In many SPECT and PET tomographic geometries, the $m \times n$ $(m \geq n)$ system response $((P_{j|i}))$ is a sparse matrix, i.e., its number of nonzero elements is only $O(n)$ as compared to $O(n^2)$ for the nonsparse case. Note, however, that even when the system response matrix is sparse, the matrix A (25) is not generally sparse, and it would appear that the recursive algorithm (10) of Corollary 1 requires $O(n^2)$ memory storage to store the $n \times n$ matrix A. In the present case, however, we only require $O(n)$ memory storage since it is seen that, using (25) in (10), the recursion collapses into a set of p vector recursions which only require storing the n parameters of the vector $\underline{\theta}$, the np entries of $\beta^{(k)}$, and the $O(n)$ nonzero entries of the sparse matrix $((P_{j|i}))$. Because of this feature, we have been able to implement this recursive CR bound on relatively large image reconstruction problems [13].

The rate of convergence of the recursive CR bound algorithm is determined by the maximum eigenvalue $\rho(A)$ of A specified by (25). For a fixed system matrix $((P_{j|i}))$, the magnitude of this eigenvalue will depend on the image intensity $\underline{\theta}$. Assume for simplicity that with probability 1 any emitted gamma ray is detected at some detector, i.e., $\Sigma_{d=1}^{m} P_{d|b} = 1$ for all b. Since $\text{trace}(A) = \Sigma_{i=1}^{n} \lambda_i$, where $\{\lambda_i\}_{i=1}^{n}$ are the eigenvalues of A, using (25) it is seen that the maximum eigenvalue $\rho(A)$ must satisfy

$$\frac{1}{n}\,\text{trace}\,(A) = 1 - \frac{1}{n}\sum_{l=1}^{m} \frac{\sum_{i=1}^{n} P_{l|i}^2 \theta_i}{\sum_{j=1}^{n} P_{l|j}\theta_j} \leq \rho(A) < 1. \quad (26)$$

A consequence of the inequality $(\Sigma_i P_{l|i}\theta_i)^2 \leq \Sigma_i \theta_i \cdot \Sigma_i \theta_i P_{l|i}^2 \theta_i$ is

$$\frac{1}{n}\,\text{trace}\,(A) \leq 1 - \frac{1}{n}. \quad (27)$$

where equality occurs if $P_{l|i}$ is independent of i. On the other hand, as the intensity $\underline{\theta}$ concentrates an increasing proportion $1 - \epsilon$ of its mass on a single pixel k_o, e.g.,

$$\theta_i = \begin{cases} (1-\epsilon)\dfrac{n-1}{n}\displaystyle\sum_{b=1}^{n}\theta_b, & i = k_o, \\ \epsilon\dfrac{1}{n}\displaystyle\sum_{b=1}^{n}\theta_b, & i \neq k_o \end{cases},$$

we obtain $(1/n)\,\text{trace}\,(A) = 1 - 1/n + O(\epsilon)$. Thus for this case we have, from (26), $1 - 1/n + O(\epsilon) \leq \rho(A) < 1$. Since the number of pixels n is typically very large, this implies that the asymptotic convergence rate of the recursive algorithm will suffer for image intensities which approach that of an ideal point source, at least for this particular choice of splitting matrix F_X.

V. Conclusion and Future Work

We have given a recursive algorithm which can be used to compute submatrices of the CR lower bound F_Y^{-1} on unbiased multidimensional parameter estimation error covariance. The algorithm successively approximates the inverse Fisher information matrix F_Y^{-1} via a monotonically convergent splitting matrix iteration. We have also given a statistical methodology for selecting an appropriate splitting matrix F which involves application of a data processing theorem to a complete-data–incomplete-data formulation of the estimation problem. We are developing analogous recursive algorithms to compute matrix CR-type bounds for constrained and biased estimation, such as those developed in [14], [15].

References

[1] A. R. Kuruc, "Lower bounds on multiple-source direction finding in the presence of direction-dependent antenna-array-calibration errors," M.I.T. Lincoln Laboratory, Tech. Rep. 799, Oct. 1989.

[2] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore: The Johns Hopkins University Press, 1989.

[3] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.

[4] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. MI-1, no. 2, pp. 113–122, Oct. 1982.

[5] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comput. Assisted Tomogr.*, vol. 8, no. 2, pp. 306–316, Apr. 1984.

[6] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag, 1981.

[7] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic, 1970.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge: Cambrige Univ. Press, 1985.

[9] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. R. Stat. Soc., Ser. B*, vol. 44, no. 2, pp. 226–233, 1982.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc., Ser. B*, vol. 39, pp. 1–38, 1977.

[11] A. O. Hero and J. A. Fessler, "Asymptotic convergence properties of EM-type algorithms," Commun. Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, Technical Report insert 282, April 1993. Also to appear in *Statistica Sinica*, Jan. 1995.

[12] A. O. Hero and L. Shao, "Information analysis of single photon computed tomography with count losses," *IEEE Trans. Med. Imag.*, vol. 9, no. 2, pp. 117–127, June 1990.

[13] A. O. Hero and J. A. Fessler, "A fast recursive algorithm for computing CR-type bounds for image reconstruction problems," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf.*, (Orlando, FL), Oct. 1992, pp. 1188–1190.

[14] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1285–1301, Nov. 1990.

[15] A. O. Hero, "A Cramer–Rao type lower bound for essentially unbiased parameter estimation," MIT Lincoln Laboratory, Lexington, MA, Tech. Rep. 890, Jan. 3, 1992.

