Additive Models, Boosting, and Inference for Generalized Divergences

John Lafferty [†]

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lafferty@cs.cmu.edu

Abstract

We present a framework for designing incremental learning algorithms derived from generalized entropy functionals. Our approach is based on the use of Bregman divergences together with the associated class of additive models constructed using the Legendre transform. A particular one-parameter family of Bregman divergences is shown to yield a family of loss functions that includes the log-likelihood criterion of logistic regression as a special case, and that closely approximates the exponential loss criterion used in the AdaBoost algorithms of Schapire *et al.*, as the natural parameter of the family varies. We also show how the quadratic approximation of the gain in Bregman divergence results in a weighted least-squares criterion. This leads to a family of incremental learning algorithms that builds upon and extends the recent interpretation of boosting in terms of additive models proposed by Friedman, Hastie, and Tibshirani.

1 Introduction

Logistic regression is a widely used statistical methodology for classification problems based upon maximum likelihood. Boosting is a popular technique for combining several simple yet "weak" learners into an accurate classifier using a voting scheme, where the learners and votes are chosen to minimize an exponential loss criterion. A recent paper of Friedman, Hastie and Tibshirani [16] presents a statistical interpretation of the AdaBoost algorithms of Freund and Schapire [15] and Schapire and Singer [23]. In particular, Friedman *et al.* show how boosting algorithms result from building additive models using Newton updates of the exponential loss function, making a comparison between boosting and stepwise logistic regression methods, which are formalized in the "LogitBoost" algorithm.

The perspective and some of the main conclusions of [16] can be summarized graphically. When predicting a random variable $Y \in \{-1, +1\}$ in the two-class case, the AdaBoost and LogitBoost algorithms each fit a model of the form $p(1\,|\,x) = \frac{e^{F(x)}}{e^{F(x)}+e^{-F(x)}}$, with $F(x) = \sum_m f_m(x)$. However, the AdaBoost algorithms use the exponential loss function $E\left[e^{-yF(x)}\right]$ while logistic regression uses the log-likelihood criterion, given by $E\left[\log\left(1 + e^{-2yF(x)}\right)\right]$, where $E\left[\cdot\right]$ denotes expectation with respect to the empirical sample. A plot of these loss functions as a function of $yF(x)$, is shown in Figure 1, together with the error and mean-squared error.

We note that the plot corresponding to Figure 1 is shown in [16] using $\log\left(1 + e^{-yF(x)}\right)$ rather than the log-likelihood $\log\left(1 + e^{-2yF(x)}\right)$ for the log-likelihood loss, resulting in the log-likelihood curve falling *below* the exponential criterion in the error region $-1 \leq yF(x) < 0$, whereas it lies above the exponential criterion in this region in our plot. While the two additive models differ only by a constant multiplicative factor, the plot given in Figure 1 is compares the two loss functions for the *same* additive model specified by $F(x) = \frac{1}{2}\log\frac{p(1\,|\,x)}{p(-1\,|\,x)}$. The curves are shown for the region $-1 < yF(x) < 1$, which is the appropriate domain for using confidence-rated predictions that are probabilities. Of course, the exponential curve lies above the log-likelihood for sufficiently large negative values of $yF(x)$.

While the exponential and log-likelihood loss functions have similar properties, there is a significant gap between them in the region $-1 < yF(x) < 1$, which could lead to qualitatively and quantitatively different behavior on real data. Indeed, the experiments carried out in [16] demonstrate that the LogitBoost and AdaBoost algorithms can yield significantly different models. The main result of the current paper is to show how this gap can be bridged within a unified framework for statistical inference. In particular, we show how a family of incremental learning algorithms derived from Bregman divergences can be constructed to include stepwise logistic regression (and the LogitBoost algorithm) as a special case, and to closely approximate the AdaBoost algorithm as the natural parameter of the family

[†] Research supported in part by NSF grant CCR-9805366, NSF KDI grant IIS-9873009, and an IBM University Partnership Award.

