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 $(X, \mathscr{B})$ be a measurable space. In the following, all densities on $(X, \mathscr{B})$ 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.

entropy (Kullback–Leibler divergence) $D(p\|p')$ between two densities p and p' on $(X, \mathscr{B})$:

$$D(p\|p') = \int_X p \ln \frac{p}{p'}.$$

Choose a positive integer d and let $\Phi = \{\phi_1, \phi_2, \cdots, \phi_d\}$ be a set of bounded *basis functions* ϕ_k: $X \to \mathbb{R}$. Fix also a *reference density* q^0 on $(X, \mathscr{B})$.

We will use the notation $\theta \cdot \phi(x)$ for the inner product $\Sigma_k \theta_k \phi_k(x)$. We now define the *regular exponential family* $\mathscr{E}(\Phi) = \{q_\theta : \theta \in \mathbb{R}^d\}$ of densities $q_\theta(x) = q^0(x) \exp(\theta \cdot \phi(x) - \psi(\theta))$, where the function ψ from $\mathbb{R}^d$ to $\mathbb{R}$ is defined by

$$\psi(\theta) = \ln \int_X e^{\theta \cdot \phi} q^0. \tag{1}$$

For any density p and for any $\theta \in \mathbb{R}^d$, define $\alpha(p) = (\alpha_1(p), \cdots, \alpha_d(p))$ by

$$\alpha_k(p) = E_p[\phi_k], \quad \text{for } k = 1, \cdots, d,$$

and $\alpha(\theta) = (\alpha_1(\theta), \cdots, \alpha_d(\theta))$ by

$$\alpha_k(\theta) = \alpha_k(q_\theta) = E_{q_\theta}[\phi_k], \quad \text{for } k = 1, \cdots, d.$$

If Φ is a set of linearly independent functions,[2] it is known that ψ is strictly convex (see, e.g., [3]). As a consequence, also $D(p\|q_\theta)$ is strictly convex in θ, which is seen from

$$\begin{aligned} D(p\|q_\theta) &= E_p\left[\ln \frac{1}{q_\theta}\right] - H(p) \\ &= -E_p[\theta \cdot \phi - \psi(\theta)] - E_p[\ln q^0] - H(p) \\ &= \psi(\theta) - \alpha(p) \cdot \theta + D(p\|q^0), \end{aligned} \tag{2}$$

where $H(p)$ is the entropy $E_p[-\ln p]$. Hence, if Φ is linearly independent and there exists a $\theta^* \in \mathbb{R}^d$ minimizing $D(p\|q_\theta)$, then θ^* is unique. Moreover, $\nabla D(p\|q_{\theta'}) = 0$ if and only if $\theta' = \theta^*$.

Finally, observe that for any density p and any vector $\theta \in \mathbb{R}^d$,

$$\nabla D(p\|q_\theta) = \alpha(\theta) - \alpha(p), \tag{3}$$

as can be derived from (1) and (2).

II. Description of the Strategy

We now introduce the iterative likelihood maximization strategy. Let $\|\cdot\|$ be the Euclidean norm. We assume that the strategy is parametrized with respect to the choice of the set of basis functions Φ. In order to simplify the analysis, we also restrict the range of each basis function ϕ_k $(k = 1, \cdots, d)$ in the interval $[-\sqrt{1/4d}, \sqrt{1/4d}]$. This ensures that for all nontrivial choices of the set Φ of basis functions, for any density p, and for any $x \in X$, $\|\phi(x) - \alpha(p)\| \in [0, 1)$.[3] We remark that the need for normalizing the ϕ_k's can be also interpreted via the notion of "comparison density," as pointed out by an anonymous referee.

On each run, the strategy is given as input a reference density q^0 and a random sample $x_1, \cdots, x_m$ independently drawn from a distribution P with density $p(x)$. The output consists of an infinite sequence $q^1, q^2, \cdots$ of densities in $\mathscr{E}(\Phi)$.