Bounds on Achievable Convergence Rates of Parameter Estimators via Universal Coding

Neri Merhav

Abstract—**Lower bounds on achievable convergence rates of parameter estimators towards the true parameter are derived via universal coding considerations. It is shown that for a parametric class of finite-alphabet information sources, if there exists a universal lossless code whose redundancy decays sufficiently rapidly, then it induces a limitation on the fastest achievable convergence rate of any parameter estimator, at any value of the true parameter, with a possible exception of a vanishingly small subset of parameter values. A specific choice of a universal**

Manuscript received December 1, 1992; revised December 2, 1993. This paper was presented in part at the 1994 IEEE International Symposium on Information Theory, January 1994.

The author is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9403843.

code yields a slightly different version of this result which extends easily to the continuous case.

***Index Terms*—Parameter estimation, Bayesian estimation, lower bounds, universal lossless coding.**

I. Introduction

There are essentially three approaches to the formulation of lower bounds in estimation theory. The *Bayesian* approach treats the unknown parameter as a random variable with a given prior probability density function (pdf) and sets a lower bound on the mean-square error (MSE) averaged with respect to (w.r.t.) this pdf. Well-known Bayesian bounds are those of Van Trees [1], Bhattacharyya [2], Bobrovsky and Zakai [3], Bellini and Tartara [4], Chazan *et al.* [5], and Weiss and Weinstein [6]. The fundamental weakness of the Bayesian approach is that averaging over the parameter space precludes the possibility of providing any *local* (or *pointwise*) information on achievable estimation accuracy at a certain point of the true parameter.

In the *non-Bayesian* approach, on the other hand, one sets a local bound on the MSE at any given value of the true parameter. Under this category, we find the lower bounds of Cramer and Rao [7]–[12], Bhattacharyya [2], Chapman and Robbins [13], Fraser and Guttman [14], Barankin [15], and Kiefer [16]. The main drawback of this class of bounds is that they are normally subject to certain limitations on the class of permissible estimators, in particular, the class of unbiased estimators. This restriction is posed primarily to eliminate uninteresting trivialities like an estimator that is set to a fixed parameter value θ_0, independently of the observations. This is definitely a very poor estimator, but it yields perfect estimation when the true parameter is indeed θ_0, and hence when included in the class of allowed estimators, no nontrivial lower bound can hold simultaneously for *any* estimator at *every* point. It should be pointed out that some of the above-mentioned non-Bayesian bounds have extensions that include possibly biased estimators, but these extended bounds depend on the bias function which, in turn, depends on the particular estimator being selected.

One way to remove trivialities such as the above, without restricting the class of estimators and still maintaining locality of the bounds, is suggested by the third approach, that is, the *minimax* approach (see, e.g., LeCam [17], Huber [18], Hájek [19], Ibragimov and Khas'minsky [20], Nemirovsky [21], and Nazin [22]). Here, one first derives a lower bound to the asymptotic estimation error associated with the *worst* parameter value within a neighborhood of radius $\delta > 0$ around a given point θ, and then let δ vanish. In other words, here we find lower bounds on quantities like

$$\lim_{\delta \to 0} \liminf_{n \to \infty} \sup_{|\theta' - \theta| < \delta} E_{\theta'}\{l[n^{1/2}(\hat{\theta} - \theta')]\}$$

where $l(\cdot)$ is a given loss function. The idea is that, on one hand, for a given $\delta > 0$, the bound is still not local (just like a Bayesian bound), and hence rules out the poor performance of bad estimators like the above trivial estimator $\hat{\theta} = \theta_0$. But, on the other hand, in the second step of the asymptotics, as the δ-neighborhood shrinks, the bound becomes local, and hence depends solely on θ. It should be noted, however, that the regularity conditions under which the minimax bounds are valid and attainable are quite demanding.

In this correspondence, we point out a possible alternative approach to a non-Bayesian setting. Consider again the above trivial estimator $\hat{\theta} = \theta_0$, regardless of the observed data, and note that this estimator "performs well" only when the true parameter is in a vanishingly small neighborhood of θ_0. This suggests a formulation of a lower bound on the achievable estimation error that is valid for any estimator at any point, with a possible exception of points in a vanishingly small subset of the parameter space.

In [23], a lower bound in this spirit has been developed in the context of universal data compression for parametric information sources. Specifically, it has been shown in [23, Theorem 1, part a] that under some regularity conditions on the parametric family of sources $\{p_\theta,\ \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$, and for every given uniquely decipherable lossless coding scheme operating on input strings of length n, the compression ratio cannot approach the entropy of the source p_θ uniformly faster than $0.5kn^{-1} \log n$, for all points $\theta \in \Theta$, except for a set of points whose volume (Lebesgue measure) vanishes as $n \to \infty$. Interestingly, the proof in [23] is based on the assumption that $\{p_\theta,\ \theta \in \Theta\}$ is such that there exists an $\sqrt{n}$-consistent estimator $\hat{\theta}$ of θ, i.e., an estimator for which the estimation error decays as fast as $n^{-1/2}$. At first glance, it might seem surprising that the existence of a *good* estimator draws a *limitation* on the achievable performance of universal codes w.r.t. the same class of information sources. The intuition behind this tradeoff between estimation performance and universal coding redundancy is, however, fairly simple: if a good estimator exists, this means that, typically, the likelihood function is sufficiently sensitive to small perturbations in the parameter value. Because of this sensitivity, even small encoding (quantization) errors in the parameter estimate might yield a considerable loss in the compression ratio, and hence relatively many bits should be allocated to encode the estimate accurately. Conversely, if the likelihood function is relatively insensitive to θ, this is an obstacle for estimation, but advantageous for coding. As an extreme example, consider the case where p_θ is completely *independent* of θ. Here, θ is not estimable at all, but obviously, there is a coding scheme that is optimal in the sense of uniformly attaining the entropy for every θ.