COLT '99 7/99 Santa Cruz, CA, USA

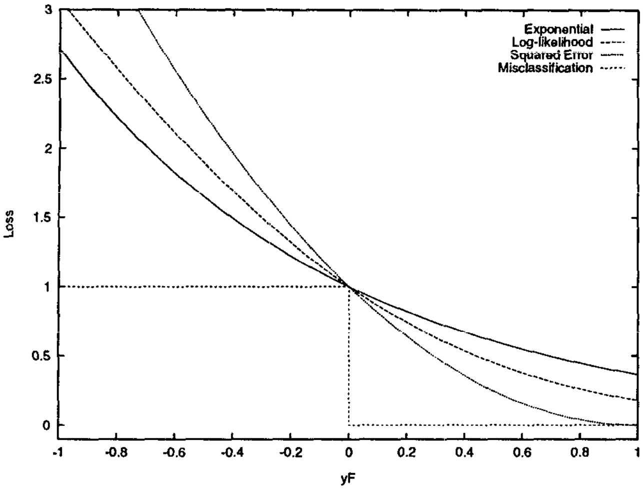

Figure 1. *Loss functions for additive models: The exponential criterion, log-likelihood, and squared error.*

varies. For appropriately chosen divergences in this family, we expect that the resulting learning algorithms and AdaBoost would be indistinguishable in practice. We also show how these learning algorithms have an interpretation in terms of weighted least squares regression, which sheds additional light on how AdaBoost differs from more standard likelihood-based approaches.

Informally, the Bregman divergence $D_\phi(p,q)$ of a convex function ϕ measures the convexity of ϕ between points p and q relative to its linear approximation at q. Bregman divergences include the Kullback-Leibler divergence as a special case, and a rich theory is associated with them based upon convex duality. In particular, the fundamental results of information geometry and the maximum entropy method generalize to Bregman divergences [9, 11]. They have recently been used in the machine learning literature in the work of Warmuth and his colleagues, as a means of obtaining loss bounds for a broad class of on-line learning algorithms [17, 19]. As we indicate in this paper, many of the bounds and techniques one can obtain for maximum likelihood estimation, based upon the Kullback-Leibler divergence for exponential families, have analogues for general Bregman divergences.

The use of statistical inference techniques based on the Bregman divergences is attractive for several reasons. This framework has been given a strong axiomatic justification by Csiszár [9]. It applies not only to inference of probability distributions, but also to inference of unnormalized measures, which can be useful in applications such as image processing. It enables a principled treatment of missing or hidden data. Since the mathematical framework of Bregman divergences draws on the geometry of convex duality [21], familiar from the theory of exponential families [1, 4], the method of alternating projections [11] can be exploited, which can be useful for establishing convergence properties.

On a historical note, our use of stepwise learning algorithms in the maximum entropy framework, and their extension to Bregman divergence minimization, began in the early 1990s at the IBM Watson Research Center. While some of this work was published [3, 12], much of the research was either unpublished or only briefly described in IBM invention disclosures, and was carried out without knowledge of boosting-style classification procedures in the machine learning literature. While the resulting feature selection algorithms for exponential models that were developed bear a strong resemblance to AdaBoost and LogitBoost, there are some important differences. Some of the relevant issues are mentioned briefly in the discussion that concludes this extended abstract.

In the following section we show how the "gain" in log-likelihood for logistic regression is approximated to quadratic order, rederiving and summarizing some of the results of Friedman *et al.* [16]. Section 3 introduces the notion of Bregman divergence, and Section 4 shows how the Legendre transform for a Bregman divergence defines a family of additive models. Section 4 also discusses the special case of a one-parameter family of divergences shown by Csiszár to have an axiomatic justification, and how the loss functions for this family closely approximate the exponential criterion. Sections 5 and 6 show how to estimate to quadratic order the gain due to adding a single feature, leading to a family of boosting-style algorithms that helps bridge the gap between logistic regression and the AdaBoost algorithms. Section 7 concludes with a brief discussion and directions for future work.

2. Approximate Gains

In this section we briefly discuss the relevant results of [16], introducing the notation and perspective that we will use throughout. Let the random variable Y denote the class label, which is to be predicted based upon a feature vector represented by a random variable $X \in \mathbb{R}^M$. We first consider exponential models of the form

$$p(Y = y \,|\, X = x) = \frac{e^{\lambda \cdot f(x,y)}}{\sum_{y'} e^{\lambda \cdot f(x,y')}}$$

