Information Theory in Probability, Statistics, Learning, and Neural Nets

Andrew R. Barron *
Department of Statistics
Yale University
New Haven, CT 06520-8290
barron@stat.yale.edu

July 5, 1997

Abstract

Probabilistic information theory determines fundamental limits in the topics of data compression, channel capacity, thermodynamics, statistical estimation, prediction, hypothesis testing and related topics. Moreover, information theory provides an illuminating perspective on the limit theorems of probability. Beginning with some identities and inequalities for relative entropy, followed by their use in solving some problems in probability and statistics, and concluding with application to neural networks, this talk reviews the role of information theory in answering interesting questions in these topics.

1 INTRODUCTION

In probability theory, basic limits properties can be better understood by examination of the role of the relative entropy or Kullback divergence $D(P||Q)$ between pairs of distributions. It provides the exponent of the probability of large deviations and in particular the tail probability in the law of large numbers. It characterizes, via a minimization, the conditional limit distribution when conditioning on large deviation events. Chain rules for D for various choices of P_n and Q_n quantify a change in $D(P_n||Q_n)$ yielding a proof of convergence of the distribution of a Markov Chain where n is the number of steps, a proof of convergence of martingales, and a proof of the central limit theorem with a stronger mode of convergence than is traditional. The monotonicity

*Supported by NSF grants DMS-9505168 and ECS-9410760

properties and especially the chain rule reveals the naturalness of the choice of relative entropy in the investigation of these limits.

In statistics and learning theory, including neural net examples, the relative entropy plays a basic role in both classical procedures for estimation and in theoretical characterization of the limits of achievable performance. Consistency of maximum likelihood and Bayes estimates of a parameter and the identity of the limiting parameter value are examined via examination of convergence of empirical versions of the relative entropy to $D(P_X||P_{X|\theta})$. Quantification of efficiency involves the Hessian J_θ of second derivatives of $D(P_X||P_{X|\theta})$ which is the Fisher information. The local self-standardization of relative entropy via this Fisher information in its quadratic expansion reveals asymptotic risk of the form $k/2n$ where k is the number of parameters and n is the sample size. The cummulative relative entropy risk of estimators reduces via a chain rule to a total relative entropy $D(P_{X_1,\ldots,X_N}||Q_{X_1,\ldots,X_N})$. From this the individual risk of order $k/2n$ and the cummulative risk of order $(k/2)\log N$ are shown to not be beaten except for a negligible set of distributions. An index of resolvability gives a simple means to demonstrate cummulative risk bounds of the right order for Bayes estimators in both parametric and nonparametric settings. The minimax cummulative relative entropy risk is the same as the minimax redundancy of universal data compression, the same as the information capacity of the channel from the parameter to the data, and it is shown to be the same to within a constant factor as the Kolmogorov ϵ-entropy at a critical seperation ϵ_N. A consequence is characterization of the minimax rate for estimation in infinite-dimensional or nonparametric problems. Sequences of parametric families achieve these optimal rates in various nonparametric problems as a tradeoff between approximation and estimation error of order $\min_k\{D_k + k/n\}$. In suitable settings both Bayes mixtures and model selection by penalized likelihood achieve the optimal rates.

In a preliminary section we give some basic definitions and tools used in subsequent sections, including relationships between information quantities, a pythagorean property of information, the chain rule, and the asymptotic equipartition property. The relative entropy is the principle quantity that arises in the topics we study. The variance of log-density ratios and a relativized Fisher information are also introduced and related to relative entropy.

The third section begins the main body of the review. There we examine the role of information quantities in basic limit theorems of probability including the law of large numbers, large deviations, a conditional limit theorem, martingales, and the central limit theorem. The fourth section discusses information-theoretic bounds for convergence of Markov chains to the stationary distribution. The fifth section addresses roles of information theory in statistics and learning with application to neural nets, and a final section interpretes some of the conclusions in terms of data compression.

For the role of information theory in many of these topics, the books by Kullback [Kul59] and Cover and Thomas [CT91] give excellent introductions,

though as the reader will see, I provide different emphases and review substantial ground not covered therein, particularly with regard to characterization of achievable performance in statistical estimation, learning, universal data compression, bounds on Bayes procedures, and the role of information theory in probability.

2 PRELIMINARIES

Here we give some basic definitions and tools used in subsequent sections, including definition of information quantities and relationships between them, a pythagorean property of information, the chain rule, and the asymptotic equipartion property.

Before discussing informational divergence or relative entropy between distributions we mention the Shannon entropy

$$H(X) = H_\lambda(P) = -\int p(x) \log p(x)$$

of a probability distribution P for a random variable X on a measurable space $\mathcal{X}$ with density $p(x)$ with respect to a reference measure λ, usually taken to be a discrete or continuous uniform measure, i.e., counting or Lebesgue measure. Here the integral is understood to be with respect to λ, and $0 \log 0 = 0$. The reference measure is assumed to be sigma-finite on $\mathcal{X}$. In the discrete case $H = \sum_{x \in \mathcal{X}} p(x) \log 1/p(x)$ is a sum of nonnegative terms so then H exists and is nonnegative (though possibly infinite). Another case in which existance is assured (though possibly minus infinity) is when λ is a finite measure, for then it is normalizable to be a probability measure for which $H_\lambda(P) = -D(P||\lambda)$ (see below).

The entropy H arose in the work of Shannon [Sha48] on the number of bits required to represent or to generate a discrete random variable X, and in statistical physics and in Shannon information theory through the asymptotic equipartition property where it characterizes the measure of the typical set (see subsection 2.7) and through the maximum entropy principle (see subsection 3.2).

2.1 Relative Entropy

In this review we will focus on the relative entropy between pairs of probability distributions. Let

$$D(P_X||Q_X) = D(p||q) = E_P \log p(X)/q(X) = \int p(x) \log p(x)/q(x)$$

denote the Kullback-Leibler divergence or relative entropy between distributions $P = P_X$ and $Q = Q_X$ for a random variable X with probability density functions

p and q with respect to some measure λ dominating P and Q, where the integral is understood to be with respect to λ. In most cases we will take λ to be a discrete or continuous uniform measure. To handle some general situations one may take $\lambda = P + Q$. The ratio $\rho(X) = p(X)/q(X)$ will be the same (almost everywhere) for any measures dominating both P and Q. Here and throughout $p(x)/q(x)$ is taken to be zero whenever $p(x)$ is zero and it is taken to be infinite whenever the denominator is zero and the numerator is positive. If the set $\{x : q(x) = 0\}$ has positive P probability then $D(P||Q)$ is infinite. Finite $D(P||Q)$ requires P to be absolutely continuous with respect to Q, in which case one may take $\lambda = Q$ and D may be expressed in terms of the general entropy as $D(P||Q) = -H_Q(P)$. A representation for relative entropy in terms of a non-negative integrand is

$$D(P||Q) = \int (p(x) \log p(x)/q(x) + q(x) - p(x)).$$

From the strict positivity of this integrand when $p(x)$ and $q(x)$ are not equal, it follows that $D(P||Q) \geq 0$ with equality if and only if $P = Q$.

We note that $D(P||Q)$ is convex in P and Q and hence the information neighborhoods $\{P : D(P||Q) \leq \delta^2\}$ and $\{Q : D(P||Q) \leq \delta^2\}$ are convex sets of distributions.

The relative entropy D occurs naturally in a number of ways that will be covered in this paper, for instance, as a probability exponent in hypothesis tests and large deviations, as a measure of risk and cumulative risk of predictive distributions, and as the excess expected codelength in data compression necessitated by lack of knowledge of the governing distribution. Optimization of D for a family of distributions identifies the conditional limit in conjunction with large deviations, characterizes the Gaussian limit in central limit theory, characterizes the limit of likelihood-based statistical procedures, bounds the resolvability and minimax risk for various measures of statistical loss, determins the minimax redundancy in universal data compression, and determines the capacity of communication channels.

2.2 Relations Between Measures of Divergence

Though principly the relative entropy will provide the answers to questions we address here, it will be useful to relate it to other notions of divergence. In particular, we consider the L_1 distance $\int |p - q|$ (which is the total variation distance between the distributions), the squared Hellinger distance $h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2$, the Chi-square distance $\int (p - q)^2/q$, and the Renyi relative entropies $(1/(\alpha - 1)) \log \int p(p/q)^{\alpha - 1}$ for $\alpha > 0$. The Renyi relative entropies approach $D(p||q)$ from below as $\alpha \nearrow 1$ and from above as $\alpha \searrow 1$. All of the above measures of divergence are monotone increasing functions of f-divergences

[Csi67, AS66] which have the representation

$$D_f(p||q) = \int pf(q/p)$$

with f convex and equal to zero when the ratio q/p is 1. From the convexity of f these divergences have the property that they are non-increasing under (deterministic or stochastic) transformation of the random variable.

For f twice differentiable with continuous and positive second derivative at 1 these divergence are locally equivalent to each other (to within multiplicative constants) at least for p/q uniformly close to one. It appears that among these divergences (with twice differentiable f) only the squared Hellinger distance $h^2(p,q)$ is a squared metric. In this way the Hellinger metric is essential to the behavior of the local topology with any of these divergences, including relative entropy.

Other measures of discrepancy between densities that we shall encounter include the second moment of the log density ratio $\int p(\log p/q)^2$, and, for random vectors in R^d with everywhere differentiable densities, a relativized Fisher information defined by $J(p||q) = \int p(x)||\nabla \log p(x)/q(x)||^2 dx$.

The relative entropy D is less than the Chi-square, greater than the squared Hellinger, and greater than $(1/2)(L_1)^2$, see [Csi67, Kul67]. A number of additional inequalities can be established when there is a bound on the density ratio. For instance, if $q(x)/p(x) \leq v$ for all x in the support of p, then as in [YB97a],

$$\frac{1}{2+\log v} \int p(\log p/q)^2 \leq D(p||q) \leq \frac{v}{2} \int p(\log p/q)^2.$$

For sequences of pairs of distributions with densities, convergence of D to zero implies Hellinger and L_1 convergence of the difference of the densities, entailing total variation convergence of the difference of the distributions. The expected absolute value $E_P|\log p(X)/q(X)|$ is not greater than $D + \sqrt{2D}$, see [Pin64, Bar86]. Thus convergence of D to zero also implies an $L_1(P)$ convergence of $\log p(X)/q(X)$ to zero.

For any sequence of pairs of density functions with ratio $p(x)/q(x)$ converging uniformly to 1,

$$D(p||q) \sim 2h^2(p,q)$$

in the sense that the ratio of the two sides tends to one. Similarly, $D \sim (1/2)\int (p-q)^2/q$ and $D \sim (1/2)\int p(\log p - \log q)^2$, again as p/q converges uniformly to one.

2.3 Relative Entropy and Relativized Fisher Information

The relativized Fisher information will arise in cases where we take the increment of relative entropy associated with the addition of a small multiple of a random vector Z independent of X. Supppose Z has covariance equal to the

identity matrix I and suppose X also has finite second moments. The two distributions P and Q for X are assumed to have densities p and q, with respect to Lebesgue measure on R^d. Let p_τ and q_τ denote the density of $X + \sqrt{\tau}Z$ when X has density p or q, respectively. Under regularity conditions on derivatives of the densities p and q one finds that the derivative of $D(p_\tau||q_\tau)$ with respect to τ is continuous and that evaluated at $\tau = 0$ the derivative is equal to -$(1/2)J(p||q)$, where

$$J(p||q) = \int p(x)||\nabla_x \log p(x)/q(x)||^2 dx$$

is the relativized Fisher information. Thus

$$D(p_\tau||q_\tau) = D(p||q) - \frac{\tau}{2} J(p||q) + O(\tau^2).$$

When Z is a standard normal random vector and P and Q are arbitrary distributions on R^d (not necessarily continuous) with finite second moments, we may replace X by the random vector $X + \sqrt{\tau}Z$ with fixed $\tau > 0$, which has smooth densities p_τ and q_τ that satisfy the required regularity. Now adding an additional independent normal vector preserves the form of the distribution (it amounts to increasing τ). Thus we find for all $\tau > 0$ that

$$\frac{\partial}{\partial \tau} D(p_\tau||q_\tau) = -\frac{1}{2} J(p_\tau||q_\tau).$$

Now $D(p_\tau||q_\tau)$ converges to $D(P||Q)$ as $\tau \to 0$ and to 0 as $\tau \to \infty$. Consequently, the following integral representation holds

$$D(P||Q) = \frac{1}{2} \int_{\{\tau > 0\}} J(p_\tau||q_\tau) d\tau.$$

This integral identity extends the result in [Bar86] which was for the case that the second measure Q is normal, though the proofs are much the same.

Similar identities for relative entropy arises in conjuction with a process X_τ that evolves for $\tau \geq 0$ according to the stochastic differential equation

$$dX_\tau = (1/2)\nabla \log p(X_\tau) d\tau + dZ_\tau,$$

where Z_τ is a standard Brownian motion. Suppose here that $p(x)$ is continuously differentiable. Let Q_τ and $P_\tau = P$ be the marginal distributions for X_τ when the initial distribution for X_0 is either Q or P, respectively. Then Q_τ approaches the stationary distribution P in a manner analogous to a discrete-time stochastic gradient model studied in section 3.2. Examination of the discrete-time approximation suggests that via stochastic calculus we again have $\frac{\partial}{\partial \tau} D(P||Q_\tau) = -(1/2)J(P||Q_\tau)$ and $\frac{\partial}{\partial \tau} D(Q_\tau||P) = -(1/2)J(Q_\tau||P)$, yielding

$$D(P||Q) = \frac{1}{2} \int_0^\infty J(P||Q_\tau) d\tau$$

and

$$D(Q||P) = \frac{1}{2} \int_0^\infty J(Q_\tau||P) d\tau.$$

2.4 Pythagorean Relation

Though it is not a squared metric, the relationship between relative entropy and other squared distances and Pythagorean identities stated here show that relative entropy behaves geometrically in a way analogous to squared Euclidean distance.

If $\mathcal{Q}$ is a convex set of distributions and we denote $D(\mathcal{Q}||P) = \inf_{\mathrm{Q}\in\mathcal{Q}} D(\mathrm{Q}||P)$ as the divergence of P from the set $\mathcal{Q}$, then as shown in [Top79, Csi84] there is a unique information projection P^* such that for all $\mathrm{Q} \in \mathcal{Q}$,

$$D(\mathrm{Q}||P) \geq D(\mathrm{Q}||P^*) + D(\mathcal{Q}||P),$$

and consequently any sequence of distributions Q_n in $\mathcal{Q}$ for which the divergence tends to the infimum $D(\mathcal{Q}||P)$ must have Q_n converging to P^* (indeed $D(Q_n||P^*) \to 0$). Equality holds in the Pythagorean relation

$$D(\mathrm{Q}||P) = D(\mathrm{Q}||P^*) + D(P^*||P)$$

when $\mathcal{Q}$ is a hyperplane of distributions such as $\{\mathrm{Q} : E_{\mathrm{Q}} f(X) = a\}$ and P^* is in $\mathcal{Q}$ achieving $D(P^*||P) = D(\mathcal{Q}||P)$, in which case there is an exponential family characterization of this information projection [Csi75].

2.5 Chain Rule

When random variables $X_1, X_2, \ldots, X_n$ have joint densities that factor as a product of conditionals $p(X_1, \ldots, X_n) = \prod_{i=1}^n p(X_i|X^{i-1})$ and $q(X_1, \ldots, X_n) = \prod_{i=1}^n q(X_i|X^{i-1})$ the chain rule yields

$$E_{P_{X^n}} \log \frac{p(X_1, X_2, \ldots, X_n)}{q(X_1, X_2, \ldots, X_n)} = \sum_{i=1}^n E_{P_{X^n}} \log \frac{p(X_i|X^{i-1})}{q(X_i|X^{i-1})}.$$

Thus the total relative entropy between the joint distributions is a sum of expected relative entropies between the conditional distributions

$$D(P_{X^n}||Q_{X^n}) = \sum_{i=1}^n E_{P_{X^{i-1}}} D(P_{X_i|X^{i-1}}||Q_{X_i|X^{i-1}}).$$

2.6 Asymptotic Equipartition and Hypothesis Tests

Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed (i.i.d.) with marginal distribution either P or Q. By the Chain rule $(1/n)D(P_{X^n}||Q_{X^n}) = D(P||Q)$ and if the random variables are distributed according to P then by the law of large numbers $(1/n)\log p(X^n)/q(X^n)$ converges with probability one to $D = D(P||Q)$. A suitable notion of divergence $\mathcal{D}$ between processes, as a limit of the expected conditional relative entropies in the chain rule decomposition,

and a corresponding almost sure limit theorem for $(1/n)\log p(X^n)/q(X^n)$ are given in [Ore85, Bar85] when the process is stationary and ergodic with repect to the governing measures P_{X^n} under restrictions on the dominating measures Q_{X^n}, extending earlier work of Shannon, McMillan, and Breiman in which Q_{X^n} was restricted to be discrete uniform.

Let $\rho(X^n) = p(X^n)/q(X^n)$ be the density ratio and suppose $(1/n)\log\rho(X^n)$ converges to $\mathcal{D}$ in P_{X^n}-probability, with finite $\mathcal{D}$. The key consequence of this convergence is the role of the set of X^n typical for P against Q, defined for $\epsilon > 0$ by

$$A_{n,\epsilon} = \{e^{n(\mathcal{D}-\epsilon)} \leq \rho(X_1, \ldots X_n) \leq e^{n(\mathcal{D}+\epsilon)}\}.$$

The distribution is nearly concentrated on this typical set, that is $\lim_n P(A_{n,\epsilon}) = 1$, it is nearly equipartitioned (nearly uniformly distributed relative to Q_{X^n}) within the typical set, the $Q_{X^n}(A_{n,\epsilon})$ measure of the typical set is between $e^{-n(\mathcal{D}+\epsilon)}$ and $e^{-n(\mathcal{D}+\epsilon)}$, and no sequence of high P probability sets can have asymptotically smaller Q measure, that is, if $P_{X^n}(B_n) \geq 1-\alpha$ for some $0 < \alpha < 1$ then $Q_{X^n}(B_n) \geq e^{-n(\mathcal{D}+o(1))}$. A one-sided version of the typical set $\{\rho(X^n) \geq e^{n(\mathcal{D}-\epsilon)}\}$ is a set of smallest Q_{X^n} measure among sets that share its high P_{X^n} probability (in accordance with the Neyman Pearson Lemma in the hypothesis testing interpretation) and often it can be used in place of $A_{n,\epsilon}$.