The natural question that arises now is: Does this interesting phenomenon work in the other way as well, namely, does the existence of a good universal code for the class of sources $\{p_\theta,\ \theta \in \Theta\}$ induce a limitation on the estimation accuracy of θ? As we show here, the answer is yes. The primary purpose of this work is not necessarily to present a new powerful technique for deriving a tight lower bound on the estimation error, but rather to point out and to explore this other direction of tradeoff between universal coding redundancy and order of convergence of estimators. Specifically, we show a simple relation between the coding redundancy of the best universal code and the achievable estimation precision. Similarly to [23], the lower bound on the convergence rate of the estimation error is stated locally in the parameter space with a possible exclusion of an exception set that shrinks as the sample size grows. The proof of this result is dual to that of [23]. A specific choice of a universal code leads to a more explicit, and in some sense a stronger, statement which has a straightforward generalization from the finite-alphabet to the continuous-alphabet case.

It should be stressed, however, that this result provides information merely on the best attainable *order* of the convergence rate, without specifying the leading constant, and further work is still needed in this direction. Nevertheless, the required regularity conditions of our results are considerably weaker than those of the local minimax bounds [17]–[22].

We also demonstrate that a slight refinement in the analysis may lead to explicit lower bounds on any moment of the estimation error (rather than just asymptotic convergence rates), at the

expense of a nonshrinking exception set. In this case, there is a tradeoff between the level of the lower bound on the estimation error and the upper bound on the volume of the exception set. The behavior of this tradeoff, although not necessarily tight, seems sharper than the one that can be explored from a simple Bayesian argument. Several examples of specific estimation problems are provided.

II. Results

Let $x^n = (x_1, x_2, \cdots, x_n)$ be a vector of observations, where each component x_i takes on values in a finite set X whose size is $|X| = X$. The set of all n-dimensional observation vectors will be denoted X^n. It is assumed that x^n is drawn from a probabilistic information source with a probability mass function (PMF) $p_\theta(x^n)$, indexed by a parameter θ that takes on values in a set $\Theta \subseteq \mathbb{R}^k$, referred to as the *parameter space*. An *estimator* $\hat{\theta}_n = f_n(x^n)$ for θ is a measurable map, independent of θ, from X^n to Θ. A *length function* $L_n(x^n)$ of a lossless block code is a map from X^n to the positive reals, where $L_n(x^n)$ represents the codeword length (in nats) for a source vector x^n. We shall require the length function to be induced from a *uniquely decipherable* code (see, e.g., [24, ch. 10]), i.e., to satisfy the Kraft–McMillan inequality

$$\sum_{x^n \in X^n} e^{-L_n(x^n)} \le 1. \tag{1}$$

The nth-order normalized entropy is $H_n(\theta) \triangleq -n^{-1} E_\theta \ln p_\theta(x^n)$, where $E_\theta\{\cdot\}$ denotes expectation w.r.t. p_θ, and is well known to be a lower bound on the compression ratio $n^{-1}E_\theta L_n(x^n)$ associated with any uniquely decipherable lossless code operating on blocks of length n. The difference $R_n(L_n, \theta) = n^{-1}E_\theta L_n(x^n) - H_n(\theta)$ is referred to as the *redundancy* of the code $L_n(\cdot)$ at θ. A sequence of uniquely decipherable lossless codes is called *strongly minimax universal* w.r.t. the class of sources $\{p_\theta, \theta \in \Theta\}$ [25] if $\lim_{n\to\infty} R_n(L_n, \theta) = 0$ uniformly over $\theta \in \Theta$, namely, if $\lim_{n\to\infty} \sup_{\theta\in\Theta} R_n(L_n, \theta) = 0$.

The following theorem relates the existence of a sequence of strongly minimax universal codes w.r.t. $\{p_\theta, \theta \in \Theta\}$ to the achievable convergence rate of any estimator $\hat{\theta}_n = f_n(x^n)$, namely, to the stochastic order of the estimation error $\|\hat{\theta}_n - \theta\|$, where $\|\cdot\|$ denotes the Euclidean norm in the k-dimensional parameter space.

Theorem 1: Let $L_n(\cdot)$ be a length function of a uniquely decipherable lossless code such that $n \cdot R_n(L_n, \theta) \le k \cdot \mu_n$ for all large n, every $\theta \in \Theta$, and some monotone sequence $\{\mu_n\}_{n\ge 1}$ satisfying $\mu_n/n \to 0$. Let $\{\lambda_n\}_{n\ge 1}$ be any monotonically nondecreasing positive sequence such that $\log \lambda_n \ge \zeta \mu_n$ for some $\zeta > 1$ and all large n. Then, for any given estimator $\hat{\theta}_n = f_n(x^n)$, every positive constant C, all $0 < \epsilon \le \epsilon(\zeta)$, and all large n,

$$\Pr\{x^n : \|\hat{\theta}_n - \theta\| > C|\theta\} \ge \epsilon \cdot \frac{\mu_n}{n} \tag{2}$$

for all points $\theta \in \Theta$, except for points in a set $A_\epsilon(n) \subseteq \Theta$ whose volume tends to zero as $n \to \infty$.

The proof appears in the Appendix.