where $\lambda \cdot f(x,y) = \sum_{i=1}^{n} \lambda_i f_i(x,y)$. We will refer to the functions $f_i(x,y) \in \mathbb{R}$ as *features*, understanding that they may be compound features that are built up from more elementary features using decision trees, for instance. We denote by $\widetilde{p}(x,y)$ the *empirical distribution* determined by a collection of training examples $\{(x_i, y_i)\}_{i=1}^{N}$, so that $\widetilde{p}(x,y) = \frac{1}{N}\sum_{i=1}^{N} \delta(x_i, x)\,\delta(y_i, y)$. Let $D(p \,\|\, q)$ denote the average Kullback-Leibler divergence between two conditional distributions $p(y \,|\, x)$ and $q(y \,|\, x)$ with respect to $\widetilde{p}(x)$:

$$D(p \,\|\, q) = \sum_{x} \widetilde{p}(x) \sum_{y} p(y \,|\, x) \log \frac{p(y \,|\, x)}{q(y \,|\, x)}.$$

We will use the notation $p[f]$ to denote the expectation of f with respect to p. In particular, $\widetilde{p}[f] = \sum_{x,y} \widetilde{p}(x,y)\, f(x,y)$ denotes the expectation with respect to $\widetilde{p}$, $q[f \,|\, x]$ denotes the conditional expectation $q[f \,|\, x] = \sum_y q(y \,|\, x)\, f(x,y)$, and $q[f] = \sum_x \widetilde{p}(x)\, q[f \,|\, x]$ is the expectation with respect to $\widetilde{p}(x)\, q(y \,|\, x)$. We use $\mathcal{L}(\widetilde{p}, q)$ to denote the log-likelihood $\mathcal{L}(\widetilde{p}, q) = \sum_{x,y} \widetilde{p}(x,y) \log q(y \,|\, x)$.

Now, let q be a reference model, which in general need not be an additive model, and let f be a feature. We define the *gain* $\mathcal{G}(f;\widetilde{p},q)$ of the feature f with respect to q to be the largest improvement in likelihood that can result from adding f to q:

$$\begin{aligned}\mathcal{G}(f;\widetilde{p},q) &\overset{\text{def}}{=} \max_{\lambda} \mathcal{L}(\widetilde{p},q_\lambda) - \mathcal{L}(\widetilde{p},q) \\ &= \max_{\lambda} D(\widetilde{p}\,\|\,q) - D(\widetilde{p}\,\|\,q_\lambda)\end{aligned}$$

where $q_\lambda(y\,|\,x) = \frac{e^{\lambda f(x,y)}}{Z_\lambda(x)} q(y\,|\,x)$. The normalizing term $Z_\lambda(x)$ is given by $Z_\lambda(x) = \sum_{y'} e^{\lambda f(x,y')} q(y'\,|\,x)$.

The following result shows how the change in log-likelihood $\mathcal{L}(\widetilde{p},q_\lambda) - \mathcal{L}(\widetilde{p},q)$ has a simple interpretation when it is approximated to quadratic order in λ. We will derive an analogous result for Bregman divergences in Section 5.

Proposition 2.1. *To quadratic order, the gain $\mathcal{G}(f;\widetilde{p},q)$ is approximated by*

$$\mathcal{G}(f;\widetilde{p},q) \approx \mathcal{G}_2(f;\widetilde{p},q) \overset{\text{def}}{=} \frac{1}{2}\frac{(q[f]-\widetilde{p}[f])^2}{q\,[f^2 - q[f\,|\,x]^2]}. \tag{2.1}$$

Proof. Since $\log q_\lambda(y\,|\,x) = \lambda f(x,y) - \log Z_\lambda(x) + \log q(y\,|\,x)$, we have that $\mathcal{L}(\widetilde{p},q_\lambda) - \mathcal{L}(\widetilde{p},q) = \lambda\widetilde{p}[f] - \widetilde{p}\,[\log Z_\lambda]$. Calculating the Taylor expansion of $\log Z_\lambda(x)$ around $\lambda = 0$, we see that

$$\begin{aligned}&\log Z_\lambda(x) = \\ &= \lambda q[f\,|\,x] + \frac{\lambda^2}{2}\left(q[f^2\,|\,x] - q[f\,|\,x]^2\right) + O(\lambda^3).\end{aligned}$$

As a result,

$$\begin{aligned}&\mathcal{L}(\widetilde{p},q_\lambda) - \mathcal{L}(\widetilde{p},q) = \\ &= \lambda(\widetilde{p}[f] - q[f]) - \frac{\lambda^2}{2}\widetilde{p}\,[q[f^2\,|\,x] - q[f\,|\,x]^2] + O(\lambda^3) \\ &= \lambda(\widetilde{p}[f] - q[f]) - \frac{\lambda^2}{2}q[f^2 - q[f\,|\,x]^2] + O(\lambda^3).\end{aligned}$$

