Bounds on Approximate Steepest Descent for Likelihood Maximization in Exponential Families

*Nicolò Cesa-Bianchi**
Computer Science Department
Università di Milano
Via Comelico 39/41, 20135 Milano (Italy)
cesabian@imiucca.csi.unimi.it

Anders Krogh
Computer Science Department
University of California
Santa Cruz CA 95064
krogh@spica.ucsc.edu

Manfred K. Warmuth
Computer Science Department
University of California
Santa Cruz CA 95064
manfred@mira.ucsc.edu

December 17, 1992

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.

*This research was done while this author was visiting UC Santa Cruz partially supported by the "Progetto finalizzato sistemi informatici e calcolo parallelo" of CNR under grant 91.00884.69.115.09672.

1 Introduction

Consider the following problem: Given a random sample $x_1, \ldots, x_m$ 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. [Kul59, Csi75]). 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 entropy (Kullback-Leibler divergence) $D(p\|p')$ between two densities p and p' on $(X, \mathcal{B})$:

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

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

We will use the notation $\theta \cdot \phi(x)$ for the inner product $\sum_k \theta_k \phi_k(x)$.

We now define the regular exponential family $\mathcal{E}(\Phi) = \left\{q_\theta : \theta \in \mathbb{R}^d\right\}$ 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 $\alpha(p) = (\alpha_1(p), \ldots, \alpha_d(p))$ by

$$\alpha_k(p) = \mathbf{E}_p[\phi_k] \qquad \text{for } k = 1, \ldots, d$$

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

$$\alpha_k(\theta) = \mathbf{E}_{q_\theta}[\phi_k] \qquad \text{for } k = 1, \ldots, d.$$

If Φ is a set of linearly independent functions[2], it is known that ψ is strictly convex (see e.g. [Bro86]). As a consequence, also $D(p\|q_\theta)$ is strictly convex in θ, which is seen from

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

where $H(p)$ is the entropy $\mathbf{E}_p[-\ln p]$. 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 $\theta' = \theta^*$.

[1] The strategy was originally introduced in [LLW91] as an iterative method for the solution of sparse systems of linear equations.

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

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

2 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 $(k = 1, \ldots, d)$ in the interval $[-\sqrt{1/4d}, \sqrt{1/4d}]$. This ensures that for any density p and for any $x \in X$, $\|\phi(x) - \alpha(p)\| \in [0, 1]$.

On each run, the strategy is given as input a reference density q^0 and a random sample $x_1, \ldots, x_m$ independently drawn from a distribution P with density $p(x)$. The output consists in a infinite sequence $q^1, q^2, \ldots$ of densities in $\mathcal{E}(\Phi)$.
Let $\alpha^t = (\alpha_1^t, \ldots, \alpha_d^t)$ such that

$$\alpha_k^t = \mathbf{E}_{q^t}[\phi_k] \qquad \text{for } k = 1, \ldots, d$$

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

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

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

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

where θ^t is the parameter vector after the t-th 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 $\mathcal{E}(\Phi)$.

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.

3 Analysis

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

$$\begin{aligned} \ell(\theta) &= -\ln \prod_{i=1}^{m} q_\theta(x_i) \\ &= m\,(\theta \cdot \tilde{\alpha} - \psi(\theta)). \end{aligned} \tag{6}$$

Hence, for a set Φ of linearly independent basis functions, the maximum likelihood estimate $q_{\hat{\theta}}$ in the family $\mathcal{E}(\Phi)$ 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 [BS90, Cra76].

Using equations (2), (6) and (7), we can rewrite the Kullback-Leibler divergence as $D(q_{\hat{\theta}} \| q_\theta) = \hat{\theta} \cdot \tilde{\alpha} - \psi(\hat{\theta}) - \ell(\theta)/m$. 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 equation (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 $x \in [-1, 1]$

$$e^{kx} \leq \cosh(k) + x \sinh(k). \tag{8}$$

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

$$x \tanh^{-1}(x) \leq \ln \frac{1}{1 - x^2}. \tag{9}$$

The first inequality can be proven by applying Jensen's inequality. The second inequality is proven in the Appendix.

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

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

$$\begin{aligned}
\frac{1}{2}\|\nabla D(q_{\hat{\theta}} \| q^t)\|^2 &\leq \frac{1}{2} \ln \frac{1}{1 - \|\nabla D(q_{\hat{\theta}} \| q^t)\|^2} && (10)\\
&\leq D(q_{\hat{\theta}} \| q^t) - D(q_{\hat{\theta}} \| q^{t+1}) && (11)\\
&\leq \|\nabla D(q_{\hat{\theta}} \| q^t)\| \tanh^{-1}(\|\nabla D(q_{\hat{\theta}} \| q^t)\|) && (12)\\
&\leq \ln \frac{1}{1 - \|\nabla D(q_{\hat{\theta}} \| q^t)\|^2}. && (13)
\end{aligned}$$