This collection of results is known as the asymptotic equipartition property (AEP) of information theory and statistical mechanics (usually with Q_{X^n} uniform). In hypothesis testing it is known as the Chernoff-Stein Lemma characterizing $e^{-n(D+o(1))}$ as the best exponential convergence of Q probability of error in a test of P versus Q [Che56]. See also [CT91] for these interpretations.

2.7 Divergence from Mixtures, Mutual Information, and Capacity

In information theory and statistics a role is often played by the divergence $D(P_{X|\theta}||Q_X)$ between members of a family of distributions $P_{X|\theta}, \theta \in \Theta$ and fixed distributions Q_X on X, often taken in the form of mixtures $P_X = P_X^{(W)} = \int P_{X|\theta} W(d\theta)$ for probability measures W on Θ. Here Q_X may be interpreted as a distribution for X in the absense of knowledge of θ. The average divergence $\int D(P_{X|\theta}||Q_X)W(d\theta)$ is the relative entropy $D(P_{\theta,X}||P_\theta \times Q_X)$ of the joint distribution for θ and X from the product distribution $P_\theta \times Q_X$ when $P_\theta = W$. Chain rule expansion shows it to equal $D(P_{\theta,X}||P_\theta \times P_X) + D(P_X||Q_X)$, which is uniquely minimized by the choice of $Q_X = P_X$. The resulting minimized average divergence takes the form $D(P_{\theta,X}||P_\theta \times P_X)$ which is known as the Shannon mutual information $I(\theta; X)$ between θ and X. The minimax divergence is $V = \min_{Q_X} \max_\theta D(P_{X|\theta}||Q_X)$ and the maximin average divergence is $C = \max_W \min_{Q_X} \int D(P_{X|\theta}||Q_X)W(d\theta) = \max_W I_W(\theta; X)$ which is known as the Shannon information capacity of the family of distributions $\{P_{X|\theta}, \theta \in \Theta\}$ (aka

channel). This game-theoretic examination of relative entropy is from [Dav73] in a data compression context.

As to be expected from decision theory and game theory for convex and lower-semicontinuous loss functions [Fer67], the minimax divergence V and the maximin average divergence C are the same [Gal79, DLG80], even in the generality that $\mathcal{X}$ is a complete separable metric space [Hau95]. To continue the decision theoretic terminology, a probability measure W on θ is a prior, the mixture $P_X^{(W)}$ is a Bayes strategy, and such mixtures are admissible (for any such mixture P_X there is no Q_X for which $D(P_{X|\theta}||Q_X)$ is made everywhere as small and somewhere smaller than with P_X).

The divergence from mixtures $D(P_{X|\theta}||P_X)$ arises in data compression as the excess average codelength or redundancy of a code for X in the absence of knowledge of θ (see section 5), in hypothesis testing as a probability of error exponent for simple versus composite hypotheses (see [CB90]), in prediction with $X = (X_1, \ldots, X_N)$ as the cumulative relative entropy risk (see section 4), in gambling as the cumulative expected log wealth regret, and (averaging over θ) it arises in communication channels as an achievable rate of communication.

An index of resolvability will be used in section 4 to give upper bounds on divergence from mixtures. The following simple inequality is used to give lower bounds.

2.8 Inequality between Relative Entropy and Probabilities of Events

For any distributions P and Q for a random variable X and any measurable $A \subset \mathcal{X}$,

$$D(P||Q) \geq P(A) \log 1/Q(A) - \log 2.$$

This inequality is a consequence of the monotonicity of D under transformation. Indeed, the chain rule applied to (X, T) where $T = 1_A(X)$ gives $D(P_X||Q_X)$ not less than the binary relative entropy $D(P_T||Q_T) = P(A) \log P(A)/Q(A) + (1 - P(A)) \log(1 - P(A))/(1 - Q(A))$, which is further lower bounded by throwing away the $(1 - P(A)) \log 1/(1 - Q(A))$ term and bounding the binary entropy $H_2 = -P(A) \log P(A) - (1 - P(A)) \log(1 - P(A))$ by $\log 2$. In the applications of the inequality I give here the $\log 2$ may be ignored since it will typically be small compared to the total relative entropy between distributions for a sequence $X = (X_1, \ldots, X_n)$.

The above inequality gives a lower bound on $D(P||Q)$ in terms of $\log 1/Q(A)$ for events with $P(A)$ near one. It can be used in this form to prove Rissanen's results on the negligibility of superefficient data compression and corresponding results on the negligibility of superefficient estimation (see section 4.12).

The inequality may be rewritten in the form

$$P(A^c) \geq 1 - \frac{D(P||Q) + \log 2}{\log 1/Q(A)}.$$

Using the inequality in this form with $A = A_{r,\theta} = \{X : L(\hat{\theta}(X), \theta) \leq r\}$ for a loss function L and estimator $\hat{\theta}$ with $P = P_{X|\theta}$ one obtains a lower bound on the tail probability $P_{X|\theta}\{L(\hat{\theta}, \theta) \leq r\}$ and lower bounds on the risk $E_{P_{X|\theta}} L(\hat{\theta}, \theta) \geq rP_{X|\theta}(A^c)$. Here the idea is to find the critical distance r such that $D(P||Q)$ is small compared to $\log 1/Q(A_{r,\theta})$. In the above form the inequality is an analog of Fano's inequality in which the distribution Q is the mixture with respect to a uniform distribution on a finite set of values for θ, the event $A = \{\hat{\theta}(X) = \theta\}$ and one averages over θ. Such inequalities are used both in statistics and channel capacity to show that rates of estimation and communication have fundamental limits (see section 4.13 and section 5).

The inequality may also be rewritten in the form

$$Q_X(A) \geq e^{-(D(P_X||Q_X)+\log 2)/P_X(A)}.$$

Letting $X = (X_1, \ldots, X_n)$, one may use the inequality in this form as one way to prove the fact in the AEP that if A_n is any sequence of sets with $P_{X^n}(A_n)$ converging to one then $Q_{X^n}(A_n) \geq e^{-n(\mathcal{D}+o(1))}$ where $\mathcal{D} = \limsup(1/n)D(P_{X^n}||Q_{X^n})$. Thus the converse half of the AEP does not require stationarity or ergodicity of the processes.

2.9 Conditioning on an Event

Let X be a random variable (or vector) and let B be a set for which the distribution P_X assigns positive probability. The conditional distribution $P_{X|B}$ has density $1_B(x)/P_X(B)$ with respect to P_X and hence

$$D(P_{X|B}||P_X) = \log 1/P_X(B).$$

Thus there is an exact relative entropy expression for the probability of events

$$P\{X \in B\} = e^{-D(P_{X|B}||P_X)}.$$

3 PROBABILITY

Here we examine the role of information quantities in basic limit theorems of probability including the law of large numbers, large deviations, a conditional limit theorem, martingales, central limit theorems, and convergence of Markov chains to the stationary distribution. The chain rule and representations for increments of relative entropy will be our main tools.

3.1 Large Deviations and the Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be i.i.d. and let $B_n = \{\hat{P}_n \in \mathcal{Q}\}$ be the event that the empirical distribution is in a convex set $\mathcal{Q}$. The empirical distribution is defined

by $\hat{P}_n(A) = (1/n)\sum_{i=1}^n 1_A(X_i)$ for $A \subset \mathcal{X}$. Large deviations is concerned with the asymptotic behaviour of $P\{\hat{P}_n \in \mathcal{Q}\}$ when P is not necessarily in $\mathcal{Q}$. In particular, if $\mathcal{Q}$ equals $\{Q : E_Q f(X) \geq \mu + \epsilon\}$ or $\{Q : E_Q f(X) \leq \mu - \epsilon\}$ with $E_P f(X) = \mu$ and $\epsilon > 0$, then $P\{\hat{P}_n \in \mathcal{Q}\}$ provides the tail probabilities of the distribution of the sample average $(1/n)\sum_{i=1}^n f(X_i)$ associated with the law of large numbers. A particularly nice treatment of large deviations (and conditional limits discussed below) is in Csiszár [Csi84] using several of the probabilistic information theory ideas discussed above. From the conditioning identity and the chain rule, Csiszár showed that

$$P\{\hat{P}_n \in \mathcal{Q}\} \leq e^{-n\,D(P_{X_1|\hat{P}_n \in \mathcal{Q}}||P)} \leq e^{-n\,D(\mathcal{Q}||P)}$$

if $\mathcal{Q}$ is completely convex. When $D(\mathcal{Q}||P)$ is positive this proves exponential convergence of the probability to zero (which provides in particular a form of law of large numbers). Now $\mathrm{Q}\{|\hat{P}_n(A) - \mathrm{Q}(A)| \leq \epsilon\} \to 1$ for each set $A \subset \mathcal{X}$. Consequently, for a suitable notion of the interior $\mathcal{Q}^o$ as defined in [Csi84] one has that $\mathrm{Q}\{\hat{P}_n \in \mathcal{Q}\} \to 1$ for each Q in $\mathcal{Q}^o$. Then by the optimality of the exponent $D(Q||P)$ in the AEP or Chernoff-Stein Lemma (here with the roles of Q and P reversed) $P\{\hat{P}_n \in \mathcal{Q}\} \geq e^{-n(D(\mathrm{Q}||P)+o(1)}$ for each Q in $\mathcal{Q}^o$ and hence

$$P\{\hat{P}_n \in \mathcal{Q}\} \geq e^{-n(D(\mathcal{Q}^0||P)+o(1))}.$$

Suppose P has the same distance from both $\mathcal{Q}$ and its interior $\mathcal{Q}^0$, that is, $D(\mathcal{Q}^o||P) = D(\mathcal{Q}||P)$. (This is true for instance in the case that $\mathcal{Q} = \{\mathrm{Q} : E_{\mathrm{Q}} f(X) \geq a\}$ with a in the interior of the set of expected values achievable by the exponential family through P using exponent proportional to $f(X)$.) Then as in [Csi84] one has

$$P\{\hat{P}_n \in \mathcal{Q}\} = e^{-n(D(\mathcal{Q}||P)+o(1))}.$$

Thus relative entropy identifies the large deviations exponent. Under conditions on the range and the variance of i.i.d. random variables, Hoeffding, Bennett, and Bernstein inequalities follow via bounds on the exponent $D(\mathcal{Q}||P)$. Unmotivated use of generating functions is not needed for these large deviations proofs.

When $X_1, X_2, \ldots, X_n, \ldots$ form a Markov chain with stationary transitions, an extension given in [Sch93] of this information-theoretic technique identifies the large deviation exponent $\mathcal{D}$ for empirical measures constrained to a set $\mathcal{Q}$. It is equal to $\mathcal{D} = \min D(\mathrm{Q}_{X_1,X_2}||\mathrm{Q}_{X_1}P_{X_2|X_1})$ where the minimum is over all $\mathrm{Q}_{\mathrm{X}_1,\mathrm{X}_2}$ in $\mathcal{Q}$ that satisfy the stationarity constraint $\mathrm{Q}_{X_1} = \mathrm{Q}_{X_2}$. Armed with an evaluation of this information exponent in the case that $\mathcal{Q}$ equals $\{\mathrm{Q}_{X_1,X_2} : E_{\mathrm{Q}_{X_1}} f(X_1) \geq \mu + \epsilon\}$ or $\{\mathrm{Q}_{X_1,X_2} : E_{\mathrm{Q}_{X_1}} f(X_1) \leq \mu - \epsilon\}$ one could in principle determine how long to run a Markov chain such that with high probability the sample average of a function $f(X)$ is within ϵ of its expectation under the stationary distribution.

3.2 Conditional Limit Theorem and Thermodynamics

How should we reassess the distribution of the X_i given the empirical measurement that $\hat{P}_n$ is in the convex constraint set $\mathcal{Q}$? Continuing with the setting of the previous subsection in which $X_1, X_2, \ldots$ are i.i.d., Csiszár [Csi84] showed from the large deviations answer and from the conditioning inequality given above that the sequence of conditional distribution $P_{X_1|\hat{P}_n \in \mathcal{Q}}$ which are in $\mathcal{Q}$ have relative entropy from P converging to $D(\mathcal{Q}||P)$ and hence by the Pythagorean relation the conditional distribution $P_{X_1|\hat{P}_n \in \mathcal{Q}}$ converges to the information projection P^* in the sense that $D(P_{X_1|\hat{P}_n \in \mathcal{Q}}||P^*)$ tends to zero.

This result has interpretation as a proof of the thermodynamic principle stating that the conditional distribution on the microstates X_i for each i, given the macroscopic property $\hat{P}_n \in \mathcal{Q}$, converges in the limit of a large number n of particles to the distribution P^* that minimizes the relative entropy $D(\mathrm{Q}||P)$ among distributions Q that satisfy the macroscopic constraint $\mathrm{Q} \in \mathcal{Q}$. This conclusion agrees with the maximum entropy principle if via suitable transformation the canonical (unconditional) distribution on the microstates can be taken to be uniform. More generally, it is in agreement with Kullback's principle for estimating distributions (given that average measurements correspond to distributions in $\mathcal{Q}$) by minimum discrimination information from a previously believed distribution P.

3.3 Martingales

Basic to the understanding of relative entropy is a monotonicity and convergence property that

$$D(P_n||Q_n) \nearrow D(P||Q)$$

if P_n and Q_n are the restrictions of measures P and Q respectively to an increasing sequence of sigma-algebras of sets that in the limit generate the whole sigma-algebra of sets on which P and Q are defined. In information theory this result with sigma-fields generated by sequences of partitions has been used to show the equality of two definitions of D (one with a supremum over partitions and the other in terms of the integral of densities) arising in the work of Kolmogorov and his colleagues in the 1950's (see [Pin64]). It has also been used to characterize the mutual information $I(X_0; X_1, X_2, \ldots)$ as the limit of $I(X_0 : X_1, X_2, \ldots X_n)$ for any sequence of random variables, by taking the sigma-fields to be generated by $X_0, X_1, \ldots, X_n$ (see [Pin64, KK80]). Similar conclusions hold for conditional entropy and conditional mutual information rates. In this context it is used to demonstrate the vanishing of the gap in the sandwich proof of the Shannon-McMillan-Brieman-Moy-Orey-Barron theorem of [AC88] or the dominated convergence proof of [Ore85, Bar85]). In all of these settings the limit of the information quantities has been proven by appealing to convergence properties of martingales.

I maintain that there is a more direct understanding of the information theory property $D_n \nearrow D$ and that convergence of the appropriate martingales follow as a consequence [Bar91]. The heart of the matter is the chain rule in a suitably general setting. Consider the case that the sequence $D_n = D(P_n||Q_n)$ is bounded (when it is unbounded the convergence will still follow from a demonstration of monotonicity). Let $\rho_n(\omega)$ be the density of P_n with respect to Q_n. Then for $n > m$,

$$\log \rho_n = \log \rho_m + \log \rho_n/\rho_m.$$

Taking expectation with respect to P and using the measurability of $\log \rho_m$ with respect to the smaller sigma-algebra, we have

$$D(P_n||Q_n) = D(P_m||Q_m) + \int \rho_n \log \rho_n/\rho_m$$

where the integral is taken with respect to Q. Now since both ρ_n and ρ_m integrate to one, we have $\int \rho_n \log \rho_n/\rho_m \geq 0$. This shows that D_n is an increasing sequence of numbers. Hence it is a Cauchy sequence, so that $D_n - D_m = \int \rho_n \log \rho_n/\rho_m$ tends to zero as m tends to zero, uniformly over $n \geq m$. By inequalities in section 1.3 this implies that both $\int |\rho_n - \rho_m|$ and $E_P|\log \rho_n - \log \rho_m|$ tend to zero. Hence ρ_n is a Cauchy sequence in $L_1(Q)$ and $\log \rho_n$ is a Cauchy sequence in $L_1(P)$. Let ρ be the resulting limit of ρ_n which then exists by completeness of L_1. By $L_1(Q)$ convergence we deduce that P is absolutely continuous with respect to Q and ρ is the density of P with respect to Q. By $L_1(P)$ convergence of $\log \rho_n$ we have that $D_n = E_P \log \rho_n$ converges to $E_P \log \rho$ which by definition is $D(P||Q)$.

Thus we have a direct proof that $D(P_n||Q_n) \nearrow D(P||Q)$ with convergence of the nonnegative Q-martingales ρ_n as a corollary. The L_1 convergence for more general uniformly integrable martingales ρ_n (not necessarily positive with bounded $\int \rho_n \log \rho_n$) can be treated by reduction to this special case.

3.4 Central Limit Theory

Let $S_n = (X_1 + X_2 + \ldots + X_n)/\sqrt{n}$ be the standardized sum for i.i.d. random variables with mean zero and variance 1. Suppose it has a density function $p_n(s)$ and let $\phi(s)$ be the standard normal density. It is a familiar fact, equivalent to $D(p_n||\phi) \geq 0$, that the normal has the maximum entropy for the given variance and that equality holds only if S_n is normal [CT91]. Thus the idea that the normal should be the limit and the form that this limit should take (exponential family with exponent proportional to s^2) are naturally motived by the constrained maximum entropy principle. But does the density p_n converge to ϕ in the information sense, $D(p_n||\phi) \to 0$? Equivalently, does the entropy of S_n converge to the entropy of a normal random variable as $n \to \infty$? And is this convergence monotone? That is, is $D_n = D(p_n||\phi)$ a decreasing sequence?