Let $\alpha^t = (\alpha_1^t, \cdots, \alpha_d^t)$ such that

$$\alpha_k^t = E_{q^t}[\phi_k], \quad \text{for } k = 1, \cdots, d,$$

and $\tilde{\alpha} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_d)$ such that

$$\tilde{\alpha}_k = \frac{1}{m} \sum_{i=1}^m \phi_k(x_i), \quad \text{for } k = 1, \cdots, d.$$

The sequence of densities q^t is such that, for each $t \geq 1$ and for each $x \in X$,

$$q^{t+1}(x) = q^0(x) \exp[(\theta^t + \Delta\theta^t) \cdot \phi(x) - \psi(\theta^t + \Delta\theta^t)], \tag{4}$$

where θ^t is the parameter vector after the tth iteration (assuming $\theta^0 = 0$), and $\Delta\theta^t = \theta^{t+1} - \theta^t$ is defined by

$$\Delta\theta^t = \frac{\tanh^{-1}(\|\tilde{\alpha} - \alpha^t\|)}{\|\tilde{\alpha} - \alpha^t\|} (\tilde{\alpha} - \alpha^t). \tag{5}$$

It is easily seen that, for all $t \geq 1$, q^t is in the exponential family $\mathscr{E}(\Phi)$. Notice also that $\|\tilde{\alpha} - \alpha^t\| < 1$ for all t since the ϕ_k's have been normalized.

In the next section we show that the increment (5) corresponds to *exact* steepest descent with respect to an approximation of the Kullback–Leibler divergence along the direction of the gradient.

III. Analysis

In this section we prove bounds of the increase of the log-likelihood at each iteration. The log-likelihood function for the family $\mathscr{E}(\Phi)$ is

$$l(\theta) = \ln \prod_{i=1}^m q_\theta(x_i) = \ln \prod_{i=1}^m q^0(x_i) + m(\theta \cdot \tilde{\alpha} - \psi(\theta)). \tag{6}$$

Hence, for a set Φ of linearly independent basis functions, the maximum likelihood estimate $q_{\hat{\theta}}$ in the family $\mathscr{E}(\Phi)$ is characterized by the unique $\hat{\theta} \in \mathbb{R}^d$ satisfying the equation

$$\alpha(\hat{\theta}) = \tilde{\alpha}. \tag{7}$$

Conditions guaranteeing the existence of the MLE in exponential families can be found in [4], [5].

Using (2), (6), and (7), we can rewrite the Kullback–Leibler divergence as

$$D(q_{\hat{\theta}}\|q_\theta) = \hat{\theta} \cdot \tilde{\alpha} - \psi(\hat{\theta}) + \left(\ln \prod_{i=1}^m q^0(x_i) - l(\theta)\right) \Big/ m, \tag{8}$$

where only the last term depends on θ. Therefore, the problem of maximizing the log-likelihood function is equivalent to the problem of minimizing $D(q_{\hat{\theta}}\|q_\theta)$. Note also that (3) yields $\nabla D(q_{\hat{\theta}}\|q_\theta) = \alpha(\theta) - \tilde{\alpha}$.

We will make use of the following two inequalities.

For all $k \in \mathbb{R}$ and $x \in [-1, 1]$,

$$e^{kx} \leq \frac{e^k + e^{-k}}{2} + \frac{e^k - e^{-k}}{2} x = \cosh(k) + x \sinh(k). \tag{9}$$

For all $x \in [-1, 1]$,

$$x \tanh^{-1}(x) \leq \ln \frac{1}{1 - x^2}. \tag{10}$$

For a convex function $g(x)$, it holds that $g(x) \leq g(-1) + (1 + x)(g(1) - g(-1))/2$ for $x \in [-1, 1]$. Applying this to $g(x) = e^{kx}$ yields inequality (9). Inequality (10) is proven in the Appendix.

It follows from (8) that the increase of the log-likelihood at each iteration equals $D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1})$. We now prove that this increase is upper and lower bounded within a small constant factor by a monotone increasing function of $\|\nabla D(q_{\hat{\theta}}\|q_\theta)\|$.

[2] By linear independence of the set of functions we mean that if $(\theta - \theta') \cdot \phi(x)$ is constant almost everywhere, then $\theta = \theta'$.

[3] Notice that, since we restricted the range of the basis functions, $\|\phi(x) - \alpha(p)\| = 1$ holds only when all basis functions in Φ are constant almost everywhere. In this case, the family $\mathscr{E}(\Phi)$ reduces to $\{q^0\}$ and the MLE problem becomes vacuous. Therefore, we assume in the following that $\|\phi(x) - \alpha(p)\| \in [0, 1)$ holds.