This is maximized by taking

$$\lambda^* = \frac{\widetilde{p}[f] - q[f]}{q[f^2 - q[f\,|\,x]^2]},$$

and for this choice of λ,

$$\mathcal{L}(\widetilde{p},q_{\lambda^*}) - \mathcal{L}(\widetilde{p},q) \approx \frac{1}{2}\frac{(\widetilde{p}[f]-q[f])^2}{q[f^2 - q[f\,|\,x]^2]}$$

yielding the statement in (2.1). ▮

This approximation can be reformulated as a large deviations result for estimating the change in likelihood due to adding a feature, and to a chi-square interpretation of the gain. These facts are perhaps well-known in the statistics literature; we first learned them from Stephen and Vincent Della Pietra. The chi-square interpretation gives a useful significance test for feature selection. The simple proofs of these statements are given in an appendix.

Corollary 2.2. *Let $\mathcal{Q}_{q,f}$ be the one-parameter family of exponential models given by*

$$\mathcal{Q}_{q,f} = \left\{ q_\lambda(y\,|\,x) = \frac{e^{\lambda f(x,y)}}{Z_\lambda(x)} q(y\,|\,x)\,;\ \lambda \in \mathbb{R} \right\}$$

and let $\pi(\lambda\,|\,\mathcal{Q}_{q,f})$ be the non-conjugate prior distribution on the parameter λ which is Gaussian with mean zero and variance $\frac{1}{\sqrt{N}} < \sigma^2 \leq \infty$. Then the log-probability of N events $d = \{(x_i,y_i)\}_{i=1}^N$ is given by

$$\begin{aligned}\log p\,(d\,|\,\mathcal{Q}_{q,f}) &= \log \int p(d\,|\,q_\lambda)\,\pi(\lambda\,|\,\mathcal{Q}_{q,f})\,d\lambda \\ &= \log p(d\,|\,q) - N\mathcal{G}_2(f;\widetilde{p},q) + O(\log N).\end{aligned}$$

Corollary 2.3. *Fixing $q(y\,|\,x)$, consider $\widetilde{p}(x,y)$ to be the empirical distribution of a sample of size N drawn from $q(x,y) = \widetilde{p}(x)\,q(y\,|\,x)$. Then the quadratic approximation is distributed as*

$$\mathcal{G}_2(f;\widetilde{p},q) \sim \frac{\gamma}{2N}\chi^2$$

where

$$\gamma = \frac{q\,[(f - q[f])^2]}{q\,[(f - q[f\,|\,x])^2]}$$

and χ^2 is the chi-squared distribution with one degree of freedom.

An important special case is the two-class problem, where $y \in \{-1,+1\}$. We first consider the restriction to binary features $f(x,y) = -yf(x) \in \{-1,+1\}$. In this situation the analysis is particularly simple and intuitive.

Following some of the notation of [16], we set $q(x) = q(1\,|\,x) = 1 - q(-1\,|\,x)$, and let $y^* = (y+1)/2 \in \{0,1\}$ be the representation of y as an indicator variable. Then we can express the quadratic approximation of the gain in terms of the variance $q(x)(1-q(x))$:

$$\mathcal{G}_2(\widetilde{p};f,q) = \frac{1}{2}\frac{(\widetilde{p}[y^*f(x)] - \widetilde{p}[f(x)q(x)])^2}{\widetilde{p}[q(x)(1-q(x))]}$$

Relaxing our restriction to binary features and minimizing pointwise for $f(x) \in \mathbb{R}$ yields

$$\widehat{f}(x) = \frac{1}{2}\frac{\widetilde{p}[y^* - q(x)\,|\,x]}{\widetilde{p}[q(x)(1-q(x))]}$$

which is equivalent to a weighted least squares criterion. This analysis leads directly to the *LogitBoost* algorithm of [16].