The answer from [Bar86] is that, yes, $D(p_n||\phi) \to 0$, if and only if it is eventually finite. Moreover, $nD(p_n||\phi)$ is a subadditive sequence, hence $D(p_n||\phi)$

is convergent and it is monotone alone the powers of 2 subsequence. Doubling the sample size brings us strictly closer to the normal.

The proof that the limit of $D(p_n||\phi)$ is zero is based on the representation of D as an integral of the relativized Fisher information. This relativized Fisher information has an interpretation as a squared L_2 norm of the difference in the score functions (derivatives of the log-densities) between the smoothed density and and associated normal. Pythagorean identities for the L_2 norm and projection properties for the score functions reveal the convergence of the derivatives of the smoothed log-densities in L_2 as shown in [?] building on results of [Bro82]. The integral representation of the relative entropy in terms of the relativized Fisher information [Bar86] then proves convergence to zero of the relative entropy by application of the monotone convergence theorem.

Thus $D(p_n||\phi) \to 0$. Convergence in total variation and in distribution are corollaries. In general for the standardized sum of i.i.d. random variables with mean zero and finite variance, $D(P_n||\Phi)$ might never be finite, indeed, S_n need not have a density. Nevertheless, the classical result of convergence in distribution still follows as a corollary since convergence in distribution is equivalent to convergence in distribution for each positive ϵ for the random variables smoothed by addition of independent normals of variance ϵ, which does produce finite divergence.

3.5 Markov Chains

Let $X_1, X_2, ...$ be a Markov chain on a state space $\mathcal{X}$ with initial distribution $P^{(1)} = P_{X_1}$, transition distribution $P_{X'|X}$, and a stationary distribution P_X. Let $D_n = D(P_X||P_X^{(n)})$ be the relative entropy distance of the distribution $P_X^{(n)} = P_{X_n}$ from the stationary distribution P_X. Each step of the chain brings the distribution closer to stationarity. This can be seen by application of the chain rule applied to expand in both ways the relative entropy $D(P_{X,X'}||P_{X,X'}^{(n)})$ between the joint distribution $P_{X,X'}$ of consecutive states under stationarity and the joint distribution $P_{X,X'}^{(n)} = P_{X_n,X_{n+1}}$ of consecutive states after n steps. As in CT91, the result of this chain rule expansion is

$$D(P_X||P_X^{(n)}) = D(P_{X'}||P_{X'}^{(n+1)}) + E_{P_{X'}} D(P_{X|X'}||P_{X|X'}^{(n)})$$

where $P_{X|X'}$ is the time-reversed conditional distribution when X' has the stationary distribution and likewise $P_{X|X'}^{(n)}$ is the conditional distribution of X_n given X_{n+1}. Here if P_X and $P_X^{(n)}$ have densities $p(x)$ and $p_n(x)$ respectively and $P_{X'|X}$ has a conditional density $p(x'|x)$, then by Bayes rule the time-reversed conditional distributions have conditional densities $p(x)p(x'|x)/p(x')$ and $p_n(x)p(x'|x)/p_{n+1}(x')$, which share the common factor $p(x'|x)$.

Conclusions from the above identity are that $D_n = D(P_X||P_X^{(n)})$ is a decreasing sequence; the difference $r_n = D_n - D_{n+1}$ is itself a relative entropy

$r_n = E_{P_{X'}} D(P_{X|X'}||P^{(n)}_{X|X'})$; if D_n is finite for some n then r_n must must converge to zero; and summing over a number of steps up to n, we find that the Cesaro average satisfies

$$\frac{1}{n}\sum_{k=1}^{n} r_k \leq \frac{D_1}{n}.$$

The same analysis works for the sequence $D(P^{(n)}_X||P_X)$ which has positive increments $E_{P^{(n)}_{X'}} D(P^{(n)}_{X|X'}||P_{X|X'})$ converging to zero with Cesaro average bounded by $D(P^{(1)}_X||P_X)/n$.

To obtain convergence of $P^{(n)}_X$ to P_X, information inequalities arizing from the chain rule are used to prove a Cauchy sequence property for total variation, see [Fri73, Ken64, Ren60].

To obtain bounds for how close to zero is the divergence $D(P^{(n)}_X||P_X)$, what one would look for is for the result of applying the marginal $P^{(n)}_X$ to the two transitions $P^{(n)}_{X|X'}$ and $P_{X|X'}$ to be near each other only if $P^{(n)}_X$ is near P_X. Note the rough similarity with what is required (see section 2.1) to have a not too small exponent for the tail probability of sample averages (large deviations) in a Markov Chain. Bounds for certain types of chains are developed in the next section.

3.6 Probability Reprise

To summarize this probability section, we have seen that the chain rule in conjuction with other information-theoretic inequalities provides simple proofs for the large deviations exponent, the conditional limit theorem, the convergence of martingales, the central limit theorem, and the analysis of Markov chains. The quantities of information theory are natural in that they characterize the large deviations exponent and they characterize the identity of the limit in the conditional limit theorem and the central limit theorem. Moreover, they provide monotonicity of convergence in the central limit theorem and monotonicity of convergence of martingales and Markov chains.

4 Metropolis Chains, Stochastic Gradients, and Distance from Stationarity

In this section I present work in progress on Metropolis chains and stochastic gradient models. The aim is to identify rates of convergence of p_n to p and to determine bounds on measures of distance between them. Depending on the nature of such bounds, we would like to interprete X_n as approximately a draw from the stationary density p for polynomially large n. The eventual goal is to apply such bounds to computational statistics or computational learning

problems in the case that p is the posterior density that arises in Bayes models, so that provably accurate posterior mean functions can be rapidly computed from Monte Carlo averages using several independent runs of the chain.

4.1 Metropolis Chains

Let $p(x)$ be a given density with respect to Lebesque measure on R^d. Metropolis chains are constructed to have a transition $P_{X'|X}$ for which the distribution with density p is the stationary distribution. We choose a pilot distribution for the transitions that is uniform on a small cube of sidelength proportional to δ centered at the current state x. What is essential here is that the pilot distribution has mean equal to the current state, step size bounded by order δ, and covariance of order δ^2 times the identity matrix. For definiteness I choose this covariance to equal $\delta^2 I$. If x' is a candidate point drawn from the pilot distribution then the chain steps there with probability $\min\{p(x')/p(x), 1\}$ and otherwise stays at x.

Assume that $\log p(x)$ and $\log p_n(x)$ have a bound κ on the absolute values of the second derivatives $\frac{\partial^2}{\partial x_i^2} \log p(x)$ and $\frac{\partial^2}{\partial x_i^2} \log p_n(x)$ and the square of the first derivatives $\frac{\partial}{\partial x_i} \log p(x)$ and $\frac{\partial}{\partial x_i} \log p_n(x)$.

The density ratio between $P_{X|X'}$ and $P^{(n)}_{X|X'}$, restricted to x in the δ cube around x', is the ratio of $p(x)/p(x')$ and $p_n(x)/p_{n+1}(x')$ which are uniformly close to one for small δ. From section 1.3 this permits the approximation

$$E_{P_{X'}} D(P_{X|X'}||P^{(n)}_{X|X'}) \sim (1/2) E_{P_{X,X'}} \left(\log \frac{p(X)/p(X')}{p_n(X)/p_{n+1}(X')} \right)^2$$

to within terms that will be seen to be of order δ^3. Examining quantities that arise inside the square on the right side, one can show that $\log p_{n+1}(x')/p_n(x')$ is of order δ^2, which is negligible compared to the difference between $\log p(x)/p(x')$ and $\log p_n(x)/p_n(x')$ for which the first order Taylor expansion equals

$$(x - x')^T \nabla \log p(x')/p_n(x') + O(\delta^2).$$

In this way we find that with $D_n = D(p||p_n)$,

$$D_n - D_{n+1} = (1/2)\delta^2 J(p||p_n) - O(\delta^3)$$

and, in particular, there is a positive universal constant C such that

$$D_n - D_{n+1} \geq (1/2)\delta^2 J(p||p_n) - C d \kappa \delta^3,$$

where $J(p||q) = \int p(x)||\nabla \log p(x)/q(x)||^2 dx$ is the relativized Fisher information and d is the dimension of X. The same representation holds for the differences in the sequence $D(p_n||p)$ but with $J(p_n||p)$ in place of $J(p||p_n)$.

Note the similarity of this representation of the increment of relative entropy to the case in section 1.3 where one adds a small independent random vector of covariance equal to τI with $\tau = \delta^2$. Here the Metropolis perterbations are not independent of the site X, but the representation holds nonetheless. An alternative proof of the above identity is to note that $D_n - D_{n+1} = E_{P_X} \log p_{n+1}(X)/p_n(X)$ and then expand $\log p_{n+1}(X)/p_n(X)$ to second order using the definition of $p_{n+1}(X)$. The terms one gets appear somewhat more complicated, but several of them vanish employing an integration by parts.

The characterization of the drop in relative entropy is here based on a small step size δ. It reveals a positive drop as long as $J(p||p_n)$ remains at least a multiple of δ. To allow via these bounds a demonstration that $J_n = J(p||p_n)$ tends to zero it will be necessary to allow δ_n to shrink slowly with n. Note that this produces time-inhomogeneous transition probabilities but maintains the stationarity of the distribution P.

Now ideally one may have an inequality relationship $D(p||q) \leq \gamma J(p||q)$ (which may be interpreted as a Sobolev inequality [Maz85]), though the constant γ may in some cases be quite large (especially for multimodal p on R^d). If that inequality holds then one may deduce by induction a bound of order $D_n = O(1/\sqrt{n})$, indeed $D_n \leq \max\{D_1, 8Cd\kappa\gamma_p^{5/2}\}/\sqrt{n}$, in conjuction with a choice of $\delta_n = 2\gamma/\sqrt{n}$. The idea in that case is that the decrease in D_n will be at least a certain small positive multiple of D_n. Nonetheless, we are interested here in what error bounds we can extract for the convergence of p_n to p even in the absence of a relationship $D(p||q) \leq \gamma J(p||q)$.

Suppose we set δ_k to decrease to zero fairly slowly, such that the sequence δ_k^2 is not summable, but in such a way that δ_k^3 is a summable sequence. For instance let $\delta_k = (1/k)^r$ with $1/3 < r < 1/2$ and let the sum of cubes be $\xi = \sum_k \delta_k^3$. Summing the inequality above and then dividing by n we have that

$$\frac{1}{n} \sum_{k=1}^{n} \delta_k^2 J_k \leq \frac{A}{n}.$$

where $J_k = J(p||p_k)$ and $A = 2D_1 + C\xi d\kappa$. Consequently,

$$\frac{1}{n} \sum_{k=1}^{n} J_k \leq \frac{A}{n^{(1-2r)}}.$$

This means that if we choose a random number of steps K_n between 1 and n the expected value of the distance J_{K_n} is bounded by $A/n^{(1-2r)}$. In particular there must be a $k_n \leq n$ such that J_{k_n} is not greater than $A/n^{(1-2r)}$. If also J_n is nonincreasing, then $J_n \leq A/n^{(1-2r)}$.

A similar but slightly better bound is obtained by drawing K_n on $\{1, \ldots, n\}$ with probability mass function $\delta_k^2 / \sum_{k'=1}^n \delta_{k'}^2$, which can focus on a somewhat smaller number of steps and achieves $EJ_{K_n} \leq A / \sum_{k=1}^n \delta_k^2$. For a given n we may set $\delta_k = (1/n)^{1/3}$ for $k \leq n$ and make no steps thereafter ($\delta_k = 0$ for

$k > n$). Then the sum of cubes is $\xi = 1$ and the sum of squares is $n^{1/3}$ so that we achieve

$$EJ_{K_n} \leq \frac{A}{n^{1/3}}.$$

Thus we have convergence of the probability densities p_{K_n} to p in the relatively strong sense of $L_2(P)$ convergence of the gradient of the log-density, with an explicit bound on the error. The bound permits determination of a number of iterations n of the Metropolis algorithm, usually polynomial in the dimension d, such that we have approximately a sample from the density p in the sense of small J on the average for $k \leq n$.

The bound given here is a very promising bound, particularly because there is no assumption of log-concavity or unimodality of the target density p. Nevertheless, to be useful in an application there remains the task of showing that small J is sufficient for the task at hand. I also remind the reader of the requirement we have not verified here, that $\log p_k(x)$ as well as $\log p(x)$ have second derivatives and squares of first derivatives that remain bounded by κ for every $1 \leq k \leq n$. Assuming one has a target density p for which the derivatives are bounded in this way and that the initial distribution is chosen to be one for which these derivative bounds hold, then it is conceivable that the bounds continue to hold for $k \geq 1$.

4.2 Stochastic Gradient

Here we consider a stochastic gradient process designed to have transitions with the same conditional means and variance (to first order) as the Metropolis chain. Specifically, starting from an initial distribution for X_1, let X_n evolve according to the stochastic difference equation

$$X_{n+1} = X_n + \frac{1}{2}\delta_n^2 \nabla \log p(X_n) + \delta_n Z_n,$$

where the vector Z_n has mean zero, identity covariance and is independent of X_n, and ∇ denotes the gradient. Here I will take Z_n to be i.i.d. standard normal random vectors, though other choices such as uniformly distributed on a cube may also yield the same conclusions. In this chain we have lost the property of the density p providing an exact stationary distribution. Nevertheless, the aim is to reveal that for certain sequences δ_n the marginal distribution of the process will approach p at a reasonable rate.

Let p_n denote the density for X_n. Using Taylor expansions in the integral defining p_{n+1}, the expansion that should hold here under reasonable conditions on the densities is that

$$\log p_{n+1}/p_n = \frac{\delta_n^2}{2}\left[(\nabla \log p_n)^T \nabla(\log p_n/p) + \nabla^T \nabla \log p_n/p\right] - O(\delta_n^4),$$

where $\nabla^T \nabla \log p_n/p$ denotes the Laplacian or trace of the Hessian of $\log p_n/p$. I find this notation useful to facilitate the multivariate integration by parts

$\int p\nabla^T\nabla(\log p_n/p) = -\int(\nabla p)^T\nabla(\log p_n/p)$ which equals - $\int p(\nabla \log p)^T\nabla(\log p_n/p)$ and which combines nicely with the other term. Note that this expansion reveals the near stationarity of the density p to within order δ^4.

The corresponding drop that we would expect in the relative entropy, namely $D(p||p_n) - D(p||p_{n+1}) = E_{P_X} \log p_{n+1}(X)/p_n(X)$, can then be obtained by taking the expected value in the above identity and using the integration by parts. It also can be anticipated as an analogue of the differential identity for relative entropy in terms of relativized Fisher information. The result is a now familiar conclusion

$$D(p||p_n) - D(p||p_{n+1}) = \frac{1}{2}\delta_n^2 J(p||p_n) - 0(\delta_n^4).$$

Similarly

$$D(p_n||p) - D(p_{n+1}||p) = \frac{1}{2}\delta_n^2 J(p_n||p) - 0(\delta_n^4).$$

Here in this preliminary assessment I have not yet identified the most natural conditions on the densities for the validity of this expansion, nor have I identified the form of satisfactory constants in the $O(\delta_n^4)$ term. Nevertheless this direction is again promising for exhibiting bounds on distance from stationarity. Once again we would take δ_n to decrease to zero, but now we permit somewhat more gradual descent $(1/n)^r$ with $r > 1/4$ to make δ_n^4 summable and obtain in the same manner as before $E_{K_n} J(p||p_{K_n}) = O(1/n)^{(1-2r)}$. With δ_k held fixed at $1/n^{1/4}$ for $k \leq n$ and K_n uniformly distributed on $\{1, \ldots, n\}$ we would achieve

$$E_{K_n} J(p||p_{K_n}) = O(1/n)^{1/2}.$$

4.3 Reprise

Let's summarize this section on convergence bounds for Metropolis chains and stochastic gradients. We have seen that chain rules for relative entropy reveal the role of the relativized Fisher informations $J(p||p_n)$ and $J(p_n||p)$ for chains that make small steps. For discrete time Metropolis chains and stochastic gradients these informations converge to zero at polynomial rates in the number of steps, with apparently manageable constants. The implications of the convergence of J for the convergence of $D(p||p_n)$ and $D(p_n||p)$ or other measures of distance are less clear. Appealing to general Sobolev and Poincaré inequalities suggests that one would expect the sort of constants (related to conductance) that give rise in some cases to exponential numbers of steps with dimension in Chi-square bounds. It remains possible that improved Sobolev inequalities hold for log-density ratios. Integral relationships between D and J provide tactics for examining this issue. Additional work is desired to determine whether the relative entropy based analysis yields practical conclusions for Monte Carlo Markov chain methods in some practical multimodal and high-dimensional settings.

5 STATISTICS

Here we examine the role of relative entropy in characterizing achievable asymptotic performance of parametric and nonparametric estimators. We also look at practical bounds for risk of Bayes and related estimators. The first two subsections are of a classical flavor, but I include them since optimality properties of Bayes and maximum likelihood estimators and the nature of the optimal risk sequence as $(1/2)\#$ *parameters / sample size* seem often to be neglected in computational learning theory. The results characterizating optimal rates for uniformly accurate estimators in nonparametric settings should also be of general interest in computational learning theory.