Theorem 1. For all $t \in \mathbb{N}$,

$$\frac{1}{2}\|\nabla D(q_{\hat{\theta}}\|q^t)\|^2 \le \frac{1}{2}\ln\frac{1}{1-\|\nabla D(q_{\hat{\theta}}\|q^t)\|^2} \tag{11}$$

$$\le D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}) \tag{12}$$

$$\le \|\nabla D(q_{\hat{\theta}}\|q^t)\|\tanh^{-1}(\|\nabla D(q_{\hat{\theta}}\|q^t)\|) \tag{13}$$

$$\le \ln\frac{1}{1-\|\nabla D(q_{\hat{\theta}}\|q^t)\|^2}. \tag{14}$$

Proof: Inequality (11) is easily derived from Taylor's Theorem. For proving (12) we follow [6]: Let $S \subset X$ be the finite support of the empirical measure on X induced by the sample $x_1,\cdots,x_m$. Observe that because of the normalization of the ϕ_k's, both $\|\phi(x)-\tilde{\alpha}\|$ and $\|\alpha(\theta)-\tilde{\alpha}\|$ lie in $[0,1)$ for all $x \in X$ and $\theta \in \mathbb{R}^d$. Rewrite (4) as

$$q^{t+1}(x) = q^t(x)\frac{e^{\Delta\theta^t\cdot\phi(x)}}{Z_{t+1}}, \tag{15}$$

where

$$Z_{t+1} = \int_X e^{\Delta\theta^t\cdot\phi}q^t = \exp[\psi(\theta^{t+1}) - \psi(\theta^t)]. \tag{16}$$

Using (5), (15), (16), and (9), we can show

$$\begin{aligned} D(q_{\hat{\theta}}&\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}) \\ &= \sum_{x\in S}(\Delta\theta^t\cdot\phi(x))\tilde{p}(x) - \ln Z_{t+1} \\ &= \Delta\theta^t\cdot\tilde{\alpha} - \ln\int_X \exp(\Delta\theta^t\cdot\phi)q^t \\ &= -\ln\int_X \exp[\Delta\theta^t\cdot(\phi-\tilde{\alpha})]q^t \qquad (17) \\ &\ge -\ln\left[\cosh(\|\Delta\theta^t\|) + \sinh(\|\Delta\theta^t\|)\frac{\Delta\theta^t}{\|\Delta\theta^t\|}\cdot(\alpha^t-\tilde{\alpha})\right]. \qquad (18) \end{aligned}$$

Let G be $\|\nabla D(q_{\hat{\theta}}\|q_\theta)\|$, where the gradient is evaluated at $\theta = \theta^t$. Then $G = \|\alpha^t - \tilde{\alpha}\|$ by (3). By (5),

$$\|\Delta\theta^t\| = \tanh^{-1}(G) = \ln\sqrt{\frac{1+G}{1-G}}, \tag{19}$$

since it is well known that

$$\tanh^{-1}(x) = \ln\sqrt{\frac{1+x}{1-x}}. \tag{20}$$

From (18) and (19), after some algebra we obtain

$$\begin{aligned} D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}) &\ge -\ln\left[\frac{1}{2}\left(\sqrt{\frac{1+G}{1-G}} + \sqrt{\frac{1-G}{1+G}}\right)\right. \\ &\qquad \left. - \frac{G}{2}\left(\sqrt{\frac{1+G}{1-G}} - \sqrt{\frac{1-G}{1+G}}\right)\right] \\ &= \frac{1}{2}\ln\frac{1}{1-G^2}. \end{aligned}$$

This proves (12). Inequality (13) is proven using (17), Jensen's inequality, and (5):

$$\begin{aligned} D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}) &= -\ln\int_X \exp[\Delta\theta^t\cdot(\phi-\tilde{\alpha})]q^t \\ &\le -\int_X \Delta\theta^t\cdot(\phi-\tilde{\alpha})q^t \\ &= \Delta\theta^t\cdot(\tilde{\alpha}-\alpha^t) \\ &= G\tanh^{-1}(G). \end{aligned}$$

Finally, (14) is obtained by applying (10). □

The choice of $\Delta\theta^t$ exactly maximizes (18). To see this, note that this term is maximized when $\Delta\theta^t = -\eta(\alpha^t - \tilde{\alpha}) = -\eta\nabla D(q_{\hat{\theta}}\|q^t)$ for some choice of $\eta > 0$. To find η, we differentiate

$$\frac{\partial}{\partial\eta}[\cosh(\eta G) - G\sinh(\eta G)] = G\sinh(\eta G) - G^2\cosh(\eta G).$$

Setting the derivative equal to 0 and solving with respect to η yields $\eta = [\tanh^{-1}(G)]/G$, from which the optimal increment (5) is derived.

We conclude the section by showing a couple of applications of Theorem 1 for obtaining lower bounds on the speed of convergence of the strategy.