A similar analysis leads to the *discrete AdaBoost* algorithm of Freund and Schapire [15]. In this case we consider unnormalized models of the form

$$\bar{q}_\lambda(y\,|\,x) = e^{\lambda f(x,y)} q(y\,|\,x)$$

where $q(y\,|\,x)$ is a reference measure. The objective function now used to judge an additive model is the *exponential criterion*: $\mathcal{E}(\widetilde{p},\bar{q}_\lambda) = \sum_{x,y} \widetilde{p}(x,y)\,q(y\,|\,x)\,e^{f(x,y)}$. We again assume that the features are binary, taking the form $f(x,y) = -yf(x)$ with $f(x) \in \{-1,+1\}$. Then

$$\begin{aligned}\mathcal{E}(\widetilde{p},q) - \mathcal{E}(\widetilde{p},\overline{q}_\lambda) &= \sum_{x,y}\widetilde{p}(x,y)q(y\,|\,x)\left(1-e^{-\lambda y f(x)}\right)\\ &= \sum_{x,y}\widetilde{p}(x,y)q(y\,|\,x)\left(\lambda y f(x) - \frac{1}{2}\lambda^2\right) + o(\lambda^2).\end{aligned}$$

Maximizing this quadratic approximation as a function of λ, we find that

$$\lambda^* = \frac{\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)\,yf(x)}{\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)}.$$

The *gain in the exponential criterion* is thus estimated to be

$$\begin{aligned}\mathcal{G}_{\mathcal{E}}(\widetilde{p};f,q) &= \max_{\lambda}\mathcal{E}(\widetilde{p},q) - \mathcal{E}(\widetilde{p},\overline{q}_\lambda)\\ &\approx \frac{1}{2}\frac{\left(\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)\,yf(x)\right)^2}{\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)}.\end{aligned}$$

Choosing $f(x)$ pointwise to maximize this approximation leads to the choice

$$\begin{aligned}\widehat{f}(x) &= \arg\max_{f(x)}\frac{1}{2}\frac{\left(\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)\,yf(x)\right)^2}{\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)}\\ &= \arg\min_{f(x)}\frac{\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)\,(y-f(x))^2}{\sum_{x,y}\widetilde{p}(x,y)\,q(y\,|\,x)}\end{aligned}$$

using the facts that $y^2 = f(x)^2 = 1$. In boosting-style algorithms, $\widehat{f}$ is selected using a regression technique such as decision trees, and the weight λ^* is then chosen to maximize the full gain $\mathcal{E}(\widetilde{p};f,q) - \mathcal{E}(\widetilde{p};f,\overline{q}_\lambda)$, rather than its quadratic approximation, leading to

$$\lambda^* = \frac{1}{2}\log\frac{q[\delta(1,yf(x))]}{1-q[\delta(1,yf(x))]}$$

which specifies the discrete AdaBoost algorithm. In the following sections we will show how this analysis can be extended to a broad class of loss functions, allowing us to "reconcile" the AdaBoost and LogitBoost algorithms.

3. Generalized Divergences

If $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is a strictly convex C^1 function, the *Bregman divergence* $D_\phi(p,q)$ between discrete measures $p(x)$ and $q(x)$ is defined by

$$D_\phi(p,q) = \sum_x \phi(p(x)) - \phi(q(x)) - \phi'(q(x))(p(x)-q(x)).$$

When $\phi(x) = x^2$, we obtain the mean-squared error

$$D_{x^2}(p,q) = \sum_x (p(x) - q(x))^2$$

and when $\phi(x) = x\log x$ we obtain the *I-divergence*, or extended Kullback-Leibler divergence

$$D_{x\log x}(p,q) = \sum_x\left(p(x)\log\frac{p(x)}{q(x)} - p(x) + q(x)\right).$$

More generally, if $\Phi : \mathbb{R}^r \longrightarrow \mathbb{R}$ is strictly convex and C^1, the Bregman divergence $D_\Phi(p,q)$ is defined as

$$D_\Phi(p,q) = \Phi(p) - \Phi(q) - \nabla\Phi(q)\cdot(p-q)$$

where $p = (p_1,\ldots,p_r)$.

The Kullback-Leibler divergence between models in an exponential family can be interpreted as a Bregman divergence defined on the parameter vector. That is, if $p_\theta(x)$ is an exponential model of the form $p_\theta(x) = Z(\theta)^{-1}\,e^{\theta\cdot f(x)}$ then the divergence $D(p_{\theta_1}\,\|\,p_{\theta_2})$ is equal to the Bregman divergence for $\phi(\theta) = -\log Z(\theta)$:

$$D(p_{\theta_1}\,\|\,p_{\theta_2}) = (\theta_1-\theta_2)\cdot p_{\theta_1}[f] - (\log Z(\theta_1) - \log Z(\theta_2)).$$

For a tutorial introduction to these generalized entropy measures we refer to [11]. An elementary proof of the basic duality theorem and Pythagorean property for Bregman divergences was given in [13], following the proof in the case of the I-divergence that was given in [12]. A practical algorithm for minimizing Bregman divergences subject to linear constraints—and therefore solving generalized maximum entropy problems—was presented in [14]. In this paper a framework for incremental, or "stepwise" learning using these divergences was also presented.