Proof. Inequality (10) is easily derived from Taylor's Theorem. For proving inequality (11) we follow [LLW91]: Let $S \subset X$ be the finite support of the empirical measure on X induced by the sample $x_1, \ldots, x_m$. 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 equation (4) as

$$q^{t+1}(x) = q^t(x) \frac{e^{\Delta\theta^t \cdot \phi(x)}}{Z_{t+1}} \tag{14}$$

where

$$Z_{t+1} = \int_X e^{\Delta\theta^t \cdot \phi} q^t. \tag{15}$$

Using equations (5), (14), (15) and inequality (8) we can show

$$D(q_{\hat{\theta}} \| q^t) - D(q_{\hat{\theta}} \| q^{t+1}) = \sum_{x \in S} (\Delta\theta^t \cdot \phi(x)) \tilde{p}(x) - \ln Z_{t+1}$$

$$
\begin{aligned}
&= \Delta\theta^t \cdot \tilde{\alpha} - \ln \int_X \exp\left[\Delta\theta^t \cdot \phi\right] q^t \\
&= -\ln \int_X \exp\left[\Delta\theta^t \cdot (\phi - \tilde{\alpha})\right] q^t \qquad (16) \\
&\geq -\ln \left[\cosh(\|\Delta\theta^t\|) + \sinh(\|\Delta\theta^t\|)\frac{\Delta\theta^t}{\|\Delta\theta^t\|} \cdot (\alpha^t - \tilde{\alpha})\right]. \qquad (17)
\end{aligned}
$$

Let $G = \|\nabla D(q_{\hat{\theta}}\|q^t)\| = \alpha^t - \tilde{\alpha}$. Replacing $\Delta\theta^t$ with the right-hand side of equation (5) and using the standard formula

$$
\tanh^{-1}(x) = \ln \sqrt{\frac{1+x}{1-x}} \qquad (18)
$$

after some algebra we obtain

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

This proves inequality (11).
Inequality (12) is proven using equation (16), Jensen's inequality and equation (5):

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

Finally, inequality (13) is obtained by applying inequality (9). □

The choice of $\Delta\theta^t$ exactly maximizes (17). 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

$$
\begin{aligned}
&\frac{\partial}{\partial\eta}\Big[\cosh(\eta G) - G\sinh(\eta G)\Big] \\
&= G\sinh(\eta G) - G^2\cosh(\eta G).
\end{aligned}
$$

Setting the derivative equal to 0 and solving with respect to η yields

$$
\eta = \frac{\tanh^{-1}(G)}{G}
$$

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

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

for $t = 1, 2, \ldots$.

Proof. From inequalities 10 of Theorem 1 we obtain

$$\sqrt{2(D(q_{\hat{\theta}}\|q^t) - D(q_{\hat{\theta}}\|q^{t+1}))} \geq \|\nabla D(q_{\hat{\theta}}\|q_{\theta^t})\|$$

which holds for any $t = 1, 2, \ldots$. Also, because of the convexity of $D(q_{\hat{\theta}}\|q_{\theta})$,

$$\begin{aligned} \|D(q_{\hat{\theta}}\|q_{\theta^t})\| &\leq \|\nabla D(q_{\hat{\theta}}\|q_{\theta^t})\|\|\hat{\theta} - \theta^t\| \\ &\leq \|\nabla D(q_{\hat{\theta}}\|q_{\theta^t})\|\|\hat{\theta}\|. \end{aligned} \tag{19}$$

A simple combination of the above inequalites then yields the corollary. □

For the second result we need a preliminary lemma.

Lemma 1 ([BS90]) *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* $\theta \in \mathbb{R}^d$

$$\|\ln q_{\theta}\|_{\infty} \leq A\|\ln q_{\theta}\|_{L_2(q)}. \tag{20}$$

Choose $\theta, \theta' \in \mathbb{R}^d$. *If* $\|\theta - \theta'\| \leq r$, *then*

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

□

We are now ready to prove a second recurrence which holds in a region close to the optimum.

Theorem 2 *Let* Φ *be orthogonal with respect a log-bounded density* q *and such that inequality (20) is satisfied for some constant* $A < \infty$. *Then there are positive constants* a *and* b *such that for all* $\theta^t \in \mathbb{R}^d$, *if* $\|\theta^t - \hat{\theta}\| \leq r$, *then the following recurrence holds for all* $t = 1, 2, \ldots$

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

Proof. The theorem is proven by considering the following chain of inequalities.

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

The first inequality is again a consequence of Theorem 1, the second is an application of inequality 19 in Corollary 1, and the third is derived from Lemma 1. This concludes the proof. □

4 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 bee 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 been also shown.

References

[Bro86] L.D. Brown. *Fundamentals of Statistical Exponential Families*, volume 9 of *Lecture Notes - Monograph Series*. Institute of Math. Stat., 1986.

[BS90] A.R. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. Technical Report 8, University of Illinois at Urbana-Champaign, Department of Statistics, 1990. To appear in the *Annals of Statistics.*

[Cra76] B.R. Crain. Exponential models, maximum likelihood estimation, and the Haar condition. *Journal of the American Statistical Association*, 71:737–740, 1976.

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

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

[LLW91] N. Littlestone, P.M. Long, and M.K. Warmuth. On-line learning of linear functions. Technical Report UCSC-CRL-91-29, UC Santa Cruz, 1991. An extended abstract appeared in: *Proceedings of the 23rd ACM Symposium on the Theory of Computation.*

Appendix

Proof of inequality (9). Using the equivalence (18) we show that the function

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

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

$$\begin{aligned} f'(x) &= \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right) - \frac{x}{1-x^2} \\ &= \tanh^{-1}(x) - \frac{x}{1-x^2}; \end{aligned}$$

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. □