Corollary 1: For all t $\ge$ 1,

$$D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}) \ge \frac{D(q_{\hat{\theta}}\|q^t)^2}{2\|\hat{\theta}-\theta^t\|}.$$

Proof: From (11) and (12) in Theorem 1, we obtain $\sqrt{2(D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}))} \ge \|\nabla D(q_{\hat{\theta}}\|q^t)\|$, which holds for any $t \ge 1$. Also, because of the convexity of $D(q_{\hat{\theta}}\|q_\theta)$ in θ, by a simple geometrical argument we have $\|\nabla D(q_{\hat{\theta}}\|q^t)\| \ge D(q_{\hat{\theta}}\|q^t)/\|\hat{\theta}-\theta^t\|$, for all $t \ge 1$. This completes the proof. □

For the second result we need a preliminary lemma.

Lemma 1 ([4]): Assume Φ is an orthonormal basis with respect to a density q whose log-density $\ln q$ is bounded. Let A be such that for all $\theta \in \mathbb{R}^d$,

$$\|\theta\cdot\phi\|_\infty \le A\|\theta\cdot\phi\|_{L_2(q)}. \tag{21}$$

Then, for any θ, $\theta' \in \mathbb{R}^d$,

$$D(q_\theta\|q_{\theta'}) \ge \frac{1}{2}\|\theta-\theta'\|^2\exp\left(-\|\ln\frac{q}{q_\theta}\|_\infty - 2A\|\theta-\theta'\|\right).$$

We are now ready to prove a second recurrence.

Theorem 2: Let Φ be orthonormal with respect to a log-bounded density q and such that (21) is satisfied for all $\theta \in \mathbb{R}^d$ and for some constant $A < \infty$. Moreover, assume $\|\ln q/q_{\hat{\theta}}\|_\infty$ is bounded. Then there are positive constants $a = 2\exp(\|\ln q/q_{\hat{\theta}}\|)$ and $b = e^{2A}$ such that for all $r > 0$,

$$D(q_{\hat{\theta}}\|q_{\theta^t}) - D(q_{\hat{\theta}}\|q_{\theta^{t+1}}) \ge \frac{D(q_{\hat{\theta}}\|q_{\theta^t})}{2ae^{br}}$$

holds for all $t \ge 1$ such that $\|\theta^t - \hat{\theta}\| \le r$.

Proof: Fix $r > 0$ and assume $t \ge 1$ is such that $\|\theta^t - \hat{\theta}\| \le r$. The theorem is proven by considering the following chain of inequalities.

$$\begin{aligned} \sqrt{2(D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}))} &\ge \|\nabla D(q_{\hat{\theta}}\|q^t)\| \\ &\ge \frac{D(q_{\hat{\theta}}\|q^t)}{\|\hat{\theta}-\theta^t\|} \\ &\ge \frac{D(q_{\hat{\theta}}\|q^t)}{\sqrt{D(q_{\hat{\theta}}\|q^t)ae^{br}}}. \end{aligned}$$

The first inequality is again a consequence of Theorem 1, the second is an application of Corollary 1, and the third is derived from Lemma 1. □

Finally, we show a corollary asserting exponentially fast convergence of our strategy in a region close to the optimum.

Corollary 2: If $\|\theta^t - \hat{\theta}\| \le r$ holds for all $t \ge t_0$ and for some $r > 0$, $t_0 \ge 1$. Then, under the same assumptions of Theorem 2,

$$D(q_{\hat{\theta}}\|q^t) \le \left(1 - \frac{1}{2ae^{br}}\right)^{t-t_0} D(q_{\hat{\theta}}\|q^{t_0})$$

for all $t \ge t_0$.

IV. Conclusions

In this paper we have described a strategy for likelihood maximization (relative entropy minimization) in families of exponential densities, assuming that the log-densities are spanned by a set of bounded basis functions. Our strategy is shown to perform steepest descent on an approximation of the relative entropy function. Upper and lower bounds on the decrease of the relative entropy at each iteration have been proven. Our bounds are expressed in terms of a function of the norm of the gradient and are tight within a constant factor of $\frac{1}{2}$. Bounds on the speed of convergence of our strategy have also been shown. An interesting open problem is to show that $\|\hat{\theta} - \theta^{t+1}\| < \|\hat{\theta} - \theta^t\|$ holds for all $t \ge t_0$ and for some $t_0 \ge 1$.