4. Legendre Transforms and Additive Models

The basic tool in our analysis is the Legendre transform. This transform defines the class of additive models that we work with for a given Bregman divergence. It is central to convex optimization for these divergences, since it provides the link between the primal and dual problems [21]. The transform will be defined with respect to a set $\mathcal{S} \subset \mathbb{R}^r$ which is either the collection of probability distributions on r events, denoted by Δ_r, or the set of all positive measures, denoted by $\mathbb{R}^r_+$. Before proceeding, we need to be more precise in our use of the term additive model.

Definition 3.1. *Let $\mathcal{S} \subset \mathbb{R}^r$ be a set of measures. An additive model for $\mathcal{S}$ is defined by an action $\gamma : \mathbb{R}^r \times \mathcal{S} \longrightarrow \mathcal{S}$ satisfying the homomorphism property $\gamma(r_1 + r_2, s) = \gamma(r_1, \gamma(r_2, s))$ for all $r_1, r_2 \in \mathbb{R}^r$ and $s \in \mathcal{S}$.*

Lemma 3.2. *Given a convex function $\Phi : \mathbb{R}^r \longrightarrow \mathbb{R}$, let D_Φ be the Bregman divergence defined on measures $p = (p_1,\ldots,p_r) \in \mathcal{S} \subset \mathbb{R}^r$. Define the Legendre transform $v \circ_\Phi q$ by*

$$v \circ_\Phi q = \arg\max_{p\in\mathcal{S}}\; v\cdot p - D_\Phi(p,q).$$

Then the map $(v,q) \mapsto v \circ_\Phi q$ defines an additive model for $\mathcal{S} = \Delta_r$ and $\mathcal{S} = \mathbb{R}^r_+$.

To help clarify notation, we should point out that when p is a probability distribution, the dot product $v\cdot p$ with a vector $(v_1, v_2 \ldots, v_r) \in \mathbb{R}^r$ is the same as the expectation $E_p[v] = p[v]$ viewing v as the random variable $v(x_i) = v_i$.

Proof of Lemma 3.2. We use a calculus of variations argument to characterize $v \circ_\Phi q$. First suppose that $\mathcal{S} = \mathbb{R}^r$, the collection of all measures. Let $p(t) \in \mathbb{R}^r$ be a one-parameter family of measures with $p(0) = v \circ_\Phi q$ and $\left.\frac{dp}{dt}\right|_{t=0} = \delta p$. From the definition of the Legendre transform we have that

$$v \cdot \delta p - \nabla\Phi(v \circ_\Phi q) \cdot \delta p + \nabla\Phi(q) \cdot \delta p = 0.$$

Since the derivative δp is arbitrary, we see that the distribution $v \circ_\Phi q$ is uniquely determined by the condition

$$\nabla\Phi(v \circ_\Phi q) = \nabla\Phi(q) + v.$$

It follows that

$$\begin{aligned} \nabla(v \circ_\Phi (w \circ_\Phi q)) &= \nabla\Phi(w \circ_\Phi q) + v \\ &= \nabla\Phi(q) + w + v \\ &= \nabla\Phi((v+w) \circ_\Phi q) \end{aligned}$$

showing that $(v+w) \circ_\Phi q = v \circ_\Phi (w \circ_\Phi q)$.

If now $\mathcal{S} = \Delta_r$ is the collection of all probability distributions, we need to carry out a constrained maximization; a similar calculation will apply for $\mathcal{S} = \mathbb{R}^r_+$. Introducing a Lagrange multiplier $\psi(v, q)$ for the constraint $p \cdot \bar{1} = 1$, where $\bar{1}$ denotes the vector of all 1s, we see that the following equation must be satisfied at the maximum:

$$v \cdot \delta p - \psi(v,q)\,\bar{1} \cdot \delta p - \nabla\Phi(v \circ_\Phi q) \cdot \delta p + \nabla\Phi(q) \cdot \delta p = 0$$

showing that $v \circ_\Phi q$ is now uniquely determined by the condition $\nabla\Phi(v \circ_\Phi q) = \nabla\Phi(q) + v - \psi(v,q)\bar{1}$.