5.1 Consistency of Maximum Likelihood

Let $P_{X|\theta}$, $\theta \in \Theta$ be a parametric family of distributions with densities $p(x|\theta)$ and let $X_1, X_2, \ldots X_n \ldots$ be i.i.d. with distribution P having a density $p(x)$ (with respect to a measure λ dominating the members of the family). It is required that Θ be contained in a separable metric space. In typical examples it is $k-$dimensional Euclidean space. Let $\hat{\theta}_n$ be a maximum likelihood estimate and let $P_{X|\hat{\theta}_n}$ be the corresponding estimate of the distribution. Relative entropy is used in analysis of maximum likelihood to determine the set of possible limits θ^*, to prove convergence of $\hat{\theta}_n$ to it and to prove convergence of the corresponding estimates of the distribution.

Maximizing the log-likelihood $\log p(X^n|\theta)$ is mathematically equivalent to minimizing the empirical relative entropy $\hat{D}_n(\theta) = (1/n)\log p(X^n)/p(X^n|\theta)$. Note that for each θ the empirical relative entropy $\hat{D}_n(\theta)$ converges almost surely to the relative entropy $D(\theta) = D(P||P_{X|\theta})$. If Θ were a finite set it would follow from the resulting uniform convergence that $D(\hat{\theta}_n) = D(P||P_{X|\hat{\theta}_n})$ would converge almost surely to $D^* = \min_\theta D(\theta)$ and hence that $\hat{\theta}_n$ would converge to the set of minimizers θ^* of $D(\theta)$, entailing, if in particular the minimizer θ^* is unique, that $\hat{\theta}_n \to \theta^*$ almost surely.

The consistency proof of Wald [Wal49] (and others who have extended it) treats the case that Θ is not finite by a compactification, assuming a domination of the log-densities, a continuity in θ and a convergence to zero of the $p(X^k|\theta)$ for any sequence of θ divergent from each compact subset of Θ. Though phrased in [Wal49] as separate conditions on domination in each sufficiently small ball in a compact subset and domination outside the compact set, the domination conditions are tantamount to the key assumption that for some $k \geq 1$, the function $p_{max}(X^k) = \max_\theta p(X^k|\theta)$ satisfies $E_P \log p(X^k)/p_{max}(X^k) > -\infty$. In particular if $p_{max}(X^k)$ is integrable then it is normalizable to be a probability density $q(X^k) = p_{max}(X^k)/c_k$ with $c_k = \int p_{max}(x^k)\lambda^k(dx^k)$ and hence $E_P \log p(X^k)/p_{max}(X^k) = D(P_{X^k}||Q_{X^k}) - \log c_k > -\infty$ and Wald's domination conditions are satisfied. (The quantity $\log c_k$ will arise again in universal

learning and coding).

The consequence of these assumptions is that $D(P||P_{X|\hat{\theta}_n})$ converges almost surely (and hence in probability) to $\min_\theta D(P||P_{X|\theta})$. If $D(P||P_{X|\theta})$ is strictly above this minimum outside neighborhoods of a minimizing θ^* then convergence of D implies parameter consistency $\hat{\theta}_n \to \theta^*$ a.s. as $n \to \infty$. If $P = P_{X|\theta^*}$ is in the model class then (even if there are multiple representations of it in the family) we have that almost surely,

$$D(P_{X|\theta^*}||P_{X|\hat{\theta}_n}) \to 0.$$

The convergence in probability of $D(P||P_{X|\hat{\theta}_n})$ to $D^* = \min_\theta D(P||P_{X|\theta})$ means that for every positive ϵ and δ there is an $N(\epsilon, \delta)$ such that for all $n \geq N(\epsilon, \delta)$

$$P_{X^n}\{D(P||P_{X|\hat{\theta}_n}) > D^* + \epsilon\} \leq \delta.$$

Such convergence in probability of the loss function of an estimator is called statistical consistency. For the maximum likelihood estimator and certain classes $\{P_{X|\theta} : \theta \in \Theta\}$ this convergence is not necessarily uniform over P. The uniformity of such convergence is called uniform consistency in statistics and called PAC learnability Hau93 in computational learning theory. Uniform consistency and the exhibition of polynomial bounds on $N(\epsilon, \delta)$ (which translate into an issue of minimax rate of convergence) are matters addressed in later subsections.

5.2 Efficiency of Parametric Estimators

In examination of the asymptotics of the distribution of estimators, relative entropy and related quantities arise again, not only in identifying, via the consistency analysis, the limit θ^*, which becomes the mean of the asymptotic distribution, but also in characterizing the asymptotic variance using the inverse of the Hessian of $D(P||P_{X|\theta})$ at θ^*, and even in determining the Gaussian shape of the asymptotic distribution via the entropy maximization principle in the central limit theory.

I do acknowledgement that, in the case that the true distribution is in the parametric family, characterization of the Fisher information matrix as a Hessian of twice the squared Hellinger distance is applicable to more general families than the entropy-based definition.

Suppose we have a smooth k-dimensional parametric family. Together with consistency, expansion of the gradient of the log-likelihood shows that under suitable conditions, maximum likelihood estimators achieve

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = J_{\theta^*}^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) \right) + O_P\left(\frac{1}{\sqrt{n}}\right)$$

Where J_{θ^*} is an information matrix and $g(X)$ is a score function that will be defined momentarily. The function g will have moments $E_P g(X) = 0$ and $I_{\theta^*} =$

$E_P g(X) g(X)^T$, so by application of the central limit theorem, the expansion yields convergence in distribution of the estimator to a normal. In particular,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \Rightarrow \text{Normal}(0, J_{\theta^*}^{-1} I_{\theta^*} J_{\theta^*}^{-1}).$$

The definition of $g(x)$ and J_{θ^*} for this expansion depends on which of two sorts of conditions one takes. Under local domination conditions as in Cramer Cra5?, $g(x) = \nabla \log p(x|\theta^*)$ is the gradient with respect to the parameter vector evaluated at θ^* and taken pointwise in x. Here J_{θ^*} is the expected Hessian $-E_P \nabla \nabla^T \log p(X|\theta^*)$ assumed to be positive definite. Under Cramer's conditions it is in agreement with the Hessian of the expectation $\nabla \nabla^T D(P||P_{X|\theta^*})$. In the special case that $P = P_{X|\theta^*}$ is in the family, these matrices reduce, under the local domination conditions, to the Fisher information $J_{\theta^*} = I_{\theta^*}$ and

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \Rightarrow \text{Normal}(0, I_{\theta^*}^{-1}).$$

Under a mean square differentiability condition as in [IH80, LeC86, ?, ?] one takes $g(x) = 2\zeta(x)/\sqrt{p(x|\theta^*)}$ where $\zeta(x) = \nabla \sqrt{p(x|\theta^*)}$ is the gradient in $L_2(\lambda)$ of $\sqrt{p(x|\theta^*)}$ at θ^*. The Fisher information $I_{\theta^*} = E_P g(X) g(X)^T$ may also be expressed as $\int \zeta(x) \zeta^T(x) 1_{\{p(x|\theta^*)>0\}} \lambda(dx)$, assuming here that $P = P_{X|\theta^*}$ is in the family. With this assumption one may set $J_{\theta^*} = (1/2)(I_{\theta^*} + \tilde{I}_{\theta^*})$ where $\tilde{I}_{\theta^*} = \int \zeta(x)(\zeta(x))^T \lambda(dx)$. With L_2 differentiability, $\tilde{I}_{\theta^*}$ agrees with the Hessian of twice the squared Hellinger distance at θ^*. The information matrices $\tilde{I}_{\theta^*}$, I_{θ^*} and J_{θ^*} agree provided the set $\{x : \nabla \sqrt{p(X|\theta^*)} \neq 0 \text{ and } \sqrt{p(X|\theta^*)} = 0\}$ is a set of measure zero (in accordance with LeCam's notion of contiguity). The conclusion in this setting is again that,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \Rightarrow \text{Normal}(0, I_{\theta^*}^{-1}).$$

Maximum likelihood is a type of M estimator [?], which means that it optimizes an empirical loss $(1/n) \sum_{i=1}^n G(X_i, \theta)$ for a choice of $G(X, \theta)$ for which $E_{P_{\theta^*}} G(X, \theta)$ is assured to be optimized at the unknown θ^*. Such estimation is also called the method of minimum contrast [?, ?, BM93] or the method of empirical risk minimization [Vap82].

M estimators other than maximum likelihood are asymptotically normal centered at θ^*, but with a larger covariance V [?, ?]. The difference $V - I_{\theta^*}^{-1}$ is nonnegative and equality $V = I_{\theta^*}^{-1}$ holds only if the empirical contrast is the log-likelihood, as can be shown as a Cramer-Rao inequality applied to the covariance of the limit distribution. In addition to maximum likelihood there are a number of other estimators, including Bayes estimators with suitable priors, for which $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is asymptotically normal$(0, I_{\theta^*}^{-1})$ (under similar local differentiability conditions, but often under weaker conditions for consistency) [?, LeC86] Historically then, for reasons that will be amplified in subsection 4.12 below, such estimators are said to be efficient, and estimators which have a larger variance of the asymptotic distribution are said to be inefficient.

5.3 Implications for Loss and Risk

Armed with the asymptotic distribution of $\hat{\theta}_n$ for efficient estimators one can determine the asymptotics for various smooth loss functions $L(\hat{\theta}, \theta)$. In particular, when $L(\hat{\theta}, \theta) = (1/2)(\hat{\theta} - \theta)^T I_\theta (\hat{\theta} - \theta)$ and more generally when $L(\hat{\theta}, \theta)$ is minimized at $\hat{\theta} = \theta$ and twice continuously differentiable in $\hat{\theta}$ with Hessian matching the Fisher Information I_θ, we have that

$$nL(\hat{\theta}_n, \theta^*) \Rightarrow \frac{1}{2}\chi^2(k),$$

where $\chi^2(k)$ is a random variable with the Chi-square (k) distribution (the distribution of $||Z||^2$ if Z is multivariate standard normal on R^k) which has expectation k. In particular, for smooth parametric families these asymptotics hold for $D(P_{X|\theta^*}||P_{X|\hat{\theta}_n})$ for twice $H^2(P_{X|\theta^*}, P_{X|\hat{\theta}_n})$ or for any other asymptotically equivalent measure of divergence.

When the sequence $nL(\hat{\theta}_n, \theta^*)$ is uniformly integrable then from this convergence in distribution we have that the risk $r_n = E_P L(\hat{\theta}_n, \theta^*)$ satisfies

$$r_n = \left(\frac{k}{2n}\right)(1 + o(1)).$$

That is, the risk is asymptotically the number of parameters divided by twice the sample size. Results in this direction for relative entropy loss are in [?, Cen82].

The convergence in distribution discussed here concerns demonstration that the tail probability of the loss $L(\hat{\theta}_n, \theta)$ satisfies

$$P\{L(\hat{\theta}_n, \theta) > \frac{\tau}{2n}\} \rightarrow P\{\chi^2(k) > \tau\}.$$

5.4 Uniformly Valid Bounds for Loss and Risk

The asymptotics above do not give uniformly valid bounds on the tail probability of the loss. Nevertheless, they point the way for what to expect in PAC learnability. The best one can hope for is uniformity in the asymptotics of loss for efficient estimators, that is, for all positive τ and n

$$P\{L(\hat{\theta}_n, \theta) > \frac{\tau}{2n}\} \leq CP\{\chi^2(k) > \tau\}$$

for some constant $C \geq 1$. So at best, for loss not greater than ϵ with confidence level $1 - \delta$ the sample size required is

$$N(\epsilon, \delta) = \frac{\tau_{k,\delta}}{2\epsilon}$$

with $\tau_{k,\delta}$ the upper δ quantile of the $\chi^2(k)$ distribution. For large k and small δ, this $\tau_{k,\delta}$ is approximately

$$k + (2k)^{1/2}(2\log 1/\delta)^{1/2}$$

to within order $k^{1/2} \log\log 1/\delta$. Typically one is forced to give up somewhat on such refined asymptotics to have bounds that hold uniformly for all n and all P in a family, in order to yield a sample size $N(\epsilon, \delta)$ with the desired guarantee. Nevertheless, ignoring for the moment the effect of δ, one should at least seek to have $N(\epsilon, \delta)$ of order k/ϵ corresponding to loss bounded by order k/n uniformly in probability. We now discuss general results of this type.

The cancellation of the Fisher information in the asymptotics of the divergence suggests that finite Fisher information and local quadratic approximation of the likelihood are not essential for order k/n bounds.

What is essential is that k be a metric dimension of the family. Results in this direction are in [LeC73, BM93] where it is assumed that if the family $\mathcal{Q} = \{P_{X|\theta} : \theta \in \Theta\}$ is compact and has finite metric dimension k with the Hellinger metric (which means that each ball of radius $r_2 > 0$ can be covered using at most $(Cr_2/r_1)^k$ balls of smaller radius $r_1 \leq (1/2)r_2$), then for various estimators (Bayes with certain priors, maximum likelihood on nets, global maximum likihood) there exists constants c_0, c_1, c_2, c_3 (each ≥ 1) such that for all n

$$P\{h^2(P, P_{X|\hat{\theta}_n}) > \frac{c_1 k + c_2 \tau}{n}\} \leq c_3 e^{-\tau}$$

uniformly over P in $\{P_{X|\theta} : \theta \in \Theta\}$, and, moreover, if $D^*(P) = \min_\theta D(P||P_{X|\theta})$ is added to the right side in the bound, then for all n

$$P\{h^2(P, P_{X|\hat{\theta}_n}) > c_0 D^*(P) + \frac{c_1 k + c_2 \tau}{n}\} \leq c_3 e^{-\tau}$$

uniformly over all probability distributions P. Integrating such bounds over $\tau \geq 0$ shows that the risk $Eh^2(P, P_{X|\hat{\theta}_n})$ is bounded by order

$$D^*(P) + k/n.$$

Results of this type are used in [BBM97, YB97a], when a list of models is available, to show under conditions on the models that there exist constants c_1', c_2', c_3' such that penalized maximum likelihood estimators $\hat{P}$ achieve

$$P\{h^2(P, \hat{P}) > c_1' a_n(P) + c_2' \tau/n\} \leq c_3' e^{-\tau}$$

uniformly over all probability measures. Here $a_n(P)$ is the accuracy index

$$a_n(P) = \min_m \{D_m^* + \frac{k_m}{n}\}$$

where $D_m^* = \min_{\mathrm{Q} \in m} D(P||\mathrm{Q})$ is the approximation error and k_m is the metric dimension of model m. Again integrating such a bound over $\tau > 0$ shows that the risk $EH^2(P, \hat{P})$ is bounded by order $a_n(P)$. Armed with such bounds it is shown in [BBM97] that one can achieve minimax optimal convergence rates in many nonparametric as well as parametric cases. Moreover, the penalized

likelihood estimators acheve these optimal rates adaptively, that is, without prior knowledge of which function classes contain the true density and without prior knowledge of which finite-dimensional model $m_n(P)$ will achieve the best accuracy index.

Note the shift in perspective here. When a variety of possible models are considered of various dimension, once a sample size is chosen, the aim is to have an adaptive procedure which will lead to as small as possible a risk, as if the best family m_n were known. Here $a_n(P)$ and its empirical counterparts, are unknown to us in advance of seeing the data, and for some P can converge to zero arbitrarily slowly. Thus it is not possible to chose a sample size with performance guarantees (except of course by hypothetically presuming certain constraints on $a_n(P)$). Again, once a sample size is chosen, one should nolonger feel tied to these presumptions, but rather should allow the data to reveal better or worse accuracy than hypothesized.

5.5 Relative Entropy Risk for Maximum Likelihood in Exponential Families

Let $S = S_m$ be a linear space of dimension $k = k_m$ of real-valued functions on X and let $\phi_1(x), \ldots \phi_k(x)$ be a convenient basis for S which will be used in a model m for the log-density. Densities in the exponential family $p(x|\theta) = p_0(x) \exp\{\sum_{j=1}^k \theta_j \phi_j - \psi(\theta)\}$ where $\psi(\theta) = \log \int p_0 \exp\{\sum \theta_j \phi_j\}$ with $\int \phi_j(x) p(x|\theta)$ equal to α_j, say, for $j = 1, \ldots, n$ arises naturally in information-theoretic statistics as the information projection achieving the minimum $D(q||p_0)$ in the linear constraint set $\mathcal{Q} = \{q : \int \phi_j q = \alpha_j \text{ for } j = 1, \ldots, k\}$ in accordance with Kullback's minimum discrimination information principle [Kul59, Csi75].

Suppose we have empirically specified values $\alpha_j = (1/n) \sum_{i=1}^n \phi_j(X_i)$ for $j = 1, \ldots, k$. If there is a density $\hat{p}_{n,m}(x) = p(x|\hat{\theta})$ in the exponential family that matches these expected values then, as well as being the density in $\mathcal{Q}$ that minimizes the relative entropy from p_0, it is also the maximum likelihood estimator (and method of ϕ moments estimator) in the exponential family. I remark that given S_m the basis may be adjusted in any convenient way. As long as it spans the same space, it will yield the same maximum likelihood density estimator $\hat{p}_{n,m}(x)$.

Probability bounds on the relative entropy loss of this maximum likelihood estimator in exponential families are given in [BS91].

In particular suppose the data $X_1, \ldots, X_n$ are i.i.d. according to a density p that is not necessarily in the family. If there is a member of the exponential family $p_m^*(x) = p(x|\theta^*)$ that achieves expectations for $\phi_j(X), j = 1, \ldots, n$ that match the values achieved by p, then it is the member of the family that best approximates p in the sense of minimizing the relative entropy $D(p||p(\cdot|\theta))$, as can be confirmed by the Pythagorean-like identity

$$D(p||p(\cdot|\theta)) = D(p||p_m^*) + D(p_m^*||p(\cdot|\theta)).$$

Consequently, the relative entropy loss of an estimator $\hat{p}_{n,m}$ decomposes as a sum of an approximation error and an estimation error

$$D(p||\hat{p}_{n,m}) = D(p||p^*_m) + D(p^*_m||\hat{p}_{n,m}).$$

The approximation error $D(p||p^*_m)$ is related to the classic L_2 approximation error of the log-density $f(x) = \log p(x)$ by members $f_m(x)$ of S_m, when the corresponding L^∞ approximation error is bounded.