The sublinear growth rate of μ_n implies that the redundancy tends to zero as $n \to \infty$, as one would expect if $L_n(\cdot)$ is a strongly minimax universal code. Thus, roughly speaking, the theorem tells us that if one can find a universal code whose redundancy decays uniformly as fast as $k \cdot \mu_n/n$, then no estimator can converge to the true parameter at a rate that is essentially faster than $e^{-\mu_n}$, in the sense that the estimation error might exceed a threshold that decays slightly faster than $C \cdot e^{-\mu_n}$, with a probability that does not vanish too rapidly. This is true for all θ except for points in a vanishingly small set $A_\epsilon(n)$. Normally, $\mu_n = 0.5 \log n$ (see, e.g., [23]), which implies that a convergence rate faster than $n^{-1/2}$ cannot be expected, i.e., $\sqrt{n}$-consistency is usually the best one can hope for. However, this is not always the case. The advantage of the above theorem is that it holds for any universal code. The disadvantage, besides the fact that it is limited to the finite-alphabet case, is that $e^{-\mu_n}$ is the fastest achievable convergence rate in a rather weak sense because the probability in (2) is not guaranteed to converge to unity, but merely not to vanish very quickly.

It turns out that by selecting a specific length function $L_n(\cdot)$, this problem can be resolved. Specifically, consider the function

$$L_n(x^n) = -\ln \sup_{\theta\in\Theta} p_\theta(x^n) + k \cdot \mu_n \tag{3}$$

where μ_n is chosen independently of x^n such that the Kraft inequality (1) will be satisfied. It is readily seen that this choice of $L_n(\cdot)$ satisfies the inequality $n \cdot R_n(L_n, \theta) \le k \cdot \mu_n$ as well. This results in the following form of the above theorem.

Theorem 2: Suppose that $\mu_n \triangleq k^{-1} \ln[\sum_{x^n} \sup_\theta p_\theta(x^n)] < \infty$. Then, for every monotonically nondecreasing sequence $\{\lambda_n\}_{n\ge 1}$ such that $e^{-\mu_n}\lambda_n \to \infty$, every estimator $\hat{\theta}_n = f_n(x^n)$, every $C > 0$, and every $0 < \epsilon < 1$,

$$\Pr\{x^n : \lambda_n\|\hat{\theta}_n - \theta\| > C|\theta\} > \epsilon \tag{4}$$

for all $\theta \in \Theta$, except for points in a set $A_\epsilon(n) \subseteq \Theta$ whose volume tends to zero as $n \to \infty$.

The proof appears in the Appendix.

This is a more explicit formulation than the first theorem, and it has the following advantages. First, it avoids the need for guessing an efficient universal coding scheme. Second, it removes the above-mentioned weakness of (2). Finally, it has a straightforward extension to the continuous case: the same theorem holds true if p_θ is a considered a pdf and the summation in the definition of μ_n is replaced by integration. It is somewhat weaker than Theorem 1 in the sense that μ_n (and hence λ_n) induced by the specific choice made in (3) might not be the best one can find when the choice of $L_n(\cdot)$ is free as in Theorem 1.

It should be pointed out that the requirement $\mu_n < \infty$ appears crucial when the parameter space Θ is not bounded. In many cases of practical interest with an unbounded Θ, μ_n turns out to be infinite (see examples in the next section) which, strictly speaking, makes Theorem 2 meaningless. However, by looking at any bounded subset Θ_0 of Θ, we may again apply Theorem 2 with a finite μ_n and conclude that the volume of $A_\epsilon(n) \cap \Theta_0$ decays with n. In other words, although the overall volume of $A_\epsilon(n)$ may not vanish (or even be infinite) when Θ is unbounded, it appears that $A_\epsilon(n)$ exhibits a *sparseness* property which, generally speaking, means that in any bounded subregion of Θ, there is a relatively small "percentage" of exceptional points θ in the sense of violating the theorem. This point will be elaborated on in the next section.

Finally, it should be pointed out that as an alternative to the length function (3), another reasonable choice would be $L_n(x^n) = -\ln p(x^n)$ where $p(x^n) = \int_\Theta d\theta\, \pi(\theta) p_\theta(x^n)$, $\pi(\theta)$ being some prior on θ whose support is Θ. In this case, the rate sequence is given by $\mu_n = k^{-1} \sup_\theta E_\theta \ln[p_\theta(x^n)/p(x^n)] \triangleq k^{-1} \sup_\theta D(p_\theta \| p)$.

The above theorems only tell us what are the achievable convergence rates of estimators, but they do not provide explicit lower bounds on the estimation error. It turns out, however, that a slightly sharper analysis in the proof of Theorem 2 may lead to

particular bounds on any moment of the estimation error, at the expense that, now, the volume of the exception set will be bounded by a *constant*, and hence not guaranteed to vanish as n grows. The following corollary of Theorem 2 (proved in the Appendix) exhibits a tradeoff between the level of the bound and the volume of the exception set where this bound does not necessarily hold. Again, it has an immediate extension to the continuous-alphabet case.

Corollary 1: Let $\{p_\theta,\ \theta \in \Theta\}$ be such that $\lambda_n \triangleq [\sum_{x^n \in \mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x^n)]^{1/k} < \infty$. Then, for every n, every estimator $\hat{\theta}_n$, every $s > 0$, and every $B > 0$,

$$E_\theta \|\hat{\theta}_n - \theta\|^s \geq \frac{B^s}{\lambda_n^s} \tag{5}$$

for all points θ, except for points in a set $A_n(B)$ whose volume, for every n, is bounded by $\mathrm{Vol}\{A_n(B)\} \leq J(k,s) \cdot B^k$, where

$$J(k,s) \triangleq \frac{2 \cdot (4\pi)^{k/2}}{\Gamma(k/2)} \cdot \left(\frac{1}{k} + \frac{1}{s}\right) \cdot \left(1 + \frac{s}{k}\right)^{k/s}. \tag{6}$$

This formulation suggests that, in order to evaluate the performance of a given estimator $\hat{\theta}_n = f_n(x^n)$ at different points θ, one should subdivide Θ in accordance with level sets of the estimation error, i.e., $A_n(B, f_n) = \{\theta : \lambda_n^s E_\theta \|f_n(x^n) - \theta\|^s \leq B^s\}$, where B takes on values from zero to some upper limit $B_{\max}$, and a good estimator is one for which $\mathrm{Vol}\{A_n(B, f_n)\}$ is as large as possible for *every* $0 \leq B \leq B_{\max}$. In the next section, we will examine the maximum likelihood (ML) estimator of the variance of a Gaussian random variable from this point of view, and compare it to the upper bound on the volume of the exception set as given in Corollary 1.

It should be pointed out that tradeoffs between error levels and volumes of exception sets also can be explored using simple Bayesian arguments, but it seems that the resulting bounds are weaker than the above, at least for some values of B. Specifically, let Θ be a bounded set, and suppose, temporarily, that θ takes on values in Θ under a uniform pdf. Let $\hat{\theta}_n = E(\theta|x^n)$ be the optimal Bayesian estimator in the sense of minimizing $\int_\Theta d\theta\, E_\theta \|\hat{\theta}_n - \theta\|^2 \triangleq D^2 \cdot \mathrm{Vol}(\Theta)/\lambda_n^2$ (i.e., $s = 2$). We shall assume that D is a constant, which is, in fact, the case in many situations. Let $\hat{\theta}_n$ be any competing estimator, and let $A_n(B) = \{\theta : E_\theta \|\hat{\theta}_n - \theta\|^2 \leq B^2/\lambda_n^2\}$. Then

$$\begin{aligned} D^2 \cdot \mathrm{Vol}(\Theta) &\leq \lambda_n^2 \int_\Theta d\theta\, E_\theta \|\hat{\theta}_n - \theta\|^2 \\ &\leq B^2 \cdot \mathrm{Vol}\{A_n(B)\} + \lambda_n^2 \int_{\theta \notin A_n(B)} d\theta\, E_\theta \|\hat{\theta}_n - \theta\|^2. \end{aligned} \tag{7}$$

To further overbound the last expression, let us assume that $\hat{\theta}_n$ is such that for every θ, $E_\theta \|\hat{\theta}_n - \theta\|^2 \leq G^2/\lambda_n^2$ for some constant $G > B$. Then, (8) becomes

$$\begin{aligned} D^2 \cdot \mathrm{Vol}(\Theta) \leq B^2 \cdot \mathrm{Vol}\{A_n(B)\} + G^2 \\ \cdot [\mathrm{Vol}(\Theta) - \mathrm{Vol}\{A_n(B)\}], \end{aligned} \tag{8}$$

implying that

$$\mathrm{Vol}\{A_n(B)\} \leq \frac{G^2 - D^2}{G^2 - B^2} \cdot \mathrm{Vol}(\Theta). \tag{9}$$

This upper bound on the volume of the exception set is weaker than that of Corollary 1, first because we have limited ourselves to estimators with MSE uniformly less than G^2/λ_n^2, and second, the behavior of this bound, at least at extreme values of B, cannot reflect the behavior of any existing estimator. In fact, for $B \to 0$, the right-hand side of (10) tends to a constant, while the upper bound in (6) vanishes, as expected. Also, for $B \to G$, a singularity point of (10) is approached, and it becomes useless. Therefore, the technique presented in Corollary 1 appears more powerful than the Bayesian method, at least in the above-described simple form.

III. Examples

We now demonstrate the results of the previous section for several particular models.

Let $x_1, \cdots, x_n$ be independent copies of a zero-mean Gaussian random variable with variance $\theta = \sigma^2 \in [a, b] \triangleq \Theta$ to be estimated. It is easy to check that if $a > 0$ and $b < \infty$, then λ_n grows proportionally to $\sqrt{n}$, as expected. However, if either $a = 0$ or $b = \infty$, then μ_n, and hence also λ_n, is infinite. This means that there is no nontrivial limitation on the convergence rate of estimators outside a vanishingly small exception set. We next demonstrate that when $\Theta = [1 - \epsilon, \infty)$ one can indeed construct an estimator whose convergence rate is arbitrarily fast in a subset $A_\epsilon(n)$ of Θ whose Lebesgue measure is infinite. The following example is in the same spirit as that of [27, p. 405, Ex. 1.1]. Let $g(x^n) = n^{-1}\sum_{i=1}^n x_i^2$ denote the empirical variance, and define an estimator $\hat{\theta}_n$ as follows. If $|g(x^n) - j| \leq n^{-1/4}$ for some positive integer j, then set $\hat{\theta}_n = j$; otherwise, $\hat{\theta}_n = g(x^n)$. Now, suppose that θ belongs to the interval $I_j = [j - \xi_n, j + \xi_n]$ for some integer j, where $\{\xi_n\}_{n \geq 1}$ is an arbitrarily rapidly vanishing sequence. Since $|g(x^n) - \theta|$ decays at a rate $n^{-1/2}$, then it can be shown that for any $\theta \in I_j$, the probability that $|g(x^n) - j| \leq n^{-1/4}$ is very high for large n provided that ξ_n decays faster than $n^{-1/4}$. Thus, for every $\theta \in I_j$, the estimation error associated with $\hat{\theta}_n$ is less than ξ_n with high probability. It follows that, by taking a sufficiently fast decaying sequence $\{\xi_n\}_{n \geq 1}$, one can arbitrarily accelerate the convergence rate of $\hat{\theta}_n$ for every θ in $A_\epsilon(n) \triangleq \bigcup_{j=1}^\infty I_j$, which, in turn, has an infinite Lebesgue measure for all n. Note, also, that, here, $A_\epsilon(n)$ exhibits the sparseness property that has been discussed in the previous section.

Consider next the class of lth-order Markov sources with alphabet of size X. Here, the parameter vector θ consists of the $k = X^l(X - 1)$ transition probabilities from states defined by strings of length l to the next letter. It is well known (see, e.g., [26]) that there exists a universal code whose redundancy is uniformly less than $X^l(X - 1)\log n/(2n)$ up to higher order terms. This implies that λ_n can be chosen as any sequence that grows slightly faster than $n^{1/2}$. In other words, here, $n^{-1/2}$ is the fastest achievable convergence rate in the sense of Theorem 2. This convergence rate is typical in many problems of practical interest.

Note: To demonstrate Corollary 1, consider the subclass of Bernoulli sources, where θ denotes $\Pr\{x_i = 1\} = 1 - \Pr\{x_i = 0\}$. It is easy to verify (using Stirling's formula) that $\lambda_n \leq \sqrt{\pi n/2}$. Thus, Corollary 1 (with $K = 1$ and $s = 2$) tells us that $E_\theta(\hat{\theta} - \theta)^2 \geq 2B^2/\pi n$ for all θ outside a set whose Lebesgue measure does not exceed $6\sqrt{3}B$. On the other hand, the estimation of θ by the relative frequency of "1"'s results in $E_\theta(\hat{\theta} - \theta)^2 = \theta(1 - \theta)/n$, which is smaller than $2B^2/\pi n$ along two subintervals whose total length is $1 - \sqrt{1 - 8B^2/\pi}$, which in turn is considerably smaller than $6\sqrt{3}B$. (We believe that this gap should be attributed primarily to the fact that further work is needed to tighten the bound, and not to the suboptimality of the empirical relative frequency as an estimator of θ.)

ACKNOWLEDGMENT

Stimulating discussions with M. Feder are greatly appreciated.

APPENDIX

Proof of Theorem 1: The proof is, in some sense, dual to that of [23, Theorem 1, part a)]. For a given θ and a given estimator $\hat{\theta}_n$, let

$$X_n(\theta) \triangleq \{x^n : \lambda_n \|\hat{\theta}_n - \theta\| \le C\} \tag{A.1}$$

$$Q_n(\theta) \triangleq \sum_{x^n \in X_n(\theta)} e^{-L_n(x^n)} \tag{A.2}$$

and

$$P_n(\theta) \triangleq \sum_{x^n \in X_n(\theta)} p_\theta(x^n). \tag{A.3}$$

Now, observe that

$$\begin{aligned} n \cdot R_n(L_n, \theta) &= E_\theta \ln\left[\frac{p_\theta(x^n)}{e^{-L_n(x^n)}}\right] \\ &\ge \sum_{x^n \in X_n(\theta)} p_\theta(x^n) \ln \frac{p_\theta(x^n)}{e^{-L_n(x^n)}} \\ &\quad + \sum_{x^n \in X_n^c(\theta)} p_\theta(x^n) \ln p_\theta(x^n) \\ &\ge P_n(\theta) \ln \frac{P_n(\theta)}{Q_n(\theta)} - \Delta_\theta \end{aligned} \tag{A.4}$$

where $\Delta_\theta \triangleq -\sum_{x^n \in X_n^c(\theta)} p_\theta(x^n) \ln p_\theta(x^n)$ and the last inequality follows from Jensen's inequality and the nonnegativity of the Kullback–Leibler informational divergence between two probability measures (see also [23]). Let $A_\epsilon(n)$ denote the set of points θ for which $P_n(\theta) \ge 1 - \epsilon\mu_n/n$. For $\theta \in A_\epsilon(n)$, we can overestimate Δ_θ as follows. Since $\mu_n/n \to 0$, then for all large n,

$$\begin{aligned} \Delta_\theta &= [1 - P_n(\theta)] \sum_{x^n \in X_n^c(\theta)} \frac{p_\theta(x^n)}{1 - P_n(\theta)} \ln\left[\frac{1 - P_n(\theta)}{p_\theta(x^n)}\right] \\ &\quad + [1 - P_n(\theta)] \ln \frac{1}{1 - P_n(\theta)} \\ &\le [1 - P_n(\theta)] \cdot \ln\left\{\sum_{x^n \in X_n^c(\theta)} \frac{p_\theta(x^n)}{[1 - P_n(\theta)]} \cdot \frac{[1 - P_n(\theta)]}{p_\theta(x^n)}\right\} \\ &\quad + \epsilon\frac{\mu_n}{n} \ln \frac{n}{\epsilon\mu_n} \\ &\le \epsilon\frac{\mu_n}{n} \ln |X_n^c(\theta)| + \epsilon\frac{\mu_n}{n} \ln \frac{n}{\epsilon\mu_n} \\ &\le \epsilon\frac{\mu_n}{n} \ln X^n + \epsilon\frac{\mu_n}{n} \ln \frac{n}{\epsilon\mu_n} \\ &\le 2\mu_n \epsilon \ln X \triangleq \epsilon_1 k \mu_n. \end{aligned} \tag{A.5}$$

Thus, combining (A.4) and (A.5), we find that for every $\theta \in A_\epsilon(n)$,

$$\begin{aligned} k \cdot \mu_n &\ge n \cdot R_n(L_n, \theta) \\ &\ge P_n(\theta) \ln \frac{P_n(\theta)}{Q_n(\theta)} - \epsilon_1 k \mu_n \\ &\ge \left(1 - \epsilon\frac{\mu_n}{n}\right) \cdot \ln\left[\frac{1 - \epsilon\mu_n/n}{Q_n(\theta)}\right] - \epsilon_1 k \mu_n, \end{aligned} \tag{A.6}$$

which implies that

$$Q_n(\theta) \ge \left(1 - \epsilon\frac{\mu_n}{n}\right) \cdot \exp\left\{-\frac{k(1 + \epsilon_1)\mu_n}{1 - \epsilon\mu_n/n}\right\}, \qquad \forall \theta \in A_\epsilon(n). \tag{A.7}$$

Let N_n denote the maximum number of *disjoint* spheres $S_n(\theta) \triangleq \{\theta' : \|\theta' - \theta\| \le C/\lambda_n\}$ with centers at $A_\epsilon(n)$, and let C_n denote the center set. By the Kraft inequality and by (A.8), we have

$$1 \ge \sum_{\theta \in C_n} Q_n(\theta) \ge N_n \cdot \left(1 - \epsilon\frac{\mu_n}{n}\right) \cdot \exp\left\{-\frac{k(1 + \epsilon_1)\mu_n}{1 - \epsilon\mu_n/n}\right\}, \tag{A.8}$$

which implies that

$$N_n \le \left(1 - \epsilon\frac{\mu_n}{n}\right)^{-1} \exp\left\{\frac{k(1 + \epsilon_1)\mu_n}{1 - \epsilon\mu_n/n}\right\}. \tag{A.9}$$

Finally, note that by doubling the radius of each sphere $S_n(\theta)$, we get a cover of $A_\epsilon(n)$ (see also [23]). Thus, the volume of $A_\epsilon(n)$ is overbounded as follows:

$$\begin{aligned} \mathrm{Vol}\{A_\epsilon(n)\} &\le N_n \cdot \mathrm{Vol}\left\{\theta' : \|\theta' - \theta\| \le \frac{2C}{\lambda_n}\right\} \\ &\le \left(1 - \epsilon\frac{\mu_n}{n}\right)^{-1} \exp\left\{\frac{k(1 + \epsilon_1)\mu_n}{1 - \epsilon\mu_n/n}\right\} \cdot V_k\left[\frac{2C}{\lambda_n}\right]^k \\ &= V_k(2C)^k \cdot \left(1 - \epsilon\frac{\mu_n}{n}\right)^{-1} \\ &\quad \cdot \left[\lambda_n^{-1} \exp\left\{\frac{(1 + \epsilon_1)\mu_n}{1 - \epsilon\mu_n/n}\right\}\right]^k \end{aligned} \tag{A.10}$$

where $V_k = 2\pi^{k/2}/(k\Gamma(k/2))$ is the volume of the k-dimensional unit sphere. The rightmost side of (A.11) tends to zero provided that λ_n grows faster than $e^{\zeta\mu_n}$ for, say, $\zeta = 1 + 2\epsilon_1$. This completes the proof of Theorem 1.

Proof of Theorem 2: The proof is very similar to that of Theorem 1, although it appears significantly simpler due to the specific choice of a length function. It also generalizes easily to the continuous alphabet case, just by replacing summations with integrals. Let $X_n(\theta)$ and $P_n(\theta)$ be defined as in the proof of Theorem 1, and let us redefine $Q_n(\theta)$ as

$$Q_n(\theta) \triangleq e^{-k\mu_n} \sum_{x \in X_n(\theta)} \sup_{\theta \in \Theta} p_\theta(x^n). \tag{A.11}$$

Let $A_\epsilon(n)$ be the set of points θ such that $P_n(\theta) \ge 1 - \epsilon$. Let N_n and C_n be as in the proof of Theorem 1 [but w.r.t. the present definition of $A_\epsilon(n)$]. Then, by the definition of μ_n,

$$1 \ge \sum_{\theta \in C_n} Q_n(\theta) \ge e^{-k\mu_n} \sum_{\theta \in C_n} P_n(\theta) \ge e^{-k\mu_n} N_n \cdot (1 - \epsilon), \tag{A.12}$$

implying that $N_n \le (1 - \epsilon)^{-1} e^{k\mu_n}$. Now, similarly to (A.10),

$$\mathrm{Vol}\{A_\epsilon(n)\} \le V_k(2C)^k (1 - \epsilon)^{-1} \left[\frac{e^{\mu_n}}{\lambda_n}\right]^k, \tag{A.13}$$

which again vanishes if λ_n grows faster than e^{μ_n}.

Proof of Corollary 1: Let μ_n be as in Theorem 2, and let $\lambda_n = e^{\mu_n}$, namely, λ_n is as defined in Corollary 1. By repeating

the same steps as in the proof of Theorem 2, and by Markov's inequality, we find that for every nonexceptional θ,

$$\epsilon \le \Pr\left\{\|\hat{\theta}_n - \theta\| > \frac{C}{\lambda_n}\middle|\theta\right\} \le \frac{\lambda_n^s}{C^s}\cdot E_\theta\|\hat{\theta}_n - \theta\|^s \quad \text{(A.14)}$$

or, equivalently, $E_\theta\|\hat{\theta}_n - \theta\|^s \ge C^s\epsilon/\lambda_n^s$, which agrees with (5) if C and ϵ are chosen such that $C^s\epsilon = B^s$. On the other hand, the volume of the exception set [now denoted $A_n(B)$] when $\lambda_n = e^{\mu_n}$ is overbounded similarly to (A.13) by $\mathrm{Vol}\{A_n(B)\} \le V_k 2^k \cdot C^k/(1-\epsilon)$. By minimizing the latter expression subject to the constraint $C^s\epsilon = B^s$, the desired result is obtained.

References

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[2] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation," *Sankhya*, vol. 8, pp. 1–14, 201–208, 315–328, 1946.

[3] B. Bobrovsky and M. Zakai, "A lower bound on the estimation error for certain diffusion processes," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 45–52, 1976.

[4] S. Bellini and G. Tartara, "Bounds on errors in signal parameter estimation," *IEEE Trans. Commun.*, pp. 340–342, 1974.

[5] D. Chazan, M. Zakai, and J. Ziv, "Improved lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 90–93, Jan. 1975.

[6] A. J. Weiss and E. Weinstein, "Lower bounds on the mean square error in random parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 680–682, Sept. 1985.

[7] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Phil. Trans. Roy. Soc. London*, vol. 222, p. 309, 1922.

[8] D. Dugue, "Application des properties de la limite au sens du calcul des probabilities a l'etude des diverses questions d'estimation," *Ecol. Poly.*, vol. 3, no. 4, pp. 305–372, 1937.

[9] M. Frechet, "Sur L'extension de certaines evaluations statistiques au cas de petits echantillons," *Rev. Inst. Int. Statist.*, vol. 11, pp. 182–205, 1943.

[10] G. Darmois, "Sur les limites de la dispersion de certains estimations," *Rev. Inst. Int. Statist.*, vo. 13, pp. 9–15, 1945.

[11] C. R. Rao, "Information accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81–91, 1945.

[12] H. Cramer, *Mathematical Methods in Statistics*. Princeton, NJ: Princeton Univ. Press, 1946.

[13] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumption," *Ann. Math. Statist.*, vol. 22, pp. 581–586, 1951.

[14] D. A. Fraser and I. Guttman, "Bhattacharyya bound without regularity assumptions," *Ann. Math. Statist.*, vol. 23, pp. 629–632, 1952.

[15] E. W. Barankin, "Locally best unbiased estimators," *Ann. Math. Statist.*, vol. 20, pp. 477–501, 1949.

[16] J. Kiefer, "On minimum variances estimation," *Ann. Math. Statist.*, vol. 23, pp. 627–629, 1952.

[17] L. LeCam, "On some asymptotic properties of maximum likelihood estimates and related Bayes estimates," *Univ. California Publ. Statist.*, vol. 1, pp. 277–330, 1953.

[18] P. Huber, "Strict efficiency excludes superefficiency," *Ann. Math. Statist.*, vol. 37, p. 1425 (abstract), 1966.

[19] J. Hájek, "Local asymptotic minimax and admissibility in estimation," in *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, 1972, pp. 175–194.

[20] I. A. Ibragimov and R. Z. Khas'minsky, *Statistical Estimation: Asymptotic Theory*. Berlin, Germany: Springer, 1981.

[21] A. S. Nemirovsky, "Optimization of recursive algorithms of estimation of parameters of linear plants," *Automation Remote Contr.*, vol. 42, no. 6, pp. 775–783, 1981.

[22] A. Nazin, "On minimax bound for parameter estimation in ball (bias accounting)," in V. Sazonov and T. Shervashidze, Eds., *New Trends in Probability and Statistics*. VSP/Moksals, 1991.

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

[24] R. J. McEliece, *The Theory of Information and Coding*. Cambridge, England: Cambridge Univ. Press, 1984.

[25] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[26] ——, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211–215, Mar. 1983.

[27] E. L. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.

Bounds on Approximate Steepest Descent for Likelihood Maximization in Exponential Families

Nicolò Cesa-Bianchi, Anders Krogh, and Manfred K. Warmuth

***Abstract*—An approximate steepest descent strategy converging, in families of regular exponential densities, to maximum likelihood estimates of density functions is described. These density estimates are also obtained by an application of the principle of minimum relative entropy subject to empirical constraints. We prove tight bounds on the increase of the log-likelihood at each iteration of our strategy for families of exponential densities whose log-densities are spanned by a set of bounded basis functions.**

***Index Terms*—Exponential families, minimum relative entropy estimation, steepest descent.**

I. Introduction

Consider the following problem: Given a random sample $x_1,\cdots,x_m$ drawn independently from a distribution P with density p, find the maximum likelihood estimate in a family of regular exponential densities. This problem of density estimation is also known as minimization of relative entropy (Kullback–Leibler divergence) subject to empirical constraints (see, e.g., [1], [2]). In this work we describe an approximate steepest descent strategy[1] converging to the MLE in exponential families of densities whose log-densities are linear combinations of a set of bounded basis functions. We show tight lower and upper bounds on the increase of the log-likelihood function (or, equivalently, decrease of the relative entropy) at each iteration, as a function of the norm of the gradient.

Let $(X, \mathcal{B})$ be a measurable space. In the following, all densities on $(X, \mathcal{B})$ are understood with respect to a finite dominating measure ν. We recall the definition of the relative

Manuscript received October 10, 1992; revised October 28, 1993. This research was done while N. Cesa-Bianchi and A. Krogh were visiting the University of California-Santa Cruz. N. Cesa-Bianchi was partially supported by the "Progetto finalizzato sistemi informatici e calcolo parallelo" of CNR under Grant 91.0884.69.115.09672. A. Krogh was supported by ONR Grant N00014-91-J-1162 and a grant from the Danish Natural Science Research Council. M. K. Warmuth was supported by ONR Grant N00014-91-J-1162.

N. Cesa-Bianchi is with DSI, Università di Milano, Via Comelico 39, I-20135 Milano, Italy.

A. Krogh is with CONNECT, Electronics Institute, Build 349, Technical University of Denmark, 2800 Lyngby, Denmark.

M. K. Warmuth is with the Computer Science Department, University of California, Santa Cruz, Ca 95064.

IEEE Log Number 9402576.

[1] The strategy was originally introduced in [6] as an iterative method for the solution of sparse systems of linear equations.