To prove the homomorphism property, it suffices to show that $\psi(v_1 + v_2, q) = \psi(v_2, q) + \psi(v_1, v_2 \circ_\Phi q)$. For this purpose, let us distinguish the constrained and unconstrained problems by temporarily denoting the unconstrained action by $v \odot_\Phi q$, and the constrained action, which determines a probability distribution, by $v \circ_\Phi q$. To simplify the notation, let $\psi(v_1 + v_2, q)\bar{1}$ be denoted by ζ. Then ζ is the unique constant vector such that $(v_1 + v_2 - \zeta) \odot_\Phi q$ is a probability distribution. Thus, by the homomorphism property for the unconstrained problem, proved above, we have that

$$\begin{aligned} (v_1 + v_2) \circ_\Phi q &= (v_1 + v_2 - \zeta) \odot_\Phi q \\ &= (v_1 - \zeta + \psi(v_2, q)\,\bar{1} + v_2 - \psi(v_2, q)\,\bar{1}) \odot_\Phi q \\ &= (v_1 - (\zeta - \psi(v_2, q)\,\bar{1})) \odot_\Phi ((v_2 - \psi(v_2, q)\,\bar{1}) \odot_\Phi q) \\ &= (v_1 - (\zeta - \psi(v_2, q)\,\bar{1})) \odot_\Phi (v_2 \circ_\Phi q)\,. \end{aligned}$$

As a result, $\psi(v_1, v_2 \circ_\Phi q) = \psi(v_1 + v_2, q) - \psi(v_2, q)$, proving the homomorphism property for the constrained problem. ▮

The normalizing term $\psi(v, q)$ will be referred to as the *cumulant generating function*, to be consistent with the standard terminology for exponential families. We now focus on a particular collection of Bregman divergences, which we call the *Bregman-Csiszár divergences* (or *BC divergences*). This collection will be seen to be easy to work with computationally, leading to learning algorithms that are as practical as logistic regression and boosting.

Let $\phi_\alpha(x)$ be the family of convex functions defined for $\alpha \in [0, 1]$ by

$$\phi_\alpha(x) = \begin{cases} x - \log x - 1 & \alpha = 0 \\ x \log x - x + 1 & \alpha = 1 \\ \frac{1}{\alpha(1-\alpha)}\left(-x^\alpha + \alpha x - \alpha + 1\right) & 0 < \alpha < 1. \end{cases}$$

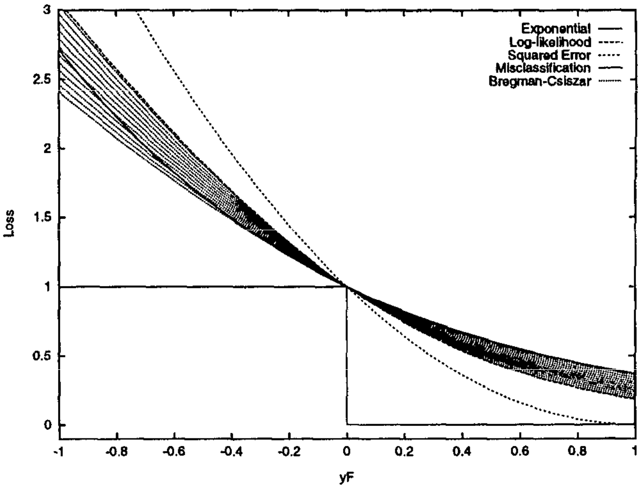

Figure 2. *Loss functions: log-likelihood, the exponential criterion, and Bregman-Csiszár divergences D_α, for several values of α in the interval $(\frac{3}{4}, 1)$, chosen in increments of $\Delta\alpha = 0.02$. The curves were generated by numerically estimating the cumulant generating functions $\psi(f, q)$ using Newton's method.*

The associated family of Bregman-Csiszár divergences $D_\alpha(p, q)$ on discrete distributions p, q is given by

$$D_\alpha(p,q) = \frac{1}{\alpha(1-\alpha)} \sum_y q^\alpha(y) - p^\alpha(y) + \alpha\, q(y)^{\alpha-1}(p(y) - q(y)).$$

For $\alpha = 0$ this specializes to the Itakura-Saito distortion

$$D_0(p,q) = \sum_y \log \frac{q(y)}{p(y)} + \frac{p(y)}{q(y)} - 1$$

and for $\alpha = 1$ it yields the extended Kullback divergence

$$D_1(p,q) = \sum_y p(y) \log \frac{p(y)}{q(y)} - p(y) + q(y)$$

Using the fact that $\lim_{\beta \to 0} \frac{x^\beta - 1}{\beta} = \log x$, it is simple to verify that this constitutes a continuous family of Bregman divergences. This family is given a strong axiomatic justification by Csiszár in [9]; however, it is out of the scope of the current paper to review this axiomatic formulation.