Appendix

Proof of Inequality (10). Using the equivalence (2), we show that the function

$$f(x) = \frac{x}{2} \ln\left(\frac{1+x}{1-x}\right) + \ln(1 - x^2)$$

is nonpositive in the interval $[-1, 1]$. Observe that

$$f'(x) = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right) - \frac{x}{1-x^2}$$
$$= \tanh^{-1}(x) - \frac{x}{1-x^2}.$$

A root of f' is 0. Also note that $f(0) = 0$. Since the second derivative

$$f''(x) = -\frac{2x^2}{(1-x^2)^2}$$

is 0 at $x = 0$ and negative elsewhere, $x = 0$ is the only extremum of f' and it is a maximum. This completes the proof of the lemma. □

References

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

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

[3] L. D. Brown, *Fundamentals of Statistical Exponential Families*, vol. 9 of Lecture Notes—Monograph Series. Institute of Math. Stat., 1986.

[4] A. R. Barron and C. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Stat.*, vol. 19(3), 1991.

[5] B. R. Crain, "Exponential models, maximum likelihood estimation, and the Haar condition," *J. Am. Stat. Ass.*, vol. 71 pp. 737–740, 1976.

[6] N. Littlestone, P. M. Long, and M. K. Warmuth, "On-line learning of linear functions," Tech. Rep. UCSC-CRL-91-29, University of California at Santa Cruz, 1991.

Bounds on the Size of Nonnegative Definite Circulant Embeddings of Positive Definite Toeplitz Matrices

G. N. Newsam and C. R. Dietrich

Abstract*—Recently Dembo *et al.* showed that an $N \times N$ positive definite Toeplitz matrix T could be embedded in a $2M \times 2M$ nonnegative definite circulant matrix S with $M = O(\kappa(T)N^2)$. This note shows that the size of the embedding can be reduced to $M = O(\kappa(T)^{1/2}N^{5/4})$ and that this is best possible for the technique presented by Dembo *et al.

***Index Terms*—Circulant embeddings, statistical simulations, Toeplitz matrices.**

I. Circulant Embeddings

An $N \times N$ symmetric Toeplitz matrix T is characterized by the vector $t = \{t_0, t_1, \cdots, t_{N-1}\}$ of elements in its first row, with $T_{mn} = t_{|m-n|}$. Let $M \ge N$ and s be any vector such that

$$s_n = t_n, \qquad n = 0, \cdots, N-1,$$
$$s_{2M-m} = s_m, \qquad m = 1, \cdots, M-1.$$

Then the $2M \times 2M$ symmetric circulant matrix S defined by $S_{mn} = s_{|m-n|}$ is termed a circulant embedding of T since any $N \times N$ block along the main diagonal of S is just a replication of T. In certain applications T is positive definite, and one would like to choose s so that S is also nonnegative definite.

Recently Dembo *et al.* [1] proved that if T is strictly positive definite then a nonnegative definite circulant embedding always exists. They did this by constructing such an embedding with $M = N + (\kappa(T)/\sqrt{6})N^2$. ($\kappa(T) \equiv \lambda_{\max}(T)/\lambda_{\min}(T)$ is the ratio of largest and smallest eigenvalues of T and is termed the condition number of T.) Nonnegative definite embeddings can be used to determine the maximum likelihood estimate of a Toeplitz covariance matrix by the entropy maximization algorithm [1], and to generate random vectors from a distribution with covariance matrix T [2], [3]. In these cases, computational efficiency demands that the embedding size M be as small as possible. The purpose of this communication is to explore the possibility of reducing the size of the embedding produced by the construction in [1]. We show that a different choice of embedding function in the construction, coupled with the use of tighter inequalities, significantly reduces M; we also show that this improved result is essentially the best possible obtainable by this construction. Finally we note that in some situations the improved construction may still produce impracticably large embeddings, and we indicate possible alternative embedding strategies for such cases.

In order to point out the improvements possible, we first review the construction as presented in [1]. The construction starts by selecting an $M \ge N$ and a symmetric continuous positive definite function $r(x)$ such that $r(0) = 1$, $r(x) > 0$ for $|x| < 1$, and $r(x) = 0$ for $|x| \ge 1$. For the moment, we do not further specify M and r; we will make particular choices when their roles become clearer later in the construction. Let $\boldsymbol{r}_\infty$ be

Manuscript received December 20, 1992; revised November 14, 1993.

G. N. Newsam is with Information Technology Division, Defence Science and Technology Organisation, P.O. Box 1500, Salisbury, S.A. 5108, Australia.

C. R. Dietrich is with the Division of Australian Environmental Sciences, Griffith University, Nathan, Qld. 4111, Australia.

IEEE Log Number 9402575.

0018-9448/94$04.00 © 1994 IEEE