MUTUAL INFORMATION AND BAYES METHODS FOR LEARNING A DISTRIBUTION

DAVID HAUSSLER
Computer and Information Sciences, UC Santa Cruz, CA 95064 USA
E-mail: haussler@cse.ucsc.edu

and

MANFRED OPPER
Universität Würzburg, E-mail: opper@physik.uni-wuerzburg.de

ABSTRACT

Each parameter w in an abstract parameter space W is associated with a different probability distribution on a set Y. A parameter w is chosen at random from W according to some *a priori* distribution on W, and n conditionally independent random variables $Y^n = Y_1, \ldots Y_n$ are observed with common distribution determined by w. Viewing W as a random variable, we obtain bounds on the mutual information between the random variable W, giving the choice of parameter, and the random variable Y^n, giving the sequence of observations. This quantity is the cumulative risk in predicting $Y_1, \ldots Y_n$ under the log loss, minus the risk if the true parameter w is known. The upper bounds are stated in terms of the Laplace transform of the rate of growth of the volume of relative entropy neighborhoods in the parameter space W, and the lower bounds are given in terms of the corresponding quantity using Hellinger neighborhoods. We show how these bounds can be interpreted in terms of an average local dimension of the parameter space W under suitable conditions.

1. Introduction

Let us assume that the state of nature is defined by a parameter w in an abstract parameter space W. We are not allowed to observe the state of nature directly, but rather must make inferences about it indirectly by observing a sequence of observations $y^n = y_1, \ldots, y_n$. These observations are the values of n random variables $Y^n = Y_1, \ldots, Y_n$ that are conditionally independent given the true state of nature w and have a common distribution determined by this true state of nature. This general setup is fundamental in statistics and related disciplines, including coding and data compression [6], and computational learning theory [8,10,9,1,14]. It is usually called *density estimation* or *parameter estimation* in statistics, depending on the nature of the parameter space W. In this paper we explore this setup from an information theoretic point of view.

We can measure our progress in learning about the true state of nature by measuring how well we are able to predict the observation y_{t+1} after seeing the previous observations $y_1, \ldots y_t$. Let us assume that Y is discrete. If, after seeing these previous observations, we produce the estimate $\hat{P}_t$ of the distribution on Y defined by the true

state of nature, then let us define our *loss* in predicting the observation y_{t+1} to be $-\log \hat{P}_t(y_{t+1})$. If the logarithm base 2 is used, then this loss can be interpreted as the number of bits needed to encode y_{t+1}, making appropriate use of the estimated distribution $\hat{P}_t$. The *total loss* in sequentially predicting the observations $y_1, \ldots, y_n$ can be defined as $-\sum_{t=1}^{n} \log \hat{P}_{t-1}(y_t)$. This can be viewed as the number of bits need to encode $y_1, \ldots, y_n$ by an adaptive coding scheme that makes use of the estimated distributions $\hat{P}_0, \ldots, \hat{P}_{n-1}$. Analogous definitions can be made for continuous Y using estimated densities.

It is useful to compare the total loss incurred by using a particular method for obtaining the estimated distributions $\hat{P}_t$ to the total loss that would be obtained if we knew the true state of nature, and thus used the true distribution on Y to encode each y_t. The difference between these two losses is called the *regret* or *net loss*. The expected regret, with respect to the random choice of $y_1, \ldots, y_n$ according to a fixed state of nature, measures the average extra number of bits we need to encode $y_1, \ldots, y_n$ over and above the average number of bits required by the optimal encoding method, which would use the true distribution. This is called the *redundancy* in coding theory, and is generally called *risk* in statistics (with respect to arbitrary loss or regret functions).