Using Lagrange multipliers, it is a simple calculation to determine the Legendre transform $f \circ_\alpha q$ associated with the Bregman-Csiszár divergence D_α, acting on the set of probability distributions Δ_r. This determines the family of additive models that we work with.

Proposition 3.3. *The Legendre transform $f \circ_\alpha q$ is given by*

$$(f \circ_\alpha q)(y) = \left(q(y)^{\alpha-1} + (\alpha - 1)(f(y) - \psi(f,q))\right)^{\frac{1}{\alpha-1}}$$

for $0 < \alpha < 1$. For $\alpha = 0, 1$ we have that

$$(f \circ_0 q)(y) = \frac{q(y)}{1 - q(y)(f(y) - \psi(f,q))}$$

and
$$(f \circ_1 q)(y) = q(y)e^{f(y)-\psi(f,q)}$$

Consider now conditional models $q(y \mid x)$ for the two-class case $y \in \{-1, +1\}$, and let $f(x,y) = -yf(x)$, as in the boosting and logistic regression models discussed earlier. We are interested in the behavior of the gain $D_\alpha(p,q) - D_\alpha(p, f \circ_\alpha q)$, as α varies. Fixing x, we want to consider how the loss varies as a function of $f(x)$. For $\alpha = 1$ we know that this is given by the scaled curve $\log\left(1 + e^{-2yf(x)}\right)$ plotted in Figure 1. For $\alpha < 1$ we cannot compute the normalization ψ explicitly, but it is easily computed numerically using Newton's method. The resulting family of loss functions is plotted in Figure 2 for several values of $\alpha \in (\frac{3}{4}, 1)$. The BC divergences very closely approximate the exponential criterion near the decision boundary when $\alpha \approx \frac{3}{4}$.

5. Duality and Information Geometry

In [12] an elementary proof was given of the fundamental duality between maximum likelihood and minimum divergence for exponential models. This simple proof extends to Bregman divergences quite easily. This fundamental fact allows us to view the same learning problem is two very different ways.

Let $f \in \mathbb{R}^r$ and let $q_0 \in \mathbb{R}^r$ be an initial measure. Define $\mathcal{P}$ and $\mathcal{Q}$ by

$$\begin{aligned}\mathcal{P} &= \{p \in \Delta_r \mid p[f] = \tilde{p}[f]\}\\ \mathcal{Q} &= \{q \in \Delta \mid q = (\lambda f) \circ_\Phi q_0 \text{ for some } \lambda \in \mathbb{R}\}.\end{aligned}$$

$\overline{\mathcal{Q}}$ will denote the closure of $\mathcal{Q}$ (in the Euclidean topology).

Theorem 4.1. *Suppose $D_\Phi(\tilde{p}, q_0) < \infty$. Then there exits a unique $q_\star \in \Delta$ satisfying*

(1) $q_\star \in \mathcal{P} \cap \overline{\mathcal{Q}}$

(2) $D_\Phi(p,q) = D_\Phi(p, q_\star) + D_\Phi(q_\star, q)$ *for any* $p \in \mathcal{P}$ *and* $q \in \overline{\mathcal{Q}}$

(3) $q_\star = \arg\min_{q \in \overline{\mathcal{Q}}} D_\Phi(\tilde{p}, q)$

(4) $q_\star = \arg\min_{p \in \mathcal{P}} D_\Phi(p, q_0)$.

Moreover, any of these four properties determines $q_\star$ uniquely.

Our proof of this duality relies on the following computational lemma for Bregman divergences. This same lemma will be used to derive the quadratic approximation to the gain.

Lemma 4.2. *Fix $f \in \mathbb{R}^r$, and consider the one-parameter family of models given by $q(t) = (tf) \circ_\alpha q$, with cumulant generