theta \in \mathscr{C}_n} Q^n(\mathscr{E}_{\theta,n}) \geq M_n \frac{\log N(\Theta, \delta_n)}{N(\Theta, \delta_n)},$$

which implies that

$$M_n \leq \frac{N(\Theta, \delta_n)}{\log N(\Theta, \delta_n)}.$$

The ratio of the covering numbers of $\mathscr{A}_n$ and Θ is thus bounded by

$$\frac{N(\mathscr{A}_n, \delta_n)}{N(\Theta, \delta_n)} \leq \frac{M_n}{N(\Theta, \delta_n)} \leq \frac{1}{\log N(\Theta, \delta_n)}, \tag{11}$$

which goes to 0 as n tends to ∞ since $\delta_n = o(1)$. Whence the sets

$$\left\{\theta \in \Theta : Q^n(\mathscr{E}_{\theta,n}) \geq \frac{\log N(\Theta, \delta_n)}{N(\Theta, \delta_n)} \text{ for all } n > n_0\right\}$$

are negligible for all n_0.

We now turn to bounding $P_\theta^n[\mathscr{E}_{\theta,n}]$. For this, define for $\eta \in (0,1)$ the sets

$$\mathscr{B}_\eta = \left\{\theta \in \Theta : \inf_n P_\theta^n[\mathscr{E}_{\theta,n}] < 1 - \eta\right\}$$

and

$$\mathscr{B}_{l,\eta} = \left\{\theta \in \Theta : \inf_n P_\theta^n\big[d\big(\tilde{\theta}_n, \theta\big) \leq l r_n\big] < 1 - \eta\right\}.$$

Since $l_1 < l_2$ implies that $\mathscr{B}_{l_1,\eta} \supset \mathscr{B}_{l_2,\eta}$, it follows that $\mathscr{B}_\eta \subset \mathscr{B}_{l,\eta}$ for all l. From Assumption C, we conclude that

$$\lim_{\varepsilon \to 0+} \frac{N(\mathscr{B}_\eta, \varepsilon)}{N(\Theta, \varepsilon)} \le \lim_{l \to \infty} \lim_{\varepsilon \to 0+} \frac{N(\mathscr{B}_{l,\eta}, \varepsilon)}{N(\Theta, \varepsilon)} = 0.$$

For $\theta \in (\mathscr{A}_n \cup \mathscr{B}_\eta)^c$, the relative entropy is lower bounded by

$$\begin{aligned} D(p_\theta^n \| q^n) &\ge P_\theta^n[\mathscr{E}_{\theta,n}] \log\left(\frac{1}{Q^n[\mathscr{E}_{\theta,n}]}\right) - \log 2 \\ &\ge (1-\eta)\left(\log \frac{N(\Theta, \delta_n)}{\log N(\Theta, \delta_n)}\right) - \log 2. \end{aligned}$$

Identifying $H(\delta_n) = \log N(\Theta, \delta_n)$, we conclude that the set

$$\{\theta \in \Theta\colon D(P_\theta^n \| q^n) \ge (1-2\eta) H(\delta_n) \text{ for all } n > n_0\} \tag{12}$$

is negligible for all $\eta \in (0,1)$ and n_0.

We conclude that the set $\{\theta \in \Theta\colon D(p_\theta^n \| q^n) < H(\delta_n)\}$ is negligible in Θ by noting that it is the limit in j of the increasing sequence of sets $\{\theta \in \Theta\colon D(p_\theta^n \| q^n) < (1-2/j)H(\delta_n)\}$, each of which is negligible in Θ. □

In general, Theorem 2 identifies $H(r_n)$ as a lower bound for the convergence rate of $D(p_\theta^n \| q^n)$. The conclusion can be more precise in the case that Θ has finite metric dimension. Indeed we show that the metric dimension k is a lower bound for $D(p_\theta^n \| q^n)/\log r_n$, except for a negligible set of θ.

THEOREM 3. *Let the parameter space $\Theta \subseteq \mathscr{M}$ have finite metric dimension k in a σ-compact metric space $(\mathscr{M}, d)$. If Assumption* C *holds, then for any sequence of distributions Q^n on $\mathbf{X}^n$, the set*

$$\left\{\theta \in \Theta\colon \limsup_{n \to \infty} \frac{D(p_\theta^n \| q^n)}{\log r_n} < k\right\}$$

is negligible in Θ.

REMARK. The conclusion of Theorem 3 remains unchanged when the rate r_n is increased to $r_n(\log(1/r_n))^\beta$ for any $\beta > 0$ (as can be seen by inspecting the proof below). Whence Assumption C can be weakened to assuming negligibility of the set

$$\mathscr{B}_{l,\eta} = \left\{\theta \in K\colon \sup_n P_\theta^n\left[d(\tilde{\theta}_n, \theta) > l \frac{r_n}{\log r_n}\right] > \eta\right\}$$

as l goes to ∞.

PROOF OF THEOREM 3. Without loss of generality, we assume that Θ is bounded. If not, write $\mathscr{M} = \bigcup_j K_j$, where K_j are compact sets with nonempty interior, and apply the argument to each of the bounded sets $\Theta \cap K_j$. Since the set Θ has finite metric entropy, the covering numbers $N(\Theta; \varepsilon)$ are of

order $(1/\varepsilon)^k$. Now apply Theorem 2 with $\alpha_n = \log(1/r_n)$ to conclude that the set of parameters for which

$$\limsup_{n\to\infty} \frac{\log 1/r_n}{\log 1/r_n - \log\log 1/r_n} \times \frac{D(p_\theta^n \| q^n)}{\log 1/r_n} < k$$

is negligible. The conclusion follows by noting that

$$\frac{\log 1/r_n}{\log 1/r_n - \log\log 1/r_n} \to 1 \quad \text{with } n \to \infty. \qquad \square$$

We now provide conditions for the existence of sequences of distributions Q^n on $\mathbf{X}^n$ having Kullback–Leibler divergence $D(p_\theta^n \| q^n)$ matching the lower bounds of Theorems 2 and 3 up to asymptotically negligible terms, and this for all $\theta \in \Theta$. In particular, in the finite metric dimension case, it reveals the efficient constant in terms of the metric dimension k.

To construct such distributions Q^n on $\mathbf{X}^n$, let W_n be a sequence of discrete prior distributions on Θ. Consider the sequence of distributions Q^n with probability densities

$$q^n(x^n) = \int_\Theta p^n(x^n|\eta) W^n(d\eta)$$

on $\mathbf{X}^n$. This choice is motivated by the fact that, for each n, Q^n minimizes the average total Kullback–Leibler loss $\int D(p_\theta^n \| q^n) W_n(d\theta)$. When W_n are judiciously chosen, the mixture $q^n(x^n)$ is a good candidate to achieve $D(p_\theta^n \| q^n)$ of order $H(r_n)$ (with best constants in the finite metric dimension case).

The following theorem gives an upper bound in terms of the covering numbers of the parameter set by ε-balls of the Kullback–Leibler divergence $n^{-1}D(p_\theta^n \| p_\eta^n)$. In the i.i.d. case this divergence reduces to $D(p_\theta \| p_\eta)$. A subset K of Θ has an ε information cover of cardinality $\mathsf{N}_n(\varepsilon)$ if there exists a set $\mathscr{C}_n = \{\eta_1, \eta_2, \ldots, \eta_{\mathsf{N}_n(\varepsilon)}\}$ of points in Θ such that for every $\theta \in \mathsf{K}$ there exists $\eta_l \in \mathscr{C}$ for which $n^{-1}D(p_\theta^n \| p_{\eta_l}^n) \le \varepsilon^2$.

THEOREM 4. *Suppose that parameter set Θ is represented as $\bigcup_j \mathsf{K}_j$, where the subsets K_j are independent of n and have ε information covering number bounded by $\mathsf{N}_n(\varepsilon)$ uniformly in j. Then for every sequence of positive numbers $\tilde{r}_n$, there exists a sequence of distributions Q^n on $\mathbf{X}^n$ such that*

$$D(p_\theta^n \| q^n) \le \log \mathsf{N}_n(\tilde{r}_n) + n\tilde{r}_n^2 + \mathsf{C}(\theta) \tag{13}$$

for some constant $\mathsf{C}(\theta)$ not depending on n or the sequence $\{\tilde{r}_n\}$.

REMARK 1. The best such bounds are obtained by choosing $\tilde{r}_n$ to (approximately) minimize $\log \mathsf{N}_n(\tilde{r}_n) + n\tilde{r}_n^2$. Combining this bound with its implications for the existence of estimators satisfying Assumption C at rate $r_n = \tilde{r}_n$, as developed in Section IV, will reveal the optimality of our upper and lower bounds.

REMARK 2. The proof applies techniques from Clarke and Barron (1990), and uses Lemma 12 in the Appendix that is implicit in Barron (1987).

PROOF OF THEOREM 4. Given $\tilde{r}_n$, let $\mathsf{N}_n = \mathsf{N}_n(\tilde{r}_n)$ and $\mathscr{C}_j = \{\eta_{j,1}, \eta_{j,2}, \ldots, \eta_{j,\mathsf{N}_n}\}$ be the set of centers in K_j such that $\mathsf{K}_j \subset \bigcup_{l=1}^{\mathsf{N}_n}\{\theta\colon n^{-1}D(p_\theta^n \| p_{\eta_{j,l}}^n) \le \tilde{r}_n^2\}$. Define a prior W_n via the following two-stage process: First draw $\mathscr{J}$, an integer-valued random variable with probability mass function $\pi(j) > 0$. Given $\mathscr{J} = j$, draw a point uniformly from the finite collection $\mathscr{C}_j$. We now apply Lemma 12 in the Appendix with $B_n = \{\eta \in \Theta\colon D(p_\theta^n \| p_\eta^n) \le n\tilde{r}_n^2\}$ to conclude that

$$\begin{aligned} D(p_\theta^n \| q^n) &\le \log \frac{1}{W_n(B_n)} + \int_{B_n} D(p_\theta^n \| p_\eta^n) \frac{W_n(d\eta)}{W_n(B_n)} \\ &\le \log \frac{1}{W_n(B_n)} + n\tilde{r}_n^2. \end{aligned}$$

By construction of the prior W_n, if $\theta \in \mathsf{K}_j$ then

$$\log \frac{1}{W_n(B_n)} \le \log \mathsf{N}_n(\tilde{r}_n) - \log \pi(j).$$

Thus the total Kullback–Leibler loss is bounded by

$$D(p_\theta^n \| q^n) \le \log \mathsf{N}_n(\tilde{r}_n) + n\tilde{r}_n^2 + \mathsf{C}(\theta),$$

where $\mathsf{C}(\theta) = \min\{-\log \pi(j)\colon \mathsf{K}_j \ni \theta\}$. □

The exhibited sequence of distributions Q^n in Theorem 4 are not Kolmogorov consistent since the prior W_n changes with n. One might expect the upper bound to be much larger when one requires the sequence Q^n to be Kolmogorov consistent. This is not the case, as we now proceed to show.

Consider the mixing distribution on Θ:

$$\tilde{W} = \sum_{m=1}^{\infty} \frac{1}{m(m+1)} W_{2^{m-1}},$$

where W_n are the mixing distributions used in Theorem 4 and $\sum_{m=1}^{\infty} 1/(m(m+1)) = 1$. The mixture using $\tilde{W}$ achieves asymptotically the same upper bound as the one obtained from the sequence W_n. Indeed, for B_n as in Theorem 4, we have that $\tilde{W}(B_n) \ge (1/m_n(m_n+1))W_{2^{m_n}}(B_n)$, where $m_n = \lfloor(\log n)/2\rfloor$, and the bound on the total Kullback–Leibler loss

$$D(p_\theta^n \| q^n) \le \log \mathsf{N}(\tilde{r}_n) + n\tilde{r}_n^2 + 3\log\log n + \mathsf{C}(\theta)$$

follows. In all the cases considered in this paper, the additional $\log\log n$ term is of smaller order than $\log \mathsf{N}(\tilde{r}_n) + n\tilde{r}_n^2$ for any choice of $\tilde{r}_n$.

The upper bound involves the covering number of the parameter set in Kullback–Leibler divergence whereas the lower bound from Theorem 2 uses covering numbers of the parameter sets using the metric d. These quantities are related in the i.i.d. case if the Kullback–Leibler divergence is locally

bounded by

$$D(p_\theta \| p_\eta) \le C_\theta \varphi(d^2(\theta, \eta)) \quad \text{for all } \eta \text{ in a neighborhood } \mathcal{N}_\theta \text{ of } \theta,$$

where φ is a continuous, monotone increasing function with $\varphi(0) = 0$. It then follows that, for small $\tilde{r}_n$, $\log \mathsf{N}(\tilde{r}_n) \le H(\varphi^{-1}(\tilde{r}_n^2/C_\theta))$. The following theorem considers an important special case of the latter.

THEOREM 5. *Suppose that the parameter set Θ has finite metric dimension k in $(\mathcal{M}, d)$ and that for each $\theta \in \Theta$ the Kullback–Leibler divergence is locally upper bounded by $D(p_\theta \| p_\eta) \le C_\theta d(\theta, \eta)^\alpha$ for all θ in a neighborhood $\mathcal{N}_\theta$. If $X_1, \ldots, X_n$ are i.i.d. P_θ, then there exists a sequence of distributions Q^n such that the total Kullback–Leibler loss is bounded by*

$$D(p_\theta^n \| q^n) \le \frac{k}{\alpha} \log n + \mathsf{C}(\theta) \tag{14}$$

for some constant $\mathsf{C}(\theta)$ not depending on n.

If the Kullback–Leibler divergence and the squared Hellinger distance are further locally lower bounded by $D(p_\theta \| p_\eta) \ge h^2(p_\theta, p_\eta) \ge c_\theta d(\theta, \eta)^\alpha$ for all $\eta \in \mathcal{N}_\theta$ and $c_\theta > 0$, then, for all sequences of distributions Q^n,

$$D(p_\theta^n \| q^n) \ge \frac{k}{\alpha} \log n$$

for all but a negligible set of θ.

We momentarily delay the proof of Theorem 5 to first exhibit sequences of estimators whose Kullback–Leibler risk converges at an appropriate rate. In Theorem 5, we use these estimators to show that Assumption C holds with a suitable rate to produce the desired conclusions. We note that efficient parameter estimation will be considered in the next section.

LEMMA 6. *Assume that the family of distributions $\{P_\theta\colon \theta \in \Theta\}$ is dominated by ν and that P_θ^n makes $X_1, \ldots, X_n$ i.i.d. with density p_θ. Then to any sequence of Kolmogorov-consistent distributions Q^n, there is a related sequence of density estimators $\tilde{p}_n$ that satisfy*

$$\mathbb{E}_\theta D(p_\theta \| \tilde{p}_n) \le \frac{1}{n} D(p_\theta^n \| q^n).$$

PROOF. Let $\hat{p}_0(x) = q^1(x)$, and for $n \ge 1$ set

$$\hat{p}_n(x) = \frac{q^{n+1}(x^n, x)}{q^n(x^n)} \equiv q_{n+1}(x | x^n)$$

and define the Césaro average density estimator

$$\tilde{p}_n(x) = \frac{1}{n} \sum_{l=0}^{n-1} \hat{p}_l(x).$$

An application of Jensen's inequality produces

$$\begin{aligned} D(p_\theta \| \tilde{p}_n) &= \int \log\left(\frac{p_\theta(x)}{n^{-1}\sum_{l=1}^{n} \hat{p}_l(x)} \right) p_\theta(x)\, \nu(dx) \\ &\le \frac{1}{n} \sum_{l=1}^{n} \int \log\left(\frac{p_\theta(x)}{\hat{p}_l(x)} \right) p_\theta(x)\, \nu(dx) \\ &= \frac{1}{n} \sum_{l=1}^{n} D\big(p_\theta \| \hat{p}_l(x|x^l)\big). \end{aligned}$$

Taking the expectation and applying the chain of information theory produces

$$\begin{aligned} \mathbb{E}_\theta\big[D(p_\theta \| \tilde{p}_n)\big] &\le \frac{1}{n} \sum_{l=0}^{n-1} \mathbb{E}_\theta\left[\log\left(\frac{p_\theta(X_{l+1})}{q\big(X_{l+1}|X^l\big)} \right) \right] \\ &= \frac{1}{n} D(p_\theta^n \| q^n). \end{aligned}$$

□

PROOF OF THEOREM 5. By Theorem 4, the total Kullback–Leibler loss is bounded by

$$D(p_\theta^n \| q^n) \le \log \mathsf{N}(\tilde{r}_n) + n\tilde{r}_n^2 + \mathsf{C}(\theta).$$

The bound on the Kullback–Leibler loss implies that

$$\big\{\eta\colon D(\theta\|\eta) \le \tilde{r}_n^2\big\} \supset \left\{ \eta\colon d(\theta,\eta)^\alpha \le \frac{\tilde{r}_n^2}{C_\theta} \right\}$$

and therefore $\log \mathsf{N}(\tilde{r}_n) \le \log(aC_\theta/\tilde{r}_n^2)^{k/\alpha}$, where $(a/\varepsilon)^k$ is a bound on the metric entropy of Θ using metric d. Whence

$$\begin{aligned} D(p_\theta^n \| q^n) &\le \log\left(\frac{aC_\theta}{\tilde{r}_n^2} \right)^{k/\alpha} + n\tilde{r}_n^2 + \mathsf{C}(\theta) \\ &\le \frac{2k}{\alpha} \log \frac{1}{\tilde{r}_n} + n\tilde{r}_n^2 + C_1(\theta) \end{aligned}$$

for some constant $C_1(\theta)$. The latter is minimized by taking $\tilde{r}_n = \sqrt{kn/\alpha}$ which leads to the claimed upper bound

$$D(P_\theta^n \| Q^n) \le \frac{k}{\alpha} \log n + C_2(\theta). \tag{15}$$

To prove the lower bound, we will show that Assumption C holds. For this, let K be a compact set. From the proof of the above upper bound, there exists a sequence of Kolmogorov-consistent distributions Q^n for which

$$D(p_\theta^n \| q^n) \le \frac{k}{\alpha} \log n + 3 \log\log n + C(\theta) \tag{16}$$

for all distributions p_θ with $\theta \in K$. As in Lemma 6, $\hat{p}_0(x) = q^1(x)$, define $\hat{p}_n(x) = q^{n+1}(x^n, x)/q^n(x^n)$ with $n \geq 1$ and

$$\tilde{p}_n = \frac{1}{n}\sum_{l=0}^{n-1} \hat{p}_l(x).$$

By Lemma 6 and (16), there exist density estimators $\tilde{p}_n$ satisfying, for every $\theta \in K$,

$$\mathbb{E}_\theta\big[D(p_\theta \| \tilde{p}_n)\big] \leq \frac{k}{\alpha}\frac{\log n}{n}(1 + o(1)).$$

Let $\tilde{\theta}_n$ be any value of $\theta \in K$ minimizing $h^2(\tilde{p}_n, p_{\tilde{\theta}_n})$ (or within a factor of 2 of the infimum if the minimum is not exactly attained). The assumed relation between d and h, the triangle inequality for the Hellinger distance and the bound $h^2 \leq D$ imply that

$$\begin{aligned}\mathbb{E}_\theta c_\theta d\big(\theta, \tilde{\theta}_n\big)^\alpha &\leq \mathbb{E}_\theta h^2(p_\theta, p_{\tilde{\theta}_n}) \leq 4\mathbb{E}_\theta h^2(p_\theta, \tilde{p}_n)\\ &\leq 4\mathbb{E}_\theta\big[D(p_\theta \| \tilde{p}_n)\big] \leq \frac{4k\log n}{\alpha n}(1 + o(1)).\end{aligned}$$

An application of Markov's inequality shows that Assumption C holds with $r_n = [(\log n)/n]^{1/\alpha}$. The second conclusion therefore follows from Theorem 3. □

4. Superefficient parametric estimation. In this section we use the results of the previous section to establish the negligibility of the set of superefficiency for estimation in finite-dimensional parametric families. First, we state the conclusions for estimation in general metric spaces. Then we specialize the conclusion to express negligibility in terms of the Hellinger metric.

THEOREM 7. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from $P_\theta \in \{P_\theta\colon \theta \in \Theta\}$, a parametric family of distributions with parameter space Θ of finite metric dimension k, satisfying Assumptions* B *and* C. *Let $\gamma(n)$ be any nonnegative sequence satisfying*

$$\frac{1}{\gamma(n)} \asymp \log\frac{r_n}{r_{n+1}} \tag{17}$$

in the sense that the ratio of the two sides converges to 1. *Then for any sequence of estimators $\hat{p}_n(x) = \hat{p}_n(x; X^n)$ of the density $p_\theta(x)$, the set of θ for which*

$$\limsup_{n\to\infty} \gamma(n)\mathbb{E}_\theta D(p_\theta \| \hat{p}_n) < k$$

is negligible in Θ. In particular, for sequences of estimators $\hat{\theta}_n$ based on X^n, the set of parameter values of θ for which $\limsup_{n\to\infty}\gamma(n)\mathbb{E}_\theta D(\theta\|\hat{\theta}_n) < k$ is negligible in Θ.

REMARK 1. For $r_n = n^{-1/2}$, one can set $\gamma(n) = 2n$ in accordance with the conclusions derived in the Introduction.

REMARK 2. The individual Kullback–Leibler risks are connected to a total Kullback–Leibler risk via the chain rule. By use of this chain rule, we give a lower bound on the individual risk for arbitrary estimators (for almost all θ).

PROOF OF THEOREM 7. To show that the set

$$\mathscr{A} \equiv \left\{\theta \in \Theta\colon \limsup_{n\to\infty} \gamma(n)\mathbb{E}_\theta D(p_\theta \| \hat{p}_n) < k\right\}$$

is negligible, consider the sets

$$\mathscr{A}_{s,m} = \left\{\theta \in \Theta\colon \gamma(n)\mathbb{E}_\theta D(p_\theta \| \hat{p}_n) < k\left(1 - \frac{1}{s}\right) \text{ for all } n > m\right\}.$$

From the definition of lim sup, it follows that $\mathscr{A} \subset \bigcup_{s=1}^{\infty} \bigcup_{m=1}^{\infty} \mathscr{A}_{s,m}$ and the union bound implies $\nu_\Theta^*(\mathscr{A}) \le \sum_{s=1}^{\infty}\sum_{m=1}^{\infty} \nu_\Theta^*(\mathscr{A}_{s,m})$. By the monotonicity $\mathscr{A}_{s,m} \subset \mathscr{A}_{s,m+1}$, it suffices to show that $\nu_\Theta^*(\mathscr{A}_{s,lm_0})$ is 0 for all large l. For this, we will apply the chain rule of information theory.

Given $s < 1$, pick $m = lm_0$ large enough such that

$$\frac{1}{\gamma(j)} \le \left(1 + \frac{1}{s}\right)\log\frac{r_j}{r_{j+1}} \quad \text{for all } j \ge m.$$

Define the probability distribution Q^n on $\mathbf{X}^n$ via its probability density

$$q(x^n) = q_0^m(x^m)\,\hat{p}_m(x_{m+1})\,\hat{p}_{m+1}(x_{m+2}) \cdots \hat{p}_{n-1}(x_n),$$

where $q_0^m(x^m)$ is the l-fold product of the densities of the measure Q^{m_0} from Assumption B. For $\theta \in \mathscr{A}_{s,m}$ and $n > m$,

$$\begin{aligned} D(p_\theta^n \| q^n) &= D(p_\theta^m \| q^m) + \sum_{j=m}^{n-1} \mathbb{E}_\theta D(p_\theta \| \hat{p}_j) \\ &\le lD(p_\theta^{m_0} \| q^{m_0}) + k\left(1 - \frac{1}{s}\right)\sum_{j=m}^{n-1} \frac{1}{\gamma(j)} \\ &\le lD(p_\theta^{m_0} \| q^{m_0}) + k\left(1 - \frac{1}{s}\right) \\ &\quad \times \sum_{j=m}^{n-1} \left(1 + \frac{1}{s}\right)\left(\log\frac{1}{r_{j+1}} - \log\frac{1}{r_j}\right) \\ &\le lD(p_\theta^{m_0} \| q^{m_0}) + k\left(1 - \frac{1}{s^2}\right)\left(\log\frac{1}{r_n} - \log\frac{1}{r_m}\right). \end{aligned} \tag{18}$$

For n large enough, the latter implies that

$$\frac{D(p_\theta^n \| q^n)}{\log 1/r_n} < \left(1 - \frac{1}{2s^2}\right)k \tag{19}$$

and the conclusion follows from Theorem 3. □

We now specialize the previous results to the problem of parametric density estimation. To express negligibility, we parametrize the ν-dominated family of distributions $\{P_\theta\colon \theta \in \Theta\}$ by their probability densities $\{p = dP_\theta/d\nu\}$ and view the latter as a subset of $\mathscr{H}$, the space of all densities dominated by ν, endowed with the Hellinger metric $h(p, q) = \{\int(\sqrt{p(x)} - \sqrt{q(x)})^2 \, d\nu(x)\}^{1/2}$. We assume that $X_1, X_2, \ldots, X_n$ are i.i.d. with density in the family.

THEOREM 8. *Suppose that the parameter set of the family of distributions* $\{P_\theta\colon \theta \in \Theta\}$ *has finite metric dimension* k *for some metric* d, *and suppose that there exists an* $\alpha > 0$ *such that for each* $\theta \in \Theta$ *there exist a constant* $C_\theta > 0$ *and an open neighborhood* $\mathscr{N}_\theta$ *such that the relative entropy is locally bounded by*

$$D(p_\theta \| p_\eta) \le C_\theta d(\theta, \eta)^\alpha \quad \text{for all } \eta \in \mathscr{N}_\theta. \tag{20}$$

Then the set of densities for which

$$\left\{p \in \mathscr{P}\colon \limsup_{n\to\infty} n\mathbb{E}_p D(p \| \hat{p}_n) < \frac{k}{\alpha}\right\}$$

is asymptotically negligible in $\mathscr{P}$ *with respect to the Hellinger metric.*

REMARK 1. The proof reveals that the upper bound on the Kullback–Leibler divergence implies Assumption C in the Hellinger metric with $r_n = \sqrt{(\log n)/n}$, and this regardless of the value of α.

REMARK 2. Negligibility in Ω with respect to the Hellinger distance does not imply negligibility in Θ with respect to d unless there is a suitable relationship between h and d.

PROOF OF THEOREM 8. The Kullback–Leibler loss is bounded from below by the square of the Hellinger distance. Thus the upper bound on the Kullback–Leibler loss implies that locally

$$h^2(p_\theta, p_\eta) \le D(p_\theta \| p_\eta) \le C_\theta d(\theta, \eta)^\alpha.$$

Whence it follows that the covering numbers of every compact subset of $\mathscr{P}$ by Hellinger balls of radius ε are bounded by $C\varepsilon^{-2k/\alpha}$, which implies that the Hellinger metric dimension of $\mathscr{P}$ is bounded by $2k/\alpha$, where k is the d-metric dimension of Θ. The lower bound on the Kullback–Leibler loss also implies that Assumption C holds in the Hellinger metric with $r_n = \sqrt{(\log n)/n}$. The argument is the same as the one presented in Theo-

rem 5. Assumption B follows from the assumed local upper bound on the Kullback–Leibler loss, and the conclusion follows from Theorem 7 with $\gamma(n) = 2n$. Indeed, this choice satisfies condition (17) since

$$\frac{1}{\gamma(n)} = \log \frac{\sqrt{(\log n)/n}}{\sqrt{(\log(n+1))/(n+1)}} = \frac{1}{2n} + O\left(\frac{1}{n \log n}\right). \qquad \square$$

5. An implication for nonparametric rates. We now present the implication of Theorem 2 for lower bounds on nonparametric rates of convergence in the Hellinger metric. As in the previous section, we view the ν-dominated family of distributions $\mathscr{P} = \{P_\theta\colon\ \theta \in \Theta\}$ as being parametrized by the densities $\Omega = \{p = dP/d\nu\colon\ P \in \mathscr{P}\}$, and take the latter as a subset of the space $\mathscr{H}$ of all densities dominated by ν endowed with the Hellinger metric. Our result identifies nonparametric rates of convergence, such that the set of densities that can be estimated at a faster rate is Hellinger negligible. Related results on minimax convergence rates are also found in Yang and Barron (1995) and Birgé (1986).

THEOREM 9. *Let $\mathscr{P}$ be a dominated collection of distributions and assume that $\mathsf{H}(\varepsilon)$, the Hellinger metric entropy of the associated set of densities, is such that, for every $c > 0$, $\limsup_{\varepsilon \to \infty} \mathsf{H}(c\varepsilon)/\mathsf{H}(\varepsilon)$ is finite. Further assume that for each $P \in \mathscr{P}$ there exist a Hellinger ball $\mathscr{N}_P$ and a positive constant C_P such that $D(p\|q) \le \mathsf{C}_P h^2(p,q)$ for all $Q \in \mathscr{N}_P \cap \mathscr{P}$. Choose r_n such that*

$$\mathsf{H}(r_n) \le n r_n^2$$

and $\limsup_{n\to\infty} r_{n+1}/r_n = 1$. Then for any estimator sequence $\{\hat{p}_n\}$, the set of densities which can be estimated in Kullback distance at rate faster than r_n^2 is Hellinger negligible. That is, for any sequence c_n decreasing to 0 such that $\limsup_{n\to\infty} c_n/r_n = 0$, the set

$$\left\{p \in \mathscr{P}\colon\ \limsup_{n\to\infty} \mathbb{E}_p D(p\|\hat{p}_n)/c_n^2 < 1\right\}$$

is Hellinger negligible.

REMARK. From the assumed equivalence between h^2 and D, it follows that the set

$$\left\{p \in \mathscr{P}\colon\ \limsup_{n} h(p, \hat{p}_n) < c_n\right\}$$

is also Hellinger negligible. In infinite-dimensional spaces, we no longer can identify best constants, but we can still reveal the best rate.

PROOF OF THEOREM 9. Combining the upper bound on the Kullback–Leibler loss and Theorem 4 shows that Assumption B holds with $m = 1$. We now parallel the argument of Theorem 7 to show that Assumption C holds with rate r_n. Indeed, the sequence of density estimators $\tilde{p}_n$ considered in Lemma 6, satisfies, for all $p \in \mathscr{P}$,

$$\begin{aligned}\mathbb{E}\left[h^2(p, \tilde{p}_n)\right] &\leq \frac{D(p^n \| q^n)}{n} \\ &\leq \frac{\log \mathsf{N}(r_n) + nr_n^2}{n} \\ &\leq \frac{\mathsf{H}(r_n) + nr_n^2}{n} \\ &\leq 2r_n^2.\end{aligned}$$

Assumption C follows from Markov's inequality.

As in Theorem 7, the conclusion will follow if the set

$$\left\{p \in \mathscr{P}: \frac{D(p^n \| q^n)}{\sum_{j=1}^n c_n^2} < 1\right\} \tag{21}$$

is Hellinger negligible. By the definition of r_n,

$$\begin{aligned}H(r_{n+1}) - H(r_n) &= (n+1)r_{n+1}^2 - nr_n^2 \\ &= nr_n^2\left((1+n^{-1})\frac{r_{n+1}^2}{r_n^2} - 1\right) \\ &= r_n^2(1 + o(1))\end{aligned}$$

and whence the sum $\sum_{j=1}^n r_n^2$ telescopes into $H(r_{n+1})(1 + o(1))$. By assumption, $\sum_{j=1}^n c_j^2 = o(\sum_{j=1}^n r_j^2)$, so that, by the monotonicity of $\mathsf{H}(\varepsilon)$, the definition of r_n and the assumption that $\limsup_{\varepsilon \to \infty} \mathsf{H}(c\varepsilon)/\mathsf{H}(\varepsilon) < \infty$ for all $c > 0$, there exists a sequence l_n increasing to ∞ such that

$$\sum_{j=1}^n c_j^2 = \mathsf{H}(l_n r_n).$$

The conclusion follows from Theorem 2. □

EXAMPLE. Let $\psi(x) = \sqrt{p(x)}$ and denote by $\mathscr{P}_{s,\alpha,C}$ the set of densities supported on $[0, 1]$ satisfying:

(i) $\int_0^1 \psi^{(k)}(x)^2\,dx \leq C$ for $k = 1, 2, \ldots, s$;
(ii) $\int_0^1 (\psi^{(s)}(x) - \psi^{(s)}(x + h))^2\,dx \leq Ch^\alpha$, $0 < \alpha \leq 1$;
(iii) $|\log \psi(x)| \leq C$.

In this example, we consider the problem of estimating $p \in \mathscr{P}_{s,\alpha,C}$. Following Lorentz (1966), we know that the Hellinger metric entropy of $\mathscr{P}_{s,\alpha,C}$ is of

order $\mathsf{H}(\varepsilon) = (1/\varepsilon)^{1/\beta}$, where $\beta = s + \alpha$. Whence the solution to $\mathsf{H}(r) = nr^2$ is $r_n = (1/n)^{\beta(1+2\beta)}$. In Theorem 9, we show that there exists a sequence of density estimators $\hat{p}_n(x)$ for which the sequence $n^{\beta(1+2\beta)}h(p, \hat{p}_n)$ is tight. By Theorem 9, we conclude that the set of densities in $\mathscr{P}_{p,\alpha,C}$ which can be estimated at a rate faster than $n^{-2\beta/(1+2\beta)}$ in Kullback distance is asymptotically Hellinger negligible. This gives another sense, other than minimax, in which the rate $n^{-\beta/(1+2\beta)}$ is optimal to estimate densities in $\mathscr{P}_{s,\alpha,\mathscr{C}}$.

APPENDIX A: TECHNICAL LEMMAS

DEFINITION 4. Let $(\mathscr{M}, d)$ be a metric space and denote by $\mathscr{B}(x, r) = \{y: d(y, x) < r\}$, the open ball of radius r and center x. A measure μ on $(\mathscr{M}, d)$ is said to be invariant if, for each r, μ assigns the same mass to $\mathscr{B}(x, r)$ for all $x \in \mathscr{M}$; that is, for each r,

$$\mu(\mathscr{B}(x, r)) \equiv \mu(\mathscr{B}(y, r)) \quad \text{for all } y \in \mathscr{M}.$$

It is said to be locally invariant if

$$\lim_{r \to 0} \frac{\mu(B(x, r))}{\mu(B(y, r))} = 1$$

for all pairs $x, y \in \mathscr{M}$.

We remark that, when $\mathscr{M}$ is a linear space, the invariant measures are translation invariant. When $(\mathscr{M}, d)$ is a space of densities with Hellinger metric, the locally invariant measure gives small Hellinger balls the same measure, an idea due to Jeffreys (1946). In particular, if $h^2(p_\theta, p_\eta) \asymp 4^{-1}(\theta - \eta)^t I(\theta)(\theta - \eta)$ as $\eta \to 0$ in $\mathbb{R}^k$, then the locally invariant measure has density on the parameter set proportional to $(\det I(\theta))^{1/2}$. That is, our results naturally allow superefficiency on the subset of the parameter space where $\det(I(\theta)) = 0$.

LEMMA 10. *Let (Θ, d) be a sigma-compact metric space with finite metric dimension. Assume there exists a locally invariant measure μ which assigns finite mass to every open ball $\mathscr{B}$. Then a set A is negligible in Θ with respect to the metric $d(\cdot, \cdot)$ if and only if it has outer μ-measure* 0.

REMARK. In $\mathbb{R}^k$, sets with zero outer μ-measure have Lebesgue measure 0.

PROOF OF LEMMA 10. We can assume that Θ is compact. If not, there exists a countable collection of compact sets with nonempty interior K_j, such that $\Theta \subset \bigcup_{j=1}^{\infty}(\Theta \cap \mathsf{K}_j)$, and the proof is done on each set $(\Theta \cap \mathsf{K}_j)$ separately.

Denote by $\mathscr{B}(x, r)$ the open ball of radius r centered at x. Since the measure μ is locally invariant, there exist functions $m(\varepsilon)$ and $M(\varepsilon)$ such that $m(\varepsilon) \le \mu(B(x, \varepsilon)) \le M(\varepsilon)$ for all $x \in K$ and such that $M(\varepsilon)/m(\varepsilon) \to 1$

as $\varepsilon \to 0$. It then follows from the definition of outer μ-measure [see Halmos (1988), Chapter 1.10] that the outer measure of any set $A \subset K$ is bounded by

$$\mu^*(A) \le N(A, \varepsilon) M(\varepsilon).$$

The minimum number of balls of radius ε covering A is less than the maximum number of balls of radius $\varepsilon/2$ packing the set $A_{\varepsilon/2} := \{y: \inf\{d(x, y): x \in A\} \le \varepsilon/2\}$. By the definition of inner measure and the local invariance of μ, this packing number is less than $\mu_*(A_{\varepsilon/2})/m(\varepsilon/2)$. Thus

$$\frac{\mu^*(A)}{M(\varepsilon)} \le N(A, \varepsilon) \le \frac{\mu\left(A_{\varepsilon/2}\right)}{m(\varepsilon/2))},$$

so that

$$\frac{m(\varepsilon/2)}{M(\varepsilon)} \frac{\mu^*(A)}{\mu(\Theta)} \le \frac{N(A, \varepsilon)}{N(\Theta, \varepsilon)} \le \frac{M(\varepsilon)}{m(\varepsilon/2)} \frac{\mu_*\left(A_{\varepsilon/2}\right)}{\mu(\Theta)}.$$

Since Θ has finite metric dimension and μ is locally invariant, $M(\varepsilon)/m(\varepsilon/2)$ converges to a strictly positive and finite constant as ε converges to 0. The conclusion follows by noting that $\mu_*(A_{\varepsilon/2}) \le \mu^*(A_{\varepsilon/2})$ and by the continuity of the outer measure as ε goes to 0. □

LEMMA 11. *Let the parameter space Θ have finite metric dimension k and a locally invariant measure μ on Θ. If for μ-a.e. θ the sequence of random variables $r_n^{-1} d(\theta, \tilde{\theta}_n)$ is tight under P_θ^n, then Assumption* C *is satisfied.*

PROOF. Let K be a compact set in Θ. Fix $0 < \eta < 1$, and consider the set

$$\mathscr{B}_l = \left\{\theta \in \Theta \cap \mathsf{K}: \sup_n P_\theta^n\left[d\left(\tilde{\theta}_n, \theta\right) > l r_n\right] > \eta\right\}.$$

From Lemma 10, there exists a constant C such that

$$\lim_{\varepsilon \to 0^+} \frac{N(\mathscr{B}_l, \varepsilon)}{N(\Theta \cap \mathsf{K}, \varepsilon)} \le \mathsf{C} \frac{\mu^*(\mathscr{B}_l)}{\mu(\Theta \cap \mathsf{K})}. \tag{22}$$

Tightness of $d(\tilde{\theta}_n, \theta)/r_n$ implies that the sets $\mathscr{B}_l$ are decreasing to a set of measure 0 as l goes to ∞. The dominated convergence theorem thus implies that the right-hand side of (22) converges to 0 with $l \to \infty$. □

LEMMA 12. *Let $\{P_\theta: \theta \in \Theta\}$ be a dominated family of distributions and assume that Θ is a measurable space. With a prior probability distribution W on Θ, define the mixture $Q(\cdot) = \int P_\eta(\cdot) W(d\eta)$. Then for any measurable $B \subset \Theta$ the Kullback–Leibler loss is bounded by*

$$\begin{aligned} D(p_\theta \| q) &\le \log \frac{1}{W(B)} + \int_B D\left(p_\theta \| p_\eta\right) \frac{W(d\eta)}{W(B)} \\ &\le \log \frac{1}{W(B)} + D(\theta \| B), \end{aligned}$$

where, for any subset $B \subset \Theta$, $D(\theta \| B) = \sup_{\eta \in B} D(p_\theta \| p_\eta)$.

PROOF. Denote by $q(x) = \int p(x|\eta)W(d\eta)$ the density of the mixture Q with respect to the dominating measure of the family of distributions $\{P_\theta: \theta \in \Theta\}$. Integrating only over the set B bounds

$$q(x) \geq \int_B p(x|\eta)W(d\eta) = W(B)\int_B p(x|\eta)\frac{W(d\eta)}{W(B)},$$

where $W(B) = \int_B W(d\eta)$. Applying this and Jensen's inequality, the Kullback–Leibler loss is bounded by

$$\begin{aligned} D(p_\theta\|q) &= \int \log\frac{p(x|\theta)}{q(x)}p(x|\theta)\nu(dx) \\ &\leq \log\frac{1}{W(B)} + \int \log\frac{p(x|\theta)}{\int_B p(x|\eta)W(d\eta)/W(B)}p(x|\theta)\nu(dx) \\ &\leq \log\frac{1}{W(B)} + \int_B \left\{\int \log\frac{p(x|\theta)}{p(x|\eta)}p(x|\theta)\nu(dx)\right\}\frac{W(d\eta)}{W(B)} \\ &= \log\frac{1}{W(B)} + \int_B D(p_\theta\|p_\eta)\frac{W(d\eta)}{W(B)}. \end{aligned}$$

By the definition of $D(\theta\|B)$, the latter is further bounded by

$$D(p_\theta\|q) \leq \log\frac{1}{W(B)} + D(\theta\|B),$$

proving the claims. □

Acknowledgments. We gratefully acknowledge the constructive comments of two anonymous referees, an Associate Editor and Larry Brown that have helped to improve the presentation of our results.

REFERENCES

ALI, S. and SILVEY, S. (1966). A general class of coefficient of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142.

BAHADUR, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.* **35** 1545–1552.

BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

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

BARRON, A. and COVER, T. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054.

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

BROWN, L. D. (1993). An information inequality for the Bayes risk under truncated squared error loss. In *Multivariate Analysis: Future Directions* (C. R. Rao, ed.). North-Holland, Amsterdam.

BROWN, L. D., LOW. M. G. and ZHAO, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25** 2607–2625.

CENCOV, N. N. (1982). Statistical decision rules and optimal inference. *Amer. Math. Soc. Transl. Ser. 2* **53**.

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

COVER, T. and THOMAS J. (1991). *Information Theory*. Wiley, New York.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318.

HALMOS, P. R. (1988). *Measure Theory*, 4th, ed. Springer, New York.

HARTIGAN, J. A. (1983). *Bayes Theory*. Springer, New York.

HARTIGAN, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A* **186** 453–461.

KOMAKI, F. (1994). On asymptotic properties of predictive distributions. Technical Report, Dept. of Mathematical Engineering and Information Physics, Faculty of Engineering, Univ. Tokyo.

KULLBACK, S. and LEIBLER, R. (1951). Information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1982). *Statistical Estimation: Asymptotical Theory. Applications of Mathematics*. Springer, New York.

LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. California Publ. Statist.* 277–330.

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

LORENTZ (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc.* **72** 903–937.

MERHAV, N. and FEDER, M. (1995). A strong version of the redundancy–capacity theorem of universal coding. *IEEE Trans. Inform. Theory* **41** 714–722.

RISSANEN, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **30** 629–636.

RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.

VOVK (1991). Asymptotic efficiency of estimators: algorithmic approach. *Theory Probab. Appl.* **36** 329–343.

YANG, Y. and BARRON, A. (1995). Information theoretic determination of minimax rates of convergence. Unpublished manuscript.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06520-8290
E-MAIL: barron@stat.yale.edu
nicolas.hengartner@yale.edu