Using for convenience in the analysis, a basis in which the $\phi_1, \ldots, \phi_k$ are orthonormal with respect to the unknown p, the estimation error $D(p^*_m||\hat{p}_{n,m})$ is shown to be related to the sum of squared error in the estimation of the expectations $\int \phi_j p$ by the sample averages $(1/n)\sum_{i=1}^n \phi_j(X_i)$ which is of order k_m/n in probability.

Indeed for families in which there is a suitable link between the L_2 and L^∞ norms in S_m (satisfied for instance by certain polynomial, spline, trigonometric, and wavelet models), it is shown in [BS91] that for a range of τ specified there

$$P\{D(p||\hat{p}_{n,m}) \geq D(p||p^*_m) + C\frac{k_m}{n}\tau\} \leq 1/\tau$$

uniformly over all p for which the L^∞ approximation error of the log density $||f - f_m||_\infty$ is bounded by some γ, where the constant C depends on γ.

Armed with classical approximation results for functions in R^d with norms on derivatives of order s, the approximation error is shown to be of order $D(p||p^*_m) = (1/m)^{2s}$ using a parameter dimension k_m of order m^d. Here the model index m represents the dimension per coordinate in a tensor product basis. It follows that $D(p||\hat{p}_{n,m})$ converges to zero in probability at the rate

$$\min_m \left(\frac{1}{m^{2s}} + \frac{m^d}{n}\right),$$

known to be minimax optimal in such function classes [YB97b].

The results given in [BS91] can be improved somewhat using the better constants in the relationship between $D(p||q)$ and $\int p(\log p/q)^2$ stated in section 1.3. Also it is possible to obtain exponential bounds on the probability tails (analogous to those stated in the preceding subsection) rather than the $1/\tau$ bound, see [YB97b].

An advantage of this treatment of exponential families is that it concerns relative entropy loss bounds rather than the weaker Hellinger loss bounds that are available for more general families.

5.6 Asymptotics of Bayes Posterior Distributions

Suppose $\{P_{X^n|\theta}, \theta \in \Theta\}$ is a family of distributions for possible data sequences X^n and suppose further that the family is dominated by a measure λ^n yielding

joint density functions $p(X^n|\theta)$. If the parameter θ is endowed with a prior probability distribution $P_\theta = W$, then Bayes rule provides the posterior distribution $P_{\theta|X^n} = W_{\theta|X^N}$ for θ given observation of X^n as

$$P_{\theta|X^n}(B) = \int_B p(X^n|\theta)W(d\theta)/p_W(X^n)$$

for $B \subset \Theta$ where

$$p_W(X^n) = \int_\Theta p(X^n|\theta)W(d\theta)$$

is the (joint) marginal density for X^n obtained by integrating out the parameter with respect to the prior.

In the Bayes model the quantity θ which is unknown is modeled probabilistically. In advance of collecting the data, at the time we are making a choice of a prior, we can assess the behavior of Bayes procedures also probabilistically by asking what will be the distributional behavior of the posterior and of Bayes estimators for various possible values of the unknown parameter. From such an investigation we can seek to know what properties of the prior determine, at least in rough, the characteristics the posterior.

Information theory plays a role here in several ways. Suppose θ^* is the unknown value of the parameter. Then with probability one $p(X^n|\theta^*)$ will be positive. Dividing it into the numerator and denominator and writing the density ratios as exponentials of their logarithms shows that the posterior probability of sets equals

$$P_{\theta|X^n}(B) = \int_B e^{-n\hat{D}(\theta^*||\theta)}W(d\theta)/e^{-n\hat{R}_n},$$

where $\hat{D}(\theta^*||\theta) = (1/n)\log(p(X^n|\theta^*)/p(X^n|\theta))$ is the empirical divergence of the distribution at θ from the distribution at θ^* and $\hat{R}_n = (1/n)\log(p(X^n|\theta^*)/p_W(X^n))$ is the empirical divergence of the Bayes mixture from the distribution at θ^*. The idea here is that $\hat{R}_n$ will be converging to zero at a certain rate as we shall see in later sections. One anticipates that the integral over B should converge to zero exponentially fast as long as it does not include a relative entropy neighborhood of θ^*, though care must be taken to deal with the potential lack of uniformity of convergence of $\hat{D}(\theta^*||\theta)$ to $D(\theta^*||\theta)$.

Examination of this representation of the posterior suggests the following principles: (1) asymptotic concentration of the posterior in relative entropy neighborhoods of θ^* for those values of θ^* for which the prior probability of such neighborhoods is positive, (2) rates of concentration of the posterior and rates of convergence of estimators determined by the tradeoff between a radius of such relative entropy neighborhoods and their prior probability, and (3) asymptotic normality of the posterior in smooth finite dimensional cases with positive prior density at θ^* via local quadratic expansion of $\hat{D}(\theta^*||\theta)$. Such conclusions are true under appropriate conditions [LeC53, ?, ?, Bar87, BSW97], with care taken

to deal with the potential lack of uniformity in θ of convergence of the empirical relative entropy $\hat{D}(\theta^*||\theta)$ for large n.

As pioneered by Le Cam and Schwartz [Sch65], a special role is played by the theory of uniformly consistent tests to reveal consistency properties of Bayes estimators in cases where that convergence need not be uniform in θ. By that means, consistency holds at θ^* for priors that give positive mass to relative entropy neighborhoods, in those cases when restricted to the parametric model, relative entropy convergence is equivalent to weak convergence of distributions, see [Bar87, CB90]. One may allow also priors in some nonparametric settings and get Hellinger or L^1 consistency provided that the prior probability is exponentially small for the set of those parameter values θ for which the models lack a certain amount of smoothness [Bar87, BSW97].

5.7 The Form of Bayes Estimation & Prediction with Relative Entropy Loss

Given a loss function for the estimation of a quantity that depends on a parameter, the task of choosing an estimator to minimize the average risk, averaging both with respect to the distribution of data $P_{X^n|\theta}$ and with respect to a prior P_θ, is accomplished by choosing for each X^n an estimate that minimizes the posterior risk. In particular, suppose we wish to estimate the distribution of an unobserved random variable U that also depends on the parameter θ. First let's suppose U is conditionally independent of X^n given θ. Assume that the family of distributions for U have densities $p(u|\theta)$ with respect to some measure. We use relative entropy loss and ask, "what is the form of the Bayes estimator?"

Let $p(u|X^n) = \int p(u|\theta) P_{\theta|X^n}(d\theta)$ be the predictive density in which we have integrated out θ with respect to the posterior distribution and let $q(u|X^n)$ be any other estimator which is nonnegative, integrates to 1, and depends on X^n. Then the estimators satisfy the chain rule

$$\int D(P_{U|\theta}||Q_{U|X^n}) P_{\theta|X^n}(d\theta) = \int D(P_{U|\theta}||P_{U|X^n}) P_{\theta|X^n}(d\theta) + D(P_{U|X^n}||Q_{U|X^n}),$$

which is uniquely minimized by setting $Q_{U|X^n} = P_{U|X^n}$, the distribution estimator corresponding to the density estimator $\hat{p}(u) = p(u|X^n)$.

So we conclude that the predictive density $p(u|X^n)$, which is already customarily used in the Bayesian's probabilistic model, is indeed the Bayes estimator with relative entropy loss [Ait75, CB90].

It is interesting to note that many other loss functions, e.g. Hellinger or L_1, lead to different estimators of the distribution that would not be the Bayesian's conditional distribution for U given X^n.

In the case that the distribution of U is not conditionally independent of the data X^n given the parameter θ, the above conclusions still hold except that $p(u|\theta)$ is replaced by $p(u|\theta, X^n)$. In that case the quantity being estimated happens to also be a function of the observed data as well as the unknown parameter.

Though such a problem deviates from traditional statistical decision theory, it remains very much relevant for prediction problems, and it is not a obstacle for these Bayesian considerations, since they are carried out conditionally given the data. We find that the Bayes estimator is $p(u|X^n) = \int p(u|\theta, X^n) P_{\theta|X^n}(d\theta)$. For prediction we principly have in mind the case that $U = X_{n+1}$ is the next observation after the data $X_1, \ldots, X_n$.

5.8 Decision Theory and Bayes Estimation

Decision-theoretic characterization of the importance of Bayes estimators arises via admissibility and via minimaxity, as well as through optimization of average risk, and some of these characterizations will arise in our information-theoretic analysis.

An estimator $\hat{p}$ of the density function $p(\cdot|\theta)$ is inadmissible if there is another estimator $\hat{p}^{better}$ such that the risk function $E_{P_{X^n|\theta}} L(p(\cdot|\theta), \hat{p}^{better})$ does not exceed and is somewhere less than $E_{P_{X^n|\theta}} L(p(\cdot|\theta), \hat{p}^{better})$ for $\theta \in \Theta$. Admissible estimators are defined as those that are not inadmissible. Bayes estimators are admissible and every admissible estimator has a risk function which agrees with the limit of risks of Bayes estimators for a sequence of priors, see for instance [Fer67].

For the estimation of a parameterized quantity μ_θ using data X^n the minimax risk using a loss function $L(\mu, \mu')$ is defined as

$$r_n = \min_{\hat{\mu}} \max_{\theta \in \Theta} E L(\mu_\theta, \hat{\mu}),$$

where the minimization is over all estimators $\hat{\mu}$ that depend on the data, the maximization is over all θ in the parameter space, and the expectation is with respect to $P_{X^n|\theta}$. For any prior W the Bayes average risk of the Bayes estimator $\hat{\mu}_W$ is

$$\int E_{P_{X^n|\theta}} L(\mu_\theta, \hat{\mu}_W) W(d\theta)$$

which is often called simply the Bayes risk. It provides a lower bound on the minimax risk. Maximizing over choices of priors W on θ leads to what is called the least favorable prior and the maximin risk

$$\underline{r}_n = \max_W \min_{\hat{\mu}} \int E_{P_{X^n|\theta}} L(\mu, \hat{\mu}) W(d\theta).$$

Under convexity and semi-continuity conditions on the loss, a fundamental theorem of games in decision theory shows that the maximin risk is the same as the minimax risk and that there is a unique minimax procedure that is Bayes with respect to a least favorable prior (though on occassion there may be several priors which yield this unique maximin procedure) [Fer67]. In particular the theory of equality of minimax and maximin risks applies to total relative entropy as discussed in subsection 1.7.

5.9 Prediction and Cumulative Relative Entropy Risk

The simplest context in which to develop risk bounds using information theory is the setting of prediction with cumulative relative entropy risk. For each $n \geq 1$, after observing $X_1, \ldots, X_{n-1}$, we are asked to provide an estimate $\hat{p}_n(x) = \hat{p}_n(x; X_1, \ldots X_{n-1})$ for the density function for X_n (this includes for $n = 1$ a guess $\hat{p}_1(x)$ based on no data for the density function for X_1). We consider first the case that $X_1, X_2, \ldots, X_N$ are i.i.d. with distribution P having density p. We require that the estimate $\hat{p}_n$ be a valid probability density. For each $n = 1, \ldots, N$ we incur a loss $D(p||\hat{p}_n)$. These losses accumulate to the total loss

$$\sum_{k=1}^{N} D(p||\hat{p}_n).$$

The corresponding total risk is the expected value

$$R_N(P) = \sum_{n=1}^{N} E_{P_{X^{n-1}}} D(p||\hat{p}_n).$$

Of course one could use other measures of loss in this sum. Some such choices of loss are bounded by investigating relative entropy [HB92]. An advantage here of relative entropy is the chain rule that allows considerable simplification of the analysis of the cumulative risk.

Let Q_{X^N} be the joint distribution with conditional densities $q(X_n|X_{n-1}, \ldots X_1)$ for X_n defined by the estimator $\hat{p}(x; X_1, \ldots X_{n-1})$. The chain rule gives total risk

$$R_N(P) = \sum_{n=1}^{N} E_{P_{X^{n-1}}} D(p||\hat{p}_n) = \sum_{n=1}^{N} E \log \frac{p(X_n)}{q(X_n|X_{n-1}, \ldots X_1)} = D(P_{X^N}||Q_{X^N})$$

and Cesaro average risk

$$\bar{r}_N(P) = \frac{1}{N} \sum_{n=1}^{N} E_{P_{X^{n-1}}} D(p||\hat{p}_n) = \frac{1}{N} D(P_{X^N}||Q_{X^N}).$$

As we shall see in the next subsection it is easier to directly bound the total relative entropy $D(P_{X^N}||Q_{X^N})$ than it is to bound the individual risks $ED(p||\hat{p}_n)$ for $n = 1, \ldots, N$, especially when we use Bayes estimators. Before turning our attention to such bounds on total risk, I point out a way to get an estimator with reasonable small individual risk once a bound on the total or Cesaro average risk is available.

An estimator with risk that is not greater than $\bar{r}_N(P)$ is obtained by smoothing over sample sizes

$$\tilde{p}_N(x) = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_n(x) = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_n(x; X_1, \ldots, X_{n-1}).$$

By convexity of D we have

$$ED(p||\tilde{p}_N) \leq \frac{1}{N}\sum_{n=1}^{N} E_{P_{X^{n-1}}} D(p||\hat{p}_n) = \frac{1}{N} D(P_{X^N}||Q_{X^N}).$$

Under the i.i.d. model we will have the same risk $D(p||\tilde{p}_{\pi,N})$ with $\tilde{p}_{\pi,N} = (1/N)\sum_{n=1}^{N} \hat{p}_n(x; X_{\pi(1)}, \ldots, X_{\pi(n-1)})$ for any permutation π of $\{1, \ldots, N\}$. If one is going to take such a Cesaro average, one might find it most palatable to average over subsamples going backward (rather than forward) in the sequence so that the most recent observations contribute the most using the estimator $(1/N)\sum_{n=1}^{N} \hat{p}_n(x; X_{N-1}, \ldots, X_{N-(n-1)})$. A variant on this approach is to average over various permutations to further reduce the relative entropy risk.

If in a specific setting one is able to characterize satisfactorily the final individual risk for the original estimator $ED(p||\hat{p}_N)$, then it is not necessary to potentially weaken the quality of the estimator by averaging across sample sizes less than N.

Next suppose that some dependence in $X_1, \ldots, X_N$ is permitted. For definiteness, think of the case of a Markov chain. In the case of a stationary transition, one could envision estimation (using parametric models) of a transition density function $p(x'|x)$ for all x, x' in the state space from the batch data $X_1, \ldots, X_N$. However, for on-line prediction purposes, at each time n, having seen the previous values $X_1, \ldots X_{n-1}$, what is strictly required is to be able to estimate (by some $q_n(x'|X_1, \ldots, X_{n-1})$) the conditional density function $p_n(x'|X_{n-1})$ for the unknown X_n conditioning on the observed X_{n-1} (rather than conditioning on some other arbitrary x). Fortunately the corresponding relative entropy risks of prediction are precisely what is accumulated in the chain rule

$$R_N(P) = \sum_{n=1}^{N} E \log \frac{p_n(X_n|X_{n-1})}{q_n(X_n|X_{n-1}, \ldots, X_1)} = D(P_{X^N}||Q_{X^N}).$$

Dividing by N we have the relative entropy rate $\bar{r}_N(P) = (1/N)D(P_{X^N}||Q_{X^N})$ which equals the average for $n = 1, \ldots, N$ of the conditional relative entropies of prediction $E \log p_n(X_n|X_{n-1})/q_n(X_n|X_{n-1}, \ldots, X_1)$.

5.10 Cumulative Risk and Resolvability of Bayesian Predictions

Continuing the initial framework of the preceding subsection in which the random variables are i.i.d., suppose we have a family of densities $p(x|\theta), \theta \in \Theta$. Here we use Bayes estimators with a prior probability distribution $W = W_N$. This prior is not permitted to change with n, but it may depend on the total horizon N of all the variables which will be observed in the study. The estimators are

the Bayes predictive densities $\hat{p}_n(x) = p^{(W)}(x|X^{n-1}) = \int p(x|\theta)W_{\theta|X^{n-1}}(d\theta)$ and the chain rule gives total risk

$$R_N(P) = \sum_{n=1}^{N} E_{P_{X^{n-1}}} D(p||\hat{p}_n) = D(P_{X^N}||P^{(W)}_{X^N}),$$

and Cesaro average risk

$$\bar{r}_N(P) = \frac{1}{N}\sum_{n=1}^{N} ED(p||\hat{p}_n) = \frac{1}{N} D(P_{X^N}||P^{(W)}_{X^N}).$$

As we shall see from a simple bound this risk is made small for any P in the information support of the prior.

The information closure of the family $\{P_{X|\theta}\}$ consists of those distributions P for which the information neighborhoods $B_{\delta,P} = \{\theta : D(P||P_{X|\theta}) \leq (1/2)\delta^2\}$ are non-empty for all $\delta > 0$ and the information support of the prior consists of those P for which the information neighborhoods are assigned positive prior probability $W(B_{\delta,P}) > 0$.

The size of the risk depends on now much prior probability is given to these information balls. Indeed, the following bound holds,

$$\bar{r}_N \leq \min_{\delta>0}\{\frac{\delta^2}{2} + \frac{1}{N}\log 1/W(B_{\delta,P})\}.$$

The right side of this inequality is here called the index of resolvability of the distribution P by the mixture of distributions with prior W. An alternative expression for the index of resolvability is

$$\min_{B}\{\max_{\theta\in B} D(P||P_{X|\theta}) + \frac{1}{N}\log 1/W(B)\}.$$

This definition is an extension of the resolvability definition given in CB91 in which W was discrete and the minimization was restricted to singleton sets $\{\theta\}$ (there $L(\theta) = \log 1/W\{\theta\}$ was interpreted as an arbitrary codelength for the parameter in a two-stage rather than mixture code for the resolution of X, see section 5).

The proof of the resolvability bound on cumulative risk follows simply from $p_W(X^N) \geq W(B)\int_B p(X^N|\theta)W(d\theta|B_\delta)$ and convexity to get

$$\begin{aligned}
D(P_{X^N}||P^{(W)}_{X^N}) &= E_P \log \frac{p(X^n)}{p_W(X^N)} \\
&\leq E_P \log \frac{p(X^n)}{\int_B p(X^N|\theta)W(d\theta|B)} + \log\frac{1}{W(B)} \\
&\leq \int_B D(P_{X^N}||P_{X^N}|\theta)W(d\theta|B) + \log\frac{1}{W(B)} \\
&\leq \max_{\theta\in B} D(P_{X^N}||P_{X^N|\theta}) + \log\frac{1}{W(B).}
\end{aligned}$$

Dividing by N and optimizing over B produces the claimed bound. This proof is from [Bar86] where the point was Cesaro average consistency for all P in the information support of the prior and its use to express rates of convergence is in [Bar87]. The name resolvability for the bound is more recent.

Thus the cumulative accuracy of Bayes estimation depends only on the local behavior of the prior for sets of θ with $P_{X|\theta}$ near the distribution P_X followed by the data. This simple conclusion is to be contrasted with the behavior of the posterior distribution which to asymptotically concentrate on a neighborhood of P_X requires also global conditions (entailing existence of a uniformly consistent test against all but an a priori neglibible set of $P_{X|\theta}$ outside of the neighborhood of P_X, [Bar87]).

Allowing dependence in the model, the same bound holds for $\bar{r}_N(P) = (1/N)D(P_{X^N}||P_{X^N}^{(W)})$ (the average conditional relative entropy of prediction) when the information neighborhoods are defined more generally by

$$B_\delta(P_{X^N}) = \{\theta : (1/N)D(P_{X^N}||P_{X^N|\theta}) \leq (1/2)\delta^2\}.$$

An alternative information-theoretic development of a similar bound on $D(P_{X^N}||P_{X^N}^{(W)})$ is obtained via chain rule expansion of the total divergence between the joint distributions $P_{X^N} \times W_\theta^{(N)}$ and $P_{X^N|\theta}W_\theta$ where the approximate posterior $W_\theta^{(N)}$ is defined to have density $e^{-D(P_{X^N}||P_{X^N|\theta})}/C_N$ with respect to the prior W_θ. The result is that

$$D(P_{X^N}||P_{X^N}^{(W)}) + E_{P_{X^N}} D(W_\theta^{(N)}||W_{\theta|X^N}) = \log 1/C_N$$

where $C_N = \int e^{-D(P_{X^N}||P_{X^N|\theta})}W(d\theta)$. Thus as shown in [Bar87, HO95a] the total relative entropy risk $D(P_{X^N}||P_{X^N}^{(W)})$ is bounded by the following quantity (interpreted as a "Razor" in Bal97)

$$\log 1/\int e^{-D(P_{X^N}||P_{X^N|\theta})}W(d\theta).$$

Indeed this proof shows that the Razor equals the total relative entropy risk up to a term that gives the error in an approximation to the posterior. Restricting the integral to a neighborhood, it is seen that bound improves somewhat on the previous bound $\min_\delta\{(N/2)\delta^2 + \log 1/W(B_{\delta,P})\}$.

5.11 Parametric Bounds

Suppose we have a finite-dimensional parametric family of densities $p(x|\theta)$, $\theta \in \Theta \subset R^k$ and that $X_1, \ldots, X_N$ are i.i.d. according to $p(x|\theta^*)$ where θ^* is in the interior of Θ and suppose the relative entropy $D(\theta^*||\theta) = D(P_{X|\theta^*}||P_{X|\theta})$ is twice continuously differentiable in θ at θ^* with positive definite Hessian J_{θ^*}.

Let $\bar{J}_{\theta^*}$ locally dominate the Hessian for θ with $D(P_{X|\theta^*}||P_{X|\theta}) \leq (1/2)\delta^2$, so that for such θ,

$$D(P_{X|\theta^*}||P_{X|\theta}) \leq \frac{1}{2}(\theta - \theta^*)^T \bar{J}_{\theta^*}(\theta - \theta^*).$$

Then the information ball $B_{\delta,\theta^*} = \{\theta : D(P_{X|\theta^*}||P_{X|\theta}) \leq (1/2)\delta^2\}$ contains the ellipse

$$S_{\delta,\theta^*} = \{\theta : (\theta - \theta^*)^T \bar{J}_{\theta^*}(\theta - \theta^*) \leq \delta^2\}.$$

Suppose also that the prior W satisfies a near-absolute continuity property near θ^*, namely that there exists a positive $\underline{w}_{\theta^*}$ such that the prior probability of the ellipse S_{δ,θ^*} is at least $\underline{w}_{\theta^*}$ times its volume (as in the case of a prior with a density $w(\theta)$ locally bounded below by $\underline{w}_{\theta^*}$).

Now the prior probability of the information neighborhood satisfies

$$W(B_{\delta,\theta^*}) \geq W(S_{\delta,\theta^*}) \geq \underline{w}_{\theta^*}|\bar{J}_{\theta^*}|^{-1/2} v_k \delta^k$$

where v_k denotes the volume of the unit ball in R^k. Consequently, we have a bound for $(N/2)\delta^2 + \log 1/W(B_{\delta,\theta^*})$ which is obtimized at $\delta^2 = k/N$, yielding

$$D(P_{X^N|\theta^*}||P_{X^N}^{(W)}) \leq \frac{k}{2}\log\frac{N}{k} + \frac{k}{2} + \log\left(|\bar{J}_{\theta^*}|^{1/2}/\underline{w}_{\theta^*}\right) + \log 1/v_k$$

Clearly, similar bounds for dependent data models are possible as long as there is a local quadratic behavior for $(1/N)D(P_{X^N|\theta^*}||P_{X^N|\theta})$.

Under additional regularity assumptions on the parametric family in the i.i.d. case it is shown in [CB90] that

$$D(P_{X^N|\theta^*}||P_{X^N}^{(W)}) = \frac{k}{2}\log\frac{N}{2\pi e} + \log\left(|J_{\theta^*}|^{1/2}/w(\theta^*)\right) + o(1)$$

for θ^* in the interior of the parameter space when the prior has a density $w(\theta)$ that is positive and continuous at θ^*. These asymptotics are shown there (in [CB90] to be related to the asymptotic normality of the posterior distribution and to Laplace's method of approximation of the Bayes mixture.

With the choice of Jeffreys' prior w_J which is proportional to $|J_\theta|^{1/2}$ we see that the total relative entropy is asymptotically $(k/2)\log N$ plus a fixed constant in the interior of the parameter space. Thus all such $P_{X^N|\theta}$ are approximately equidistant from the centroid $P_{X^N}^{(w_J)}$ obtained by averaging with respect to w_J.

Indeed, it is shown in [CB94] that for each compact S internal to Θ the minimax total relative entropy satisfies

$$V_N = \max_{Q_X^N} \min_{\theta \in S} D(P_{X^N|\theta}||Q_{X^N}) = \frac{k}{2}\log\frac{N}{2\pi e} + \log C_{J,S} + o(1)$$

where $C_{J,S} = \int_S |J_\theta|^{1/2} d\theta$. Moreover, Jeffreys' prior on S is asymptotically least favorable (maximin) and sequences of modifications of it are asymptotically

maximin and asymptotically minimax. These modifications use Jeffreys' prior on a sequence of sets S_n shrinking slowly to S, to handle the boundary behavior so that the prior remains positive in neighborhoods of all points in S.

In [XB97a] the minimax value for the family of discrete distributions on the whole probability simplex is determined. There too, modifications of Jeffrey's prior are used.

The answer $(k/2)\log N$ plus a constant for the cumulative relative entropy risk $D(P_{X^N|\theta *}||P^{(W)}_{X^N}) = \sum_{n=1}^{N} ED(p^*||\hat{p}_n)$ corresponds to individual risks $ED(p^*||\hat{p}_n)$ of the form $k/2n$ plus a summable remainder, in keeping with the efficient level identified in section 4.2.

The individual risks $ED(p^*||\hat{p}_n)$ may be much smaller or larger than $k/2n$ for some $n \leq N$, but not for most such n. Indeed, the Cesaro average of the risks is $(k/2N)(\log N + \text{constant})$ in agreement with the Cesaro average of k/n.

5.12 Negligibility of Superefficiency

The key to efficient total relative entropy risk is the identification of one joint probability measure Q_{X^N} (e.g. a Bayes mixture) which is simultaneously reasonably close to all the members $P_{X^N|\theta}$ of the parametric family. Indeed in the parametric case we have seen that relative entropy rate $(1/N)D(P_{X^N|\theta}||Q_{X^N})$ can be made to be of order $(k/2N)\log N$ for every θ. We want small distance simultaneously from distributions $P_{X^N|\theta}$ that have large distance between themselves. Thus it is essential to the ability to obtain estimators of small cumulative risk that the relative entropy distance does not satisfy a triangle inequality and hence is not a metric. Nevertheless, the intuition of distance holds up to some extent. As we shall review, there is a critical relative entropy distance, such that it is not possible to make Q_{X^N} simultaneously closer than this critical distance to the members of the family $P_{X^N|\theta}$ except for a negligible set of parameter values. This was shown by Rissanen [Ris84, Ris86] with refinements in [?, ?, BH95].

The identification of the critical distance depends on the ability to estimate uniformly well the parameter θ at a certain rate. As we have seen in typical parametric problems this rate is of order $1/\sqrt{N}$.

For simplicity, suppose Θ is a compact set in R^k and that there are estimators $\hat{\theta}_n$ such the the event $A_{N,\theta} = \{X^N : ||\hat{\theta}_N - \theta|| \leq C_N/\sqrt{N}\}$ has probability $P_{X^N|\theta}(A_{N,\theta})$ at least $1-\delta$ uniformly in Θ, where C_N is a slowly growing sequence (e.g. a logarithm). From the definition of these events, they are disjoint when θ and θ' separated by at least $\epsilon = 2C_N/\sqrt{N}$. If G_ϵ is a set of such $\epsilon-$separated θ's, it is not possible to have the $Q_{X^N}(A_{N,\theta})$ probability be much larger than $1/\text{card}(G_\epsilon)$ except for a small fraction of such θ's, since the sum of $Q_{X^N}(A_{N,\theta})$ over the disjoint sets is not more than 1. Implications for the total relative entropy follow, using the fact that $D(P_{X^N|\theta}||Q_{X^N})$ is not smaller than the

divergence restricted to the event $A_{N,\theta}$ and its complement, to get

$$\begin{aligned} D(P_{X^N|\theta}||Q_{X^N}) &\geq P_{X^N|\theta}(A_{N,\theta}) \log 1/Q_{X^N}(A_{N,\theta}) - \log 2 \\ &\geq (1-\delta) \log 1/Q_{X^N}(A_{N,\theta}) - \log 2. \end{aligned}$$

Taking G_ϵ to be a largest ϵ separated set, it has log cardinality equal to the metric entropy [], known to be of order $k \log 1/\epsilon$. Consequently, for any Q_{X^N},

$$D(P_{X^N|\theta}||Q_{X^N}) \geq (1-\delta)\frac{k}{2} \log N - \text{const}$$

except for θ in a set that possesses a sparse cover and consequently has small Lebesgue measure. Thus $(k/2) \log N$ is the efficient level of cumulative relative entropy risk.

Indeed, in [Ris86, BH95] it is shown that the set of superefficient cumulative risk

$$\{\theta : \limsup_N \frac{D(P_{X^N|\theta}||Q_{X^N})}{\log N} < \frac{k}{2}\}$$

has Lebesgue measure zero. A corollary using the chain rule is that for any estimator sequence $\{\hat{P}_n\}$ the set of superefficient individual risk

$$\{\theta : \limsup_n nED(P_{X|\theta}||\hat{P}_n) < \frac{k}{2}\}$$

also has Lebesgue measure zero [BH95]. This confirms that $k/(2n)$ is the efficient level of risk for parametric estimation with relative entropy loss.

In particular this conclusion holds for plug-in estimators $\hat{P}_n = P_{X|\hat{\theta}_n}$. The conclusion in this case is in agreement with $k/(2n)$ as the efficient risk for loss functions locally equivalent to

$$(1/2)(\theta - \hat{\theta})^T J_\theta (\theta - \hat{\theta}).$$

Related negligibility of superefficiency results are in [?] using Bayesian Cramer-Rao inequalities.

The precursor of such efforts is the result of LeCam [LeC53] (based on Fatou's Lemma applied to the Bayes average of the difference in risk of a given estiamtion and the risk of a Baye estimator for certain absolutely continous priors) that for any loss function that can be expressed as a bounded function of $\sqrt{n}(\theta - \hat{\theta})$ and any efficient sequence of estimators, the set of superefficiency has measure zero. The present information-theoretic analysis does not presume the estimator to be efficient to obtain neglibility of superefficiency, it is for the potentially unbounded sequence $nED(P_{X|\theta}||P_{X|\hat{\theta}})$, and it can be used to quantify bounds on the measure of superefficiency for finite n.

5.13 Minimax Bounds and Metric Entropy

Let $\mathcal{F}$ be a class of functions, let $d(f,g)$ be a metric and let $K(\epsilon)$ be the log-cardinality of the largest ϵ-separated set (ϵ-net) of functions in $\mathcal{F}$, called the Kolmogorov ϵ-entropy, metric entropy, or ϵ-capacity of the family $\mathcal{F}$. An optimal set of functions packed at separation ϵ is also an ϵ-cover, every function in $\mathcal{F}$ is within ϵ of a function in the net.

Let $\mathcal{F}$ be large enough to satisfy $\lim_{\epsilon \to 0} K(\epsilon/2)/K(\epsilon) > 1$ which is true for instance if $K(\epsilon) \asymp (1/\epsilon)^c$, as is typical in nonparametric (infinite-dimensional) function classes, in contrast to the $k \log 1/\epsilon$ behavior in finite-dimensional cases.

Let data be i.i.d. with sample size N from a distribution in family $\{P_f : f \in \mathcal{F}\}$. I have in mind that f is the density or log-density in the case that the observations consist of independent scalars or vectors X_i, and that f is the regression function $f(x) = E[Y|X = x]$ in the case that the observations consist of input-output pairs (X_i, Y_i) for $i = 1, \ldots, N$. For notational simplicity I will write X^N for the data with the understanding that one needs to incorporate response variables in the input-output case.

Suppose the metric $d(f,g)$ is such that $d^2(f,g)$ agrees with $D(P_f||P_g)$ to within constant factors at least for f and g in $\mathcal{F}$. For instance, in density estimation, d can be Hellinger distance and $\mathcal{F}$ can be densities bounded away from zero and infinity. In regression with additive Gaussian error, d can be the L_2 distance between regression functions. In some cases the desired relationship fails for all of $\mathcal{F}$ but it holds within suitable subsets that resolve the minimax rates. As shown in [YB97b] this is true for instance with the L_2 norm on densities.

Let the critical separation ϵ_N be defined to satisfy $\epsilon_N^2 = K(\epsilon_N)/N$, which is of the same order as $\min_\epsilon \{\epsilon^2 + K(\epsilon)/N\}$. In many, but not all cases, one can interpret $K(\epsilon)$ as a dimension of an ϵ-approximating subfamily, such that $\epsilon^2 + K(\epsilon)/N$ reminds us of the familiar tradoff in squared approximation error plus dimension over sample size (or squared bias plus variance) as in the accuracy index in subsection 4.4.

The result in [YB97b] is that the following quantities coincide to within constant factors: the minimax individual risk $\min_{\hat{f}_N} \max_{f \in \mathcal{F}} E d^2(f, \hat{f}_N)$, the minimax relative entropy risk $\min_{\hat{f}_N} \max_{f \in \mathcal{F}} E D(P_f||P_{\hat{f}_N})$, the minimax Cesaro average relative entropy risk, $(1/N)$ times the information capacity of the family of distributions for the data, the critical squared distance ϵ_N^2, the metric entropy at the critical distance divided by the sample size $K(\epsilon_N)/N$, and the index of resolvability of distributions in the family by a mixture that uses a uniform prior over a minimal cardinality ϵ_N-cover.

For the upper bound on minimax risk, the proof is based on the use of the Cesaro average of Bayes estimators (as in subsection 4.9) together with the resolvability bound (from subsection 4.10) on the Cesaro average of relative entropy risks, also known as $(1/N)$ times the total relative entropy. With a uniform prior over an ϵ_N-cover the index of resolvability is bounded (choosing

$\delta = \epsilon_N$) by a constant times ϵ_N^2 plus $K(\epsilon_N)/N$, since the logarithm of the reciprocal of the prior probability is $K(\epsilon_N)$ in this case.

For the lower bound on minimax risk, the proof is based on the use of Fano's inequality (c.f. subsection 1.8) which applied here asserts that averaging over θ with respect to a uniform distribution on an optimal r-packing set (and using the fact that for any estimator $\hat{\theta}$ the projection to the r-net produces an error-free estimate of a net point θ when $d(\hat{\theta}, \theta)$ is less than $r/2$) we have

$$P\{d(\hat{\theta}, \theta) \geq r/2\} \geq 1 - \frac{I(\theta; X^N) + \log 2}{K(r)}$$

Now the mutual information $I(\theta; X^N)$ is not greater than the Shannon capacity C_N which by the resolvability bound on the minimax total relative entropy is not greater than a constant times $N\epsilon_N^2$ plus $K(\epsilon_N)$. (The use of Fano's inequality in this context was first made by Hasminskii [Has78] at the recommended of Mark Pinsker.) Then picking $r = r_N$ a fixed fraction of ϵ_N makes the Kolmogorov capacity $K(r)$ in the denominator at least twice the numerator, and hence makes the probability given above at least $1/2$. Consequently, still incorporating the average over θ as well as over the data,

$$Ed^2(\hat{\theta}, \theta) \geq (r/2)^2 P\{d(\hat{\theta}, \theta) \geq r/2\} \geq r^2/8.$$

Thus for any estimator $\hat{\theta}$ the maximum risk over θ is at least $r_N^2/8$ which is of the same order as $\epsilon_N^2 = K(\epsilon_N)/N$. Together the lower and upper bounds identify this to be the minimax rate.

Implications are given in [YB97b] for a variety of function classes. The proof in [YB97b] we have outlined here goes somewhat beyond the precursors in that it is the order of the metric entropy alone which the result requires for determination of the minimax order of risk of estimation. In contrast the theorem in [Bir83] (used in [Dev87],[Yat88],[BBM97]) demands of the reader determination of a local packing set (such as a hypercube) in which the diameter of the set is the same order as the separation ϵ_N of points in the net and yet the log-cardinality is of the same order as $K(\epsilon_N)$. What is essential to reveal the role of the metric entropy alone is to recognize that the average or maximum of the divergences $D(P_{X^N|\theta}||Q_{X^N})$ of distributions from a centroid $Q_{X^N} = \sum_\theta P_{X^N|\theta} W\{\theta\}$ is what is needed to best control the mutual information $I(\theta; X^N)$ in Fano's inequality (as we do using the resolvability), and not the potentially coarse bounds on this mutual information from the relative entropy diameter $\max_{\theta,\theta'} D(P_{X^N|\theta}||P_{X^N|\theta'})$ [Bir83] or from $NI(\theta; X_1)$ [Has78, BH79]. These course bounds are potentially of order N (unless one restricts to a very localized packing set) whereas $I(\theta; X^N)$ and the capacity C_N grow more slowly (at rate $N\epsilon_N^2$).

5.14 Neural Net Bounds

In this section we use single hidden layer sigmoidal network models as an example setting for presentation of resolvability bounds on cumulative risk of Bayes predictive estimators.

We will consider both dichotomous response models and Gaussian error models in which the the conditional distribution for the response Y given input $X = x$ has mean function $f(x)$ which we model using a neural net. In both cases there will be observations of $(X_i, Y_i)_{i=1}^N$. The inputs X_i will be i.i.d. with an arbitrary and possibly unknown distribution P_X on a given bounded convex set B (such as the cube $[-1, 1]^d$). The risk bounds we give will hold uniformly over all such P_X.

For the dichotomous response case we have $Y_i \in \{-1, 1\}$, with probability of getting a 1 equal to $1/2 + f(X_i)/2$. Here $f(x)$ represents the difference of the probability of getting a 1 and getting a -1 when $X = x$. For the sake of symmetry we are putting the Bernoulli distribution on $\{-1, 1\}$. We will assume in this dichotomous response case that $|f(x)| \leq 1 - \alpha$ is strictly less than 1. If necessary this can be arranged by mixing with a coin flip with probability α.

For the Gaussian error model we have $Y_i = f(X_i) + e_i$ where the e_i are i.i.d. Normal$(0, \sigma^2)$.

Consider the neural net model

$$f_m(x, \theta) = \sum_{j=1}^{m} c_k \psi(a_j \cdot x)$$

parameterized by $\theta = (a_j, c_j)_{j=1}^m$ with internal weight vectors a_j in R^{d+1} and external weights c_j, where $\psi(u)$ is an odd-symmetric sigmoid such as the hyperbolic tangent or $2\phi(u) - 1$ where $\phi(u) = e^u/(1 + e^u)$ is the logistic sigmoid. From the odd-symmetry of ψ, we restrict the c_j to be positive, without loss of generality. For simplicity an auxiliary coordinate of x is set to 1 so that the internal weights parameterize the location as well as the orientation and gain of the sigmoids. In the dichotomous response case we will clip the magnitude of $f_m(x, \theta)$ to be not greater than $1 - \alpha$.

For the function f it is assumed to have a spectral norm $C_{f,B}$ which for now is assumed to be not greater than some given v. Here $C_{f,B} = \int |\omega|_B \tilde{F}(d\omega)$ is a first moment of the Fourier magnitude distribution $\tilde{F}$ and $|\omega|_B = \sup_{x \in B} |\omega \cdot x|$ is the norm of the frequency vector that is dual to the domain B for the variable X. The consequence of this assumption (established in [Bar93]) that we need is that that there exists an approximation $f_m^*(x) = \sum_{j=1}^m c_j^* \psi(a_j^* \cdot x)$, with $\sum_{j=1}^m |c_j^*| \leq v$ and $|a_j^*|_B \leq \tau_m$ where τ_m is of order $\sqrt{m} \log m$, achieving

$$||f - f_m^*||^2 \leq \frac{(2v)^2}{m}$$

where the norm of the approximation error is taken in $L_2(P_X)$. Here the exterior weights may be fixed at $c_j = v/m$. This approximation bound holds more

generally assuming that f/v is in the closure of the convex hull of signum functions. We make the more narrow assumption of bounded spectral norm in order to have control on the magnitudes of the internal weights a_j in the model.

Let $P_{X^N,Y^N|f}$ denote the distribution of the sample $(X_i,Y_i)_{i=1}^N$ with the (unknown) true target function f, and let $P_{X^N,Y^N|f_{m,\theta}}$ or $P_{X^N,Y^N|\theta}$ denote the corresponding distribution with $f(x)$ replaced by members of the approximating family $f_{m,\theta}(x) = f_m(x,\theta)$. Let W be a prior distribution that we assign to θ and let $P^{(W)}_{X^N,Y^N}$ denote the resulting mixture.

Our information-theoretic analysis involves examination of the total relative entropy $D(P_{X^N,Y^N|f}||P^{(W)}_{X^N,Y^N})$ which is the cumulative relative entropy risk of the Bayesian predictive distributions. The resolvability bound gives for any subset A of the parameter space

$$\frac{1}{N}D(P_{X^N,Y^N|f}||P^{(W)}_{X^N,Y^N}) \leq \max_{\theta\in A} D(P_{X,Y|f}||P_{X,Y|f_{M,\theta}}) + \frac{1}{N}\log\frac{1}{W(A)}.$$

For the Gaussian error model

$$D(P_{X,Y|f}||P_{X,Y|f_{m,\theta}}) = \frac{1}{2\sigma^2}||f-f_{m,\theta}||^2,$$

and for the dichotomous response model (using the L_1^2 and Chi-square bounds on D)

$$\frac{1}{2}||f-f_{m,\theta}||^2 \leq D(P_{X,Y|f}||P_{X,Y|f_{m,\theta}}) \leq \frac{1}{\alpha}||f-f_{m,\theta}||^2$$

Thus our L_2 approximation bounds are ready made to bound the resolvability. The resolvability bounds for the cumulative risk of the Bayes estimators are comparable to that which was given for constrained least squares estimators in [Bar94].

At a suitable $\theta^* = (a_j^*)_{j=1}^m$ depending on f, with norms bounded by $|a_j^*|_B \leq \tau_m$, the approximation error $||f-f_{m,\theta^*}||$ is bounded by $2v/\sqrt{m}$. Now take A to be the neighborhood of θ^* defined by $A = \{\theta : |a_j - a_j^*|_B \leq 1/\sqrt{m}, j = 1,2,...,m\}$, and use the triangle inequality and the fact that the sigmoid ψ is Lipshitz with $|\psi(u)-\psi(u')| \leq 2|u-u'|$ to obtain for θ in A, that the approximation error $||f-f_{m,\theta}||$ is bounded by $||f-f_{m,\theta^*}|| + 2v/\sqrt{m}$ which is not greater than $4v/\sqrt{m}$. As a consequence of these bounds we have that

$$\frac{1}{N}D(P_{X^N,Y^N|f}||P^{(W)}_{X^N,Y^N}) \leq \frac{16v^2}{cm} + \frac{1}{N}\log\frac{1}{P\{\theta\in A\}},$$

where $c = 2\sigma^2$ in the Gaussian regression case, and $c = \alpha$ in the dichotomous regression case.

It remains to lower bound $P\{\theta \in A\}$ for a specific choice of the prior. Taking for instance a prior that makes the a_j independently uniformly distributed

on $\{|a_j|_B \leq \tau_m + 1/\sqrt{m}\}$ in R^{d+1}, we have $P(A) = 1/(\sqrt{m}\tau_m + 1)^{m(d+1)}$. Consequently,

$$\begin{aligned}\frac{1}{N} D(P_{X^N,Y^N|f}||P^{(W)}_{X^N,Y^N}) &\leq \frac{16v^2}{cm} + \frac{m(d+1)}{N}\log(\sqrt{m}\tau_m + 1) \\ &= O\left(v\left(\frac{d\log N}{N}\right)^{1/2}\right)\end{aligned}$$

for $m \sim v(N/(d\log N))^{1/2}$.

Note that the second term in the bound involves the ratio of the parameter dimension $k_m = m(d+1)$ and the sample size N. Thus the bound is similar to the familiar squared approximation error plus parameter dimension divided by the sample size as in section 4.4. A difference here is the log factor, presumably because the nonlinear neural net models with potentially large internal parameter values lack the homogeneous metric dimension property.

Nevertheless, the neural net model (and other similar nonlinear models [BBM97]) have a particularly nice flexibility of approximation to achieve the indicated accuracy using only order m times d parameters. In contrast, linear approximation (e.g. by tensor product expansions as in subsection 4.7) requires exponentially many terms in d to achieve comparable accuracy for functions of bounded spectral norm [Bar93]. The logarithm is a minor price to pay for the gain in the approximation versus dimension tradeoff.

Recall that the relative entropy distance between the joint distributions is related to an average relative entropy distance between $f(x)$ and the Bayes estimates $\hat{f}_{n,Bayes}(x) = \int f_m(x,\theta)p(\theta|X^n,Y^n)d\theta$, averaging over samples of size $n = 0, 1, ..., N-1$. Indeed, by the chain rule

$$\frac{1}{N} D(P_{X^N,Y^N|f}||P^{(W)}_{X^N,Y^N}) = \frac{1}{N}\sum_{n=0}^{N-1} ED(P_{X,Y|f}||P_{X,Y|f_{n,Bayes}}).$$

Let the Cesaro average of the Bayes estimates be $\hat{f}_N(x) = \frac{1}{N}\sum_{n=0}^{N-1}\hat{f}_{n,Bayes}(x)$. Then by the convexity of the relative entropy and its relationship to the squared L_2 norm, we conclude with the following bound on the mean squared error,

$$\begin{aligned}\frac{1}{\underline{c}} E||f - \hat{f}_N||^2 &\leq ED(P_{X,Y|f}||P_{X,Y|\hat{f}_N}) \\ &\leq \frac{1}{N}\sum_{n=0}^{N-1} ED(P_{X,Y|f}||P_{X,Y|f_{n,Bayes}}) \\ &= O\left(v\left(\frac{d\log N}{N}\right)^{1/2}\right),\end{aligned}$$

where $\underline{c}$ is $2\sigma^2$ in the Gaussian regression case and 2 in the dichotomous regression case.

If, as is usually the case, a bound on the spectral norm is not known in advance, one can incorporate in the prior distribution the parameter v for the sum of the external coefficients c_j. Moreover, one can mix with a prior various size models m. By such strategies, one can achieve accuracy given by the resolvability bound $C_{f,B}\left(\frac{d \log N}{N}\right)^{1/2}$, without prior knowledge of what size network is best. See also the discussion on model selection and mixing below.

This completes the information-theoretic proof of the accuracy of neural net estimators based on the Bayesian predictors. As a consequence of these bounds, it is sufficient to have a polynomially bounded sample size to obtain an accurate estimates of a target function with a polynomially bounded spectral norm.

The analysis of Bayes posterior mean estimates rather than optimization of penalized empirical risk is very much motivated by interest in computational issues of estimation. The idea is that while the multimodality of the empiricial risk surfaces creates a major obstacle to reliable optimization, there remains the possibility to obtain Monte Carlo computations of posterior means $\hat{f}_{n,Bayes}(x) = \int f_m(x,\theta)p(\theta|X^n, Y^n)d\theta$ by sampling from the posterior distribution and averaging $f_m(x,\theta)$. As discussed in section 3, stochastic gradient methods are designed for this purpose in which the posterior distribution plays the role of the target stationary distribution, but it remains to be seen whether there is a satisfactory form of rapid convergence to stationarity suitable for accurate Monte Carlo averages for these multimodal models.

5.15 Worst Case Regret

We return our attention to the prediction problem in subsection 4.9 in which for each $n = 1, 2, \ldots N$, having seen $x_1, \ldots x_{n-1}$ we are to provide a conditional density function $q(x|x_{n-1}, \ldots, x_1)$ for the next observation x_n (with respect to a reference measure λ). Our aim is to have a large value for this density when evaluated at the heretofore unseen x_n, and in particular to have a large product $q(x_1, \ldots, x_N) = \prod_{n=1}^N q(x_n|x_{n-1}, \ldots, x_1)$ compared to the best within a certain class of predictive densities $p(x|x_{n-1}, \ldots, x_1, \theta)$. For motivation, think of a weatherman declaring for each day a probability of rain, where his economic interest in accurate predictions is forced by having his wealth at the station multiplied by $q(x_n|x_{n-1}, \ldots, x_1)$ times some odds each day. For other gambling, learning, and data compression motivations see CT91,CO96,XB97b and section 5.

The game we consider here is the choice of $q_N(x_1, \ldots, x_N)$ that minimizes the maximum ratio

$$\max_{x_1,\ldots,x_N} \max_{\theta} p(x_1, \ldots, x_N|\theta)/q(x_1, \ldots, x_N)$$

subject to q having sum or integral $\int q(x^N)$ equal to 1 (where the integral is with respect to the N-fold product of the reference measure λ). With this constraint

the choice of joint probability function coincides the choice of a sequence of functions $q(x|X_{n-1},\ldots,X_1)$ for $x \in \mathcal{X}$ interpreted operationally as providing conditional probabilities that a player declares for the next outcome. Here there is no presumption of knowledge of a distribution governing the data (indeed the formulation does not even presuppose that there is a governing distribution).

The target level of performance is the best value $p_{max}(x_1,\ldots,x_N) = \max_\theta p(x_1,\ldots,x_N|\theta)$ at time horizon N among those achieved by players that use predictive densities $p(x|x_{n-1},\ldots,x_1,\theta)$ in a given family. The value $p_{max}(x_1,\ldots,x_N) = p(x_1,\ldots,x_N|\hat{\theta}_N)$ is not itself achievable. Indeed, its sum or integral will be greater than 1 and the value of $\hat{\theta}_N$ revealed with hindsight, is not available prior time N (except for one lucky player unknown to us in advance who happens to use $\theta = \hat{\theta}_N$ for all n). An equivalent game is obtained by taking the logarithm of the ratio which decomposes into a sum of logarithmic regrets for prediction for each $n \leq N$.

The solution to this game (due to Shtarkov [Sht88] in a data compression context) is to take $q_N(x_1,\ldots,x_N)$ to be the normalization of the maximum likelihood

$$p_{max}(x_1,\ldots,x_N)/c_N$$

where c_N is the normalization constant

$$c_N = \int p_{max}(x_1,\ldots,x_N).$$

With this choice the ratio $\max_\theta p(x_1,\ldots,x_N|\theta)/q(x_1,\ldots,x_N)$ is the same for all $x_1,\ldots,x_N$ and is equal to c_N. For any other joint probability assignment the density must be smaller for some sequence. It follows that c_N is the minimax value of the ratio.

This same c_N arose in the examination of consistency of maximum likelihood (subsection 4.1), where finiteness of c_N for some N is a key requirement in a set of sufficient conditions for consistency.

It is not practical to determine the conditional distributions for prediction for each $n \leq N$ based on the normalized maximum likelihood $p_{max}(x_1,\ldots,x_N)/c_N$. Statistical analysis (with an information-theoretic flair) comes to the rescue by consideration of Bayes procedures to give an approximately optimal (and often practical) solution that is easier to interprete. Here $q(x_1,\ldots,x_N)$ is chosen to be a Bayes mixture $p_W(x_1,\ldots,x_N) = \int p(x_1,\ldots,x_N|\theta)W(d\theta)$, for which the conditionals are available from posterior distributions. Analogs of the general resolvability bound are available for the logarithm of the regret, by restricting the integral to a neighborhood of $\hat{\theta}_N$ and using the maximum ratio in the neighborhood.

For smooth families and prior densities, Laplace's approximation to the ratio $p_{max}(x_1,\ldots,x_N)/\int p(x_1,\ldots,x_N|\theta)w(\theta)d\theta$ is

$$\left(\frac{N}{2\pi}\right)^{k/2} \frac{|\hat{I}(\hat{\theta}_N)|^{1/2}}{w(\hat{\theta}_N)}$$

where $\hat{I}(\theta)$ is the empirical Fisher information, defined as the Hessian of $-(1/N)\log p(x_1,\ldots,x_N|\theta)$ when the maximum likelihood estimate $\hat{\theta}_N$ is interior to the parameter space. Thus when $\hat{I}(\theta)$ is independent of the data, as holds when the family has a representation in regular exponential form, an approximate constant ratio strategy is obtained by making the prior $w(\theta)$ be proportional to $|I(\theta)|^{1/2}$, that is by the choice of Jeffreys' prior. Consequently, the approximation that should hold for the minimax ratio for θ in a set S is

$$c_N \sim \frac{N}{2\pi}^{k/2} \int_S |I(\theta)|^{1/2} d\theta$$

The corresponding minimax log regret approximation is

$$\frac{k}{2}\log\frac{N}{2\pi} + \log\int_S |I(\theta)|^{1/2} d\theta$$

which is in agreement with the result of [CB94] for minimax expected log regret (see subsection 4.11) for compact sets S interior to the parameter space (for the expected log regret case it was not necessary to restrict to exponential families).

So far, in work in progress with Takeuchi, we have confirmed that the approximation holds for exponential families with S interior to the parameter space and for one-dimensional exponential families we find modifications of the prior that give the approximate minimax answer on the whole natural parameter space.

With Qun Xie, the minimax regret has been determined by this Bayes method for the whole simplex of distributions for a finite alphabet [**?**]. In all these cases the answer agrees with the form $\frac{k}{2}\log\frac{N}{2\pi} + \log\int |I(\theta)|^1/2 d\theta$ though various modifications to Jeffreys' prior are required to handle boundary behavior.

For non-exponential families the situation is complicated by the fact that the ratio of $|\hat{I}(\hat{\theta}_N)|^{1/2}$ and $|I(\hat{\theta}_N)|^{1/2}$ is not necessarily close to one uniformly over all sequences.

Another approach to yield the asymptotics of $\int p_{max}(x_1,\ldots,x_N)$ in certain cases is a Riemann integral interpretation Sta95,CO96,Fre96. However, that approach does not reveal feasible methods for computation of the predictive distributions potentially afforded by the Bayesian mixtures.

5.16 Cumulative Risk Bounds for other Loss Functions

Suppose a response variable Y_n is to be predicted from a explanatory variable X_n and past values of the variables $(X_i, Y_i)_{i=1}^{n-1}$. We assume that it is generated according to a distribution $P_{Y_n|X^n,Y^{n-1}}$ unknown to us and possibly a member $P_{Y_n|X^n,Y^{n-1},\theta^*}$ of a parametric family of predictive distributions. We base our predictions using a predictive distribution $Q_{Y_n|X^n,Y^{n-1}}$, which in particular can be the predictive distribution associated with a Bayes mixture.

Suppose the set $\mathcal{Y}$ of possible values for Y is discrete. The rule that minimizes the probability of error if we knew the probability mass function $P\{Y_n = y|X^n, Y^{n-1}\}$ would be to pick $Y_n^* = \text{argmax}P\{Y_n = y|X^n, Y^{n-1}\}$. Not knowing P we use $\hat{Y}_n = \text{argmax}Q\{Y_n = y|X^n, Y^{n-1}\}$. The instantaneous regret is defined as the difference in the probability of errors for prediction of Y_n

$$P\{Y_n \neq \hat{Y}_n|X^n, Y^{n-1}\} - P\{Y_n \neq Y_n^*|X^n, Y^{n-1}\}.$$

The expected regret for $n = 1, 2, \ldots, N$ is defined as the expected relative frequence of errors

$$\bar{r}_N = \frac{1}{N}\sum_{n=1}^{N} E[P\{Y_n \neq \hat{Y}_n|X^n, Y^{n-1}\} - P\{Y_n \neq Y_n^*|X^n, Y^{n-1}\}].$$

Now since $\hat{Y}_n$ minimizes the predictive probability of error with respect to Q, we can upper bound the expected regret by adding the positive difference $Q\{Y_n \neq Y_n^*|X^n, Y^{n-1}\} - Q\{Y_n \neq \hat{Y}_n|X^n, Y^{n-1}\}$ inside the expected value. Then we have differences between P and Q predictive probabilities at $\hat{Y}_n$ and at Y_n^*. Summing the probability differences over all y yields a bound in terms of total variation distances between the predictive distributions which in turn is bounded using $\sqrt{2D}$.

$$\begin{aligned}
\bar{r}_N &\leq \frac{1}{N}\sum_{n=1}^{N} E \sum_y |P\{Y_n = y|X^n, Y^{n-1}\} - Q\{Y_n = y|X^n, Y^{n-1}\}| \\
&\leq \frac{1}{N}\sum_{n=1}^{N} E(2D(P_{Y_n|X^n,Y^{n-1}}||Q_{Y_n|X^n,Y^{n-1}}))^{1/2} \\
&\leq \left(2\frac{1}{N}\sum_{n=1}^{N} ED(P_{Y_n|X^n,Y^{n-1}}||Q_{Y_n|X^n,Y^{n-1}})\right)^{1/2} \\
&\leq \left(2\frac{1}{N}ED(P_{Y^N|X^N}||Q_{Y^N|X^N})\right)^{1/2}.
\end{aligned}$$

where the last two lines are by the Cauchy-Schwartz inequality and the Chain rule. The expectation is with respect to any distribution on X^N.

We conclude that the upper bounds we have obtained on Kullback-Leibler risk apply to the expected regret in prediction with a zero-one valued loss, by taking a square root. In particular in smooth finite-dimensional parametric families we have from the resolvability bound an expected regret of order

$$\left(\frac{k \log N}{2N}\right)^{1/2}.$$

Comparable lower bounds for expected regret with zero-one loss are in [**?**].

To avoid the square-root, suppose there is an $\alpha = \alpha_P$ such that the difference between $P\{Y_n = y|X^n, Y^{n-1}\}$ at the maximizer Y_n^* and at all other y is at least as large as the gap α_P. Then in the expected regret

$$\bar{r}_N = \frac{1}{N}\sum_{n=1}^{N} E[P\{Y_n = Y_n^*|X^n, Y^{n-1}\} - P\{Y_n = \hat{Y}_n|X^n, Y^{n-1}\}]$$

we obtain an upper bound by multiplying by the ratio $[P\{Y_n = Y_n^*|X^n, Y^{n-1}\} - P\{Y_n = \hat{Y}_n|X^n, Y^{n-1}\}]/\alpha_P$ inside the expection. Then proceeding as before we obtain

$$\begin{aligned}
\bar{r}_N &\leq \frac{1}{N}\sum_{n=1}^{N} E[P\{Y_n = Y_n^*|X^n, Y^{n-1}\} - P\{Y_n = \hat{Y}_n|X^n, Y^{n-1}\}]^2/\alpha_P \\
&\leq \frac{1}{N\alpha_P}\sum_{n=1}^{N} E[sum_y(P\{Y_n = y|X^n, Y^{n-1}\} - Q\{Y_n = y|X^n, Y^{n-1}\})]^2 \\
&\leq \frac{2}{N\alpha_P}\sum_{n=1}^{N} ED(P_{Y_n|X^n,Y^{n-1}}||Q_{Y_n|X^n,Y^{n-1}}) \\
&\leq \left(2\frac{1}{N\alpha_P} ED(P_{Y^N|X^N}||Q_{Y^N|X^N})\right).
\end{aligned}$$

Consequently, the expected regret with zero-one loss is bounded by using the bounds on the Kullback-Leibler risk and dividing by the gap α. In this way we avoid the square-root.

The regret bounds given here are in [HB92] (for the binary response case) and in [BCH93] for general discrete-valued Y.

5.17 Model Selection

5.18 Statistics and Learning Reprise

6 Data Compression

References

[Ait75] J. Aitchison. Goodness of prediction fit. *Biometrika*, vol.62, pp.547-554, 1975.

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings 2nd International Symposium on Information Theory* (Petrov and Csaki, eds.) Akademia Kiado, Budapest, pp.267-281, 1973.

[AC88] P. H. Algoet and T. M. Cover. A sandwich proof of the Shannon-McMillan-Breiman theorem. *Ann. Probab.* vol.16, pp.899-909, 1988.

[AS66] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Royal Statist. Society Ser. B*, vol.28, pp.131-142.

[Bah71] R. R. Bahadur. *Some Limit Theorems in Statistics* SIAM, Philadelphia, 1971.

[Bar85] A. R. Barron. The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.*, Vol.13, pp.1292-1303, 1985.

[Bar86] A. R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, Vol.14, pp.336-342, 1986.

[Bar87] A. R. Barron. Are Bayes rules consistent in information? In *Problems in Communications and Computation* (Cover and Gopinath, eds.) Springer-Verlag, New York, pp.85-91, 1987.

[Bar88] A. R. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Tech. Report 7, Univ. Illinois, Dept. Statist., 1988.

[Bar91] A. R. Barron. Information theory and martingales. Presented at the *1991 International Symposium on Information Theory.*

[Bar93] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inform. Theory* vol.39, pp.930-945, 1993.

[Bar94] A. R. Barron. Approximation and estimation bounds for artificial neural networks, *Machine Learning* vol.14, 115-133.

[BBM97] A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. To appear in *Probabability Theory and Related Fields.*

[BCH93] A. R. Barron, B. Clarke, D. Haussler, Information bounds for the risk of Bayesian predictions and the redundancy of universal codes. Proc. IEEE Intern. Symp. on Inform. Theory, p.54, 1993.

[BH95] A. R. Barron and N. Hengartner. Information theory and supereffi-ciency. Submitted 1995. In revision 1997.

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, vol.37, pp.1034-1054,1738.

[BSW97] A. R. Barron, M. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. Submitted 1997.

[BS91] A. R. Barron and C.-H. Sheu. Approximation of densities by sequences of exponential families. *Ann. Statist.*, vol.19, pp.1347-1369, 1991.

[Bir83] L. Birgé. Approximation dans les espaces metriques et theorie de l'estimation. *Z. Warsch. Verw. Gebiete*, vol.65, ppl.181-237, 1983.

[Bir86] L. Birgeé. On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, vol.71, pp.271-291, 1986.

[BM93] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, vol.97, pp.113-150, 1993.

[BM97] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* 1997.

[BH79] J. Bretagnolle and C. Huber. Estimation des densités: Risque minimax. *Z. Warsch. Verw. Gebiete*, vol.10, pp.119-137, 1979.

[Bro82] . A proof of the central limit theorem motivated by the Cramér-Rao inequality. In *Statistics and Probability: Essays in Honor of C. R. Rao* (Kallianpur, Krishnaiah, and Ghosh, eds.) North-Holland, Amsterdam, 1982.

[Cen82] N. N. Cencov. *Statistical Decision Rules and Optimal Inference.* Nauka, Moscow, 1972; Amer. Math. Soc. Transl., vol.53, Providence, R.I., 1982.

[Che56] H. Chernoff. Large sample theory – parametric case. *Ann. Math. Statist.* vol27, pp.1-22, 1956.

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, vol.36, pp.453-471, 1990.

[CB94] B. S. Clarke and A. R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Planning and Inference*, vol.41, 37-60, 1994.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory* Wiley, New York, 1991.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations *Studia Sci Math. Hangar.*, vol.2, pp.299-318, 1967.

[Csi75] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* vol.3, pp.146-158, 1975.

[Csi84] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* vol.12, pp.768-793, 1984.

[Dav73] L. D. Davisson. Universal noiseless coding. *IEEE Trans. Inform. Theory*, vol.19, pp.783-795, 1973.

[DLG80] L. Davision and A. Leon-Garcia. A source matching approach to finding minimax codes. *IEEE Trans. Inform. Theory* vol.26, pp. 166-174.

[Dev87] L. Devroye. *A Course in Density Estimation.* Birkhauser, Boston, 1987.

[EP83] S. Y. Efroimovich and M. S. Pinsker. Estimation of square-integrable probability density of a random variable. *Problems Inform. Transmission*, vol.18, pp.175-189, 1983.

[FMG92] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, vol.38, pp.1258-1270, 1992.

[Fer67] T. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York, 1967.

[Fre96] Y. Freund. Predicting a binary sequence almost as well as the optimal biased coin. *Proc. 9th Annual Workshop on Computational Learning Theory*, pp. 89-98, Morgan Kaufmann, 1996.

[Fri73] J. Fritz. An information-theortetical proof of limit theorems for reversible Markov processes. *Trans. Sixth Prague Conf. on Inform. Theory, Statist. Decision Functions, Random Processes* Prague, Sept. 1971, Academia Publ., Czech. Acad. Science, 1973.

[Has78] R. Z. Hasminskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Probab. Appl.*, vol.23, pp.794-796, 1978.

[Gal79] R. Gallager. Source coding with side information and universal coding. MIT Technical Report LIDS-P-937, 1979.

[Hau92] D. Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, vol.100, pp.78-150, 1992.

[Hau95] D. Haussler. A general minimax result for relative entropy. Manuscript, 1995.

[HB92] D. Haussler and A. R. Barron. How well do Bayes methods work for on-line prediction of $\{\pm\}$ values? *Proc. NEC Symposium on Computation and Cognition*, 1992.

[HKW97] D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. To appear in *IEEE Tranactions on Information Theory*, 1997.

[HO95a] D. Haussler and M. Opper. General bounds on the mutual information between a parameter and n conditionally independent observations, *Proc. 8th Annual Workshop on Computational Learning Theory*, Morgan Kaufmann, 1995.

[HO95b] D. Haussler and M. Opper. Mutual information, metric entropy, and risk in estimation of probability distributions. manuscript submitted 1995.

[IH80] I. Ibragimov and R. Hasminskii *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York, 1980.

[IH82] Bounds for the risks of non-parametric regression estimates. *Theory Probab. Appl.* vol.27, pp.84-99, 1982.

[Jay57] E. T. Jaynes. Information theory and statistical mechanics. I. *Phys. Rev.* vol.106, pp.620-630, 1957.

[Ken64] D. G. Kendall. *Information Theory and the limit theorem for Markov chains and processes with a countable infinity of states.* Ann. Inst. Stat. Math.} vol.15, pp.137-143, 1964.

[Kul59] S. Kullback. *Information Theory and Statistics*, Wiley, New York, 1959.

[Kul67] S. Kullback. A lower bound for discrimination in terms of variation, *IEEE Trans. Inform. Theory*, vol.13, pp.126-127, 1967.

[KK80] S. Kullback, J. C. Keegel, and J. H. Kullback. *Topics in Statistical Information Theory.* Springer-Verlag, Berlin, 1980.

[LeC53] L. Le Cam. On some asymptotic properties of maximum likelihood and related Bayes estimates. In *University of California Publications in Statistics* (Neyman, Loeve, and Struve, eds.) Cambridge University Press, London, vol.1, pp.277-329, 1953.

[LeC73] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.* vol.1, pp.38-53, 1973.

[LeC86] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York, 1986.

[Lin59] Yu. V. Linnik. An information-theoretic proof of the central limit theorem with the Lindeberg condition. *Theory Probab. Appl.* vol.4, pp.288-299, 1959.

[Maz85] V. G. Mazja. *Sobolev Spaces*, Springer-Verlag, Berlin, New York, 1985.

[MF92] N. Merhav and M. Feder. Universal sequential learning ad decisions from individual data sequences. *Proc. 5th Annual Workshop on Computational Learning Theory*, ACM Press, pp.413-427, 1992.

[Moy61] S. C. Moy. Generalizations of Shannon-McMillan theorem. *Pacific J. Math.* vol.11, 705-714, 1961.

[Ore85] S. Orey. On the Shannon-Perez-Moy theorem. *Contemp. Math.* vol.41, pp.319-327, 1985.

[Pin64] M. S. Pinsker. *Information and Information Stability of Random Variables*, Transl. by A. Feinstein. Holden-Day, San Francisco, 1964.

[Ren60] A. Rényi. On measures of entropy and information. In *Proc. Fourth Berkeley Symp. on Math. Statist. and Probab.*, vol.I, pp.547-561, 1960.

[Ris83] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Statist.* vol.11, 416-431, 1983.

[Ris84] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* vol.30, pp.629-636, 1984.

[Ris86] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.* vol.14, pp.1080-1100, 1986.

[Sch65] L. Schwartz. On Bayes procedures. *Z. Warsch. Verw. Gebiete*, vol.4, pp.10-26, 1965.

[Sch93] C. Schroeder. I-projection and conditional limit theorems for discrete parameter Markov processes. *Ann. Probab.* vol.21, pp.721-758, 1993.

[Sha48] C. Shannon. A mathematical theory of communication. *Bell System Tech. J.* vol.27, pp.379-423,623-656, 1948.

[Sht88] Yu. M. Shtarkov. Universal squential coding of single messages. *Probl. Inform. Transmission* vol.23, pp.3-17, 1988.

[Sta59] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. and Control*, vol.2, pp.101-112, 1959.

[Top79] F. Topsoe. Information theoretical optimization techniques. *Kybernetika* vol.15, pp.8-27, 1979.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, 1982.

[Vov90] V. Vovk. Aggregating strategies Proc. 3rd Annual Workshop on Computational Learning Theory, Morgan Kaufmann, pp.371-383, 1990.

[Wal49] A. Wald. Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* vol.20, pp.595-601, 1949.

[WMF94] M. J. Weinberger, N. Merhav and M. Feder, Optimal sequential probability assignment for individual sequences. *IEEE Trans. Inform. Theory* vol.40, pp.384-396, 1994.

[XB97a] Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Trans. Inform. Theory* vol.43, pp.646-657, 1997.

[XB97b] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inform. Theory*, submitted 1996, in revision 1997.

[Yam94a] K. Yamanishi. Generalized stochastic complexity and its application to learning. In *Proc. 1994 Conf. Information Science and Systems*, vol.2, pp.763-768.

[Yam94a] K. Yamanishi. The minimum L-complexity algorithm and its applications to learning non-parametric rules. In *Proc.— 7th Annual Workshop on Computational Learning Theory*, ACM Press, pp.173-182, 1994.

[YB97a] Y. Yang and A. R. Barron. An asymptotic property of model selection criteria. Revised for *IEEE Trans. Inform. Theory*, 1997.

[YB97b] Y. Yang and A. R. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, Submitted 1995, Revised 1997.

[Yat88] Y. G. Yatracos. A lower bound on the error in nonparametric regression type problems. *Ann. Statist.*, vol.16, pp.1180-1187, 1988.