The risk of any method depends on the true state of nature $w \in W$. Taking a Bayesian approach, let us define an *a priori* distribution on the parameter space W so that W itself can be viewed as a random variable. We can then quantify the performance of any method that produces estimated distributions $\hat{P}_t$ by its average risk when the true state of nature is drawn at random according to the prior distribution on W. In this case the posterior distribution $P(Y_{t+1}|y_1, \ldots, y_t)$ is the unique optimal choice for $\hat{P}_t$, and the average risk (redundancy) of this optimal method (the *Bayes method*) is called the *Bayes risk*. A simple calculation (see e.g. [8]) shows that the Bayes risk is equal to the mutual information $I(W; Y^n)$ between the random variable W and the random variable $Y^n = Y_1, \ldots, Y_n$, which can be interpreted as the average amount of information contained in the observation sequence Y^n about the true state of nature w. This leads to the other standard interpretations of the mutual information in terms of density estimation and coding theory.

Because of it's key role in several areas, the mutual information $I(W; Y^n)$ has been investigated by numerous authors. Early work by Ibragimov and Hasminskii showed that $I(W; Y^n) \approx (D/2) \log n$ when Y is real-valued and the conditional distributions on Y are a smooth family of densities indexed by real-valued parameter vectors w of dimension D, and certain other conditions apply [12]. In this case they were even able to estimate the lower order additive terms in this approximation, which involve the Fisher information and the entropy of the prior. Further related results were given by Efroimovich [7] and Clarke [4]. Clarke and Barron gave a detailed analysis, with applications, of the risk (redundancy) of the Bayes method as a function of the true state of nature [5]. Related lower bounds, which are often quoted, were obtained

by Rissanen [13], based on certain asymptotic normality assumptions. In this paper we extend work from [11,8,2] that seeks to obtain bounds on $I(W;Y^n)$ in more general settings.

In this paper we build on results given in [11]. The approach taken here is to relate the mutual information $I(W;Y^n)$ directly to a kind of average local dimension of the parameter space W. The distance between two states of nature w and w^* is measured in terms of the distance between the conditional distributions on the observation space Y that they define. This distance is defined as the Hellinger distance between the distributions. The volume of a Hellinger neighborhood of radius r around the true state of nature w^* is defined as the prior probability of all $w \in W$ within distance r of w^*. The rate at which this volume scales as a function of r is shown to be the key quantity that determines the mutual information for large n for many cases. The exponent in the rate of growth can be defined as the local dimension of W at the point w^*.

2. Basic Definitions and Statement of Main Result

We use the following notational conventions. For a random variable X, P_X denotes the distribution function for X, and if X has a density then it is denoted p_X. For countable X, $P(x)$ denotes the probability that $X = x$ and for continuous X, $p(x)$ is shorthand for $p_X(x)$. For a (measurable) function $f(x)$, $\mathbb{E}_X f(x)$ denotes the expectation of f. Given random variables X and Y, $P_{X,Y}$ denotes their joint distribution, $\mathbb{E}_{X,Y} f(x,y)$ the expectation of f with respect to this joint distribution, etc. The conditional distribution of Y given that $X = x$ is denoted $P_{Y|x}$, and $p_{Y|x}$ denotes the conditional density. $P(y|x)$ and $p(y|x)$ denote the probability and conditional density of y given $X = x$, resp. The conditional expectation of f given that $X = x$ is denoted $\mathbb{E}_{Y|x} f(x,y)$. Similar notation is used to condition on an event. Finally, the marginal distribution of Y is denoted P_Y, a specific marginal probability is denoted $P(y) = \mathbb{E}_X P(y|x)$, the marginal density is denoted p_Y, and a specific density value is denoted $p(y) = \mathbb{E}_X p(y|x)$. Throughout the paper we employ the convention that $0 \ln 0 = 0 \ln \infty = 0 \ln \frac{0}{0} = 0$.

Let $W, Y_1, Y_2, \ldots$ be random variables such that $Y_1, Y_2, \ldots$ are conditionally independent and identically distributed given W, their common distribution denoted by $P_{Y|w}$ for each $w \in W$. The random variable W takes values in an arbitrary set. For simplicity we assume that either Y takes values in a countable set, or Y is continuous[a] with conditional densities $p_{Y|w}$ defined for each $w \in W$. Our default assumption will be that Y is countable, and all our results will be stated explicitly for this case, with comments on how analogous results can be obtained for continuous Y. All functions

[a]Specifically, we assume that conditional densities $p_{Y|w}$ are defined with respect to a common sigma-finite measure on a suitable probability space, such as k-dimensional real space with the Lebesgue measure.

in our results are assumed to be measurable without explicit mention.

For each $n \geq 1$, let $Y^n = Y_1, \ldots, Y_n$ and let y^n denote a typical value of Y^n. We give upper and lower bounds on the mutual information between W and Y^n, defined for countable Y by

$$I(W; Y^n) = \mathbb{E}_{W,Y^n} \ln \frac{P(y^n|w)}{P(y^n)},$$

and similarly for continuous Y, using the corresponding densities denoted with lowercase p.

In obtaining these bounds, we use two notions of "distance" between probability distributions. Let $P = \{p_i\}$ and $Q = \{q_i\}$ be two probability mass functions on a countable set. The *Kullback-Leibler (KL) divergence* (or *relative entropy distance*) between P and Q is defined by

$$D_K(P||Q) = \sum_i p_i \ln \frac{p_i}{q_i}$$

and the *(squared) Hellinger distance* between P and Q is defined by

$$D_H(P, Q) = 2 \sum_i (\sqrt{p_i} - \sqrt{q_i})^2.$$

For densities $p(x)$ and $q(x)$, the analogous definitions are $D_K(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$ and $D_H(p, q) = 2 \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$.

Note that

$$D_H(P, Q) = 4(1 - \sum_i \sqrt{p_i q_i}), \tag{1}$$

($D_H(p, q) = 4(1 - \int \sqrt{p(x)q(x)}dx)$ in the continuous case) and therefore $D_H(P, Q) \leq 4$, in contrast to the KL divergence, which can be arbitrarily large.

We employ these notions of distance between distributions to define two corresponding notions of "distance" between the points in W. For each $w^*, w \in W$, let

$$\Delta_K(w^*, w) = D_K(P_{Y|w^*}||P_{Y|w})$$

and

$$\Delta_H(w^*, w) = D_H(P_{Y|w^*}, P_{Y|w}).$$

The main tool we use is the following result from [11].

Theorem 1 *For every $n \geq 1$,*

$$-I\!E_{W^*} \ln I\!E_W e^{-\frac{n}{4}\Delta_H(w^*,w)} \leq I(W; Y^n) \leq -I\!E_{W^*} \ln I\!E_W e^{-n\Delta_K(w^*,w)}.$$

For any random variable Z taking values in $[0,\infty)$, the function $\phi(n) = \mathbb{E}_Z e^{-nz}$ is the *Laplace transform* of Z. Thus fixing w^*, the bounds on $I(W;Y^n)$ given in Theorem 1 are determined by the logarithms of two Laplace transforms, $\phi_H(n) = \mathbb{E}_W e^{-n\Delta_H(w^*,w)}$ and $\phi_K(n) = \mathbb{E}_W e^{-n\Delta_K(w^*,w)}$, the former being evaluated at $n/4$. It is often possible to make very good estimates of the logarithm of a Laplace transform for larger values of n. This is because for large n the transform $\phi(n) = \mathbb{E}_Z e^{-nz}$ is typically almost entirely determined by the the asymptotic form of the distribution of the random variable Z at values near zero. This means that the bounds in Theorem 1 will be determined by the probability measure of the set of $w \in W$ that lie in a small neighborhood around w^*, as determined by the distance measures $\Delta_K(w^*,w)$ or $\Delta_H(w^*,w)$. It can be shown that in many cases the behavior of these two distances is similar in this small neighborhood (Lemma 2 below).

The following lemma will be used to make this more precise.

Lemma 1 *Let Z be a nonnegative random variable and let $F(x) = P(Z \le x)$ be the distribution function for Z. Then*

$$\liminf_{n\to\infty} \frac{-\ln \mathbb{E}_Z e^{-nz}}{\ln n} \ge \liminf_{n\to\infty} \frac{-\ln F(1/n)}{\ln n}$$

and

$$\limsup_{n\to\infty} \frac{-\ln \mathbb{E}_Z e^{-nz}}{\ln n} \le \limsup_{n\to\infty} \frac{-\ln F(1/n)}{\ln n}.$$

Hence if $\lim_{n\to\infty} \frac{-\ln F(1/n)}{\ln n}$ exists, then $\lim_{n\to\infty} \frac{-\ln \mathbb{E}_Z e^{-nz}}{\ln n}$ also exists and has the same value.

Proof: For any $r > 0$,

$$\begin{aligned}
-\ln \mathbb{E}_Z e^{-nz} &= -\ln\left(P(Z \le r)\mathbb{E}_{Z|z\le r} e^{-nz} + P(Z > r)\mathbb{E}_{Z|z>r} e^{-nz}\right) \\
&\ge -\ln\left(P(Z \le r) + \mathbb{E}_{Z|z>r} e^{-nz}\right) \\
&\ge -\ln\left(F(r) + e^{-nr}\right) \\
&\ge -\ln\left(2 * \max(F(r), e^{-nr})\right) \\
&= -\ln 2 + \min(-\ln F(r), nr)
\end{aligned}$$

Let $r = \frac{\ln^2 n}{n}$. Then it follows that

$$\liminf_{n\to\infty} \frac{-\ln \mathbb{E}_Z e^{-nz}}{\ln n} \ge \liminf_{n\to\infty} \min\left(\frac{-\ln F(\ln^2 n/n)}{\ln n}, \ln n\right) = \liminf_{n\to\infty} \frac{-\ln F(\ln^2 n/n)}{\ln n}.$$

Let $m = \frac{n}{\ln^2 n}$. Then

$$\liminf_{n\to\infty} \frac{-\ln F(\ln^2 n/n)}{\ln n} = \liminf_{m\to\infty} \frac{-\ln F(1/m)}{\ln m} * \frac{\ln m}{\ln n} = \liminf_{m\to\infty} \frac{-\ln F(1/m)}{\ln m}.$$

This establishes the first part.

For the second part, note that for any $r > 0$,

$$\begin{aligned} -\ln \mathbb{E}_Z e^{-nz} &\leq -\ln\left(P(Z \leq r)\mathbb{E}_{Z|z\leq r} e^{-nz}\right) \\ &\leq -\ln F(r) + nr \end{aligned}$$

Let $r = 1/n$ and the second part follows. □

For any $d \geq 0$ we say that a function $F(x)$ defined on the nonnegative reals is *of order d near zero* if

$$\lim_{x\to 0} \frac{\ln F(x)}{\ln x} = d.$$

For each $w^* \in W$, let $F_{w^*,H}(x) = P_W\{w : \Delta_H(w^*, w) \leq x\}$ and $F_{w^*,K}(x) = P_W\{w : \Delta_K(w^*, w) \leq x\}$.

Assumption 1 *We assume that for almost all $w^* \in W$ there exists a finite index $d(w^*)$ such that $F_{w^*,H}$ is of order $d(w^*)$ near zero.*

The number $d(w^*)$ is the exponent in the rate of growth of the volume (measure) of a (squared) Hellinger neighborhood of w^* as a function of its radius x, and hence is related to a kind of "local dimension" of W at w^*. Since $\Delta_H(w^*, w)$ is a squared distance, the real local dimension at w^* is $D(w^*) = 2d(w^*)$. This is the exponent in the rate of growth of the volume of a neighboorhood around w^* measured with respect to the metric $\Delta_H^{(1/2)}(w^*, w)$. Let $D = \mathbb{E}_{W^*} D(w^*)$.

Theorem 2 *Under Assumption 1,*

$$\liminf_{n\to\infty} \frac{I(W; Y^n)}{\ln n} \geq \frac{D}{2}.$$

Proof: By Assumption 1 and Lemma 1, for almost all w^*,

$$\liminf_{n\to\infty} -\frac{\ln \mathbb{E}_W e^{-\frac{n}{4}\Delta_H(w^*,w)}}{\ln n} \geq d(w^*) = \frac{D(w*)}{2}.$$

Thus, using Fatou's lemma, it follows from Theorem 1 that

$$\begin{aligned} \liminf_{n\to\infty} \frac{I(W; Y^n)}{\ln n} &\geq \liminf_{n\to\infty} \mathbb{E}_{W^*} \frac{-\ln \mathbb{E}_W e^{-\frac{n}{4}\Delta_H(w^*,w)}}{\ln n} \\ &\geq \mathbb{E}_{W^*} \liminf_{n\to\infty} \frac{-\ln \mathbb{E}_W e^{-\frac{n}{4}\Delta_H(w^*,w)}}{\ln n} \\ &\geq D/2. \end{aligned}$$

□

To show that the lower bound is tight, we make corresponding but somewhat stronger assumptions on the behavior of the volume of KL-divergence neighborhoods in W.

Assumption 2 *We assume that for almost all $w^* \in W$, $F_{w^*,K}$ is of order $d(w^*)$ near zero (the same as the order of $F_{w^*,H}$).*

The following lemma can be used to give sufficient conditions for Assumption 2 to follow from Assumption 1. It gives a direct comparison between the Hellinger distance and the Kullback-Leibler divergence (see also e.g. [3]).

Lemma 2 [11] *(see also e.g. [3]) For any z, $0 \leq z \leq \infty$, define*

$$b(z) = \frac{z - \ln z - 1}{2(1-\sqrt{z})^2} \in [1/2, \infty].$$

For all distributions $P = \{p_i\}$ and $Q = \{q_i\}$,

$$b(s) D_H(P,Q) \leq D_K(P||Q) \leq b(r) D_H(P,Q),$$

where $r = \inf_i \frac{q_i}{p_i}$ and $s = \sup_i \frac{q_i}{p_i}$. The analogous result holds for any densities p and q with $r = \inf_X \frac{q(x)}{p(x)}$ and $s = \sup_X \frac{q(x)}{p(x)}$.

It is easy to verify that the function $b(z)$ is decreasing and continuous, $b(0) = \infty$, $b(1) = 1$ and $b(\infty) = 1/2$. Therefore, the lemma shows that when the ratios q_i/p_i are near 1 then $D_K(P||Q) \approx D_H(P,Q)$.

Let $z = \inf_{w,w^* \in W}$ and $_{y \in Y} \frac{P(y|w)}{P(y|w^*)}$. If $z > 0$ then by Lemma 2, $\Delta_K(w^*, w)$ and $\Delta_H(w^*, w)$ always differ by at most a fixed constant factor $b(z)$. In this case it is clear that the orders of $F_{w^*,H}$ and $F_{w^*,K}$ near zero are the same for all w^*, and hence Assumption 2 follows from Assumption 1. It is not difficult to construct considerably weaker conditions under which Assumption 2 follows from Assumption 1 as well.

Our last assumption is the following.

Assumption 3 *We assume that there exists a positive integer n_0 and for almost all w^* there exists a positive constant $c(w^*)$ such that $\mathbb{E}_{W^*} c(w^*) < \infty$ and for all $n \geq n_0$, $F_{w^*,K}(1/n) \geq n^{-c(w^*)}$.*

Note that under Assumption 2, $\lim_{n\to\infty} \frac{-\ln F_{w^*,K}(1/n)}{\ln n} = d(w^*)$ for almost all w^*. If this convergence is uniform over w^*, and $\mathbb{E}_{W^*} d(w^*)$ is finite, then we can take $c(w^*) = d(w^*) + \epsilon$ for any positive ϵ and Assumption 3 is satisfied. Thus Assumption 2 implies Assumption 3 in this case.

Theorem 3 *Under assumptions 2 and 3,*

$$\limsup_{n\to\infty} \frac{I(W;Y^n)}{\ln n} \leq \frac{D}{2} < \infty.$$

Proof: It follows from Assumption 3 that for $n \geq n_0$ and almost all w^*,

$$\begin{aligned} -\ln \mathbb{E}_W e^{-n\Delta_K(w^*,w)} &\leq -\ln\left(P\{w : \Delta_K(w^*,w) \leq 1/n\} \mathbb{E}_{W|\Delta_K(w^*,w)\leq 1/n} e^{-n\Delta_K(w^*,w)}\right) \\ &\leq -\ln F_{w^*,K}(1/n) + 1 \\ &\leq c(w^*) \ln n + 1. \end{aligned}$$

Therefore

$$\frac{-\ln \mathbb{E}_W e^{-n\Delta_K(w^*,w)}}{\ln n} \leq c(w^*) + 1$$

for all $n \geq \max(3, n_0)$. Furthermore, by Assumption 2, for almost all w^*

$$\limsup_{n\to\infty} \frac{-\ln \mathbb{E}_W e^{-n\Delta_K(w^*,w)}}{\ln n} \leq d(w^*) = D(w^*)/2.$$

Since $\mathbb{E}_{W^*}(c(w^*)+1)$ is finite, using the bounded convergence theorem it follows from the above and Theorem 1 that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{I(W;Y^n)}{\ln n} &\leq \limsup_{n\to\infty} \mathbb{E}_{W^*} \frac{-\ln \mathbb{E}_W e^{-n\Delta_K(w^*,w)}}{\ln n} \\
&\leq \mathbb{E}_{W^*} \limsup_{n\to\infty} \frac{-\ln \mathbb{E}_W e^{-n\Delta_K(w^*,w)}}{\ln n} \\
&\leq D/2.
\end{aligned}$$

Since $\mathbb{E}_{W^*}(c(w^*)+1)$ is finite, D is also finite. □

Thus Assumptions 1, 2 and 3 together imply that either $I(W;Y^n)$ is $o(\log n)$, or $I(W;Y^n) \sim \frac{D}{2}\ln n$ for some positive D, and in the latter case they provide a formula for the constant D, giving the interpretation of D as an average local dimension of W. Other methods of bounding $I(W;Y^n)$ using Theorem 1 are given in [11].

3. Conclusion

We have given general asymptotic estimates of the mutual information $I(W;Y^n)$ in terms of the average local dimension of W. Some assumptions were required in obtaining our results; it would be interesting to see how these can be weakened. Also, we have treated only the finite dimensional case, which leads to an $O(\ln n)$ scaling law for $I(W;Y^n)$. It would be interesting to see what general scaling laws can be established in the infinite dimensional case using Theorem 1.

Acknowledgments

D. Haussler was supported by NSF grant IRI-9123692 and M. Opper by a Heisenberg fellowship of the DFG.

References

1. S. Amari and N. Murata. Statistical theory of learning curves under entropic loss. *Neural Computation*, 5:140–153, 1993.

2. A. Barron, B. Clarke, and D. Haussler. Information bounds for the risk of Bayesian predictions and the redundancy of universal codes. In *Proc. International Symposium on Information Theory*.
3. L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte gebiete*, 65:181–237, 1983.
4. B. Clarke. *Asymptotic cumulative risk and Bayes risk under entropy loss with applications*. PhD thesis, Dept. of Statistics, University of Ill., 1989.
5. B. Clarke and A. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.
6. T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.
7. S. Y. Efroimovich. Information contained in a sequence of observations. *Problems in Information Transmission*, 15:178–189, 1980.
8. D. Haussler and A. Barron. How well do Bayes methods work for on-line prediction of $\{+1, -1\}$ values? In *Proceedings of the Third NEC Symposium on Computation and Cognition*. SIAM, 1992.
9. D. Haussler, M. Kearns, and R. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14:84–114, 1994.
10. D. Haussler, J. Kivinen, and M. Warmuth. Tight worst-case loss bounds for predicting with expert advice. Technical Report UCSC-CRL-94-36, University of California at Santa Cruz, Computer and Information Sciences, 1994.
11. D. Haussler and M. Opper. General bounds on the mutual information between a parameter and n conditionally independent observations, 1995. submitted to COLT '95.
12. I. Ibragimov and R. Hasminskii. On the information in a sample about a parameter. In *Second Int. Symp. on Information Theory*, pages 295–309, 1972.
13. J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.
14. K. Yamanishi. A learning criterion for stochastic rules. *Machine Learning*, 1992. Special Issue on the Proceedings of the 3nd Workshop on Computational Learning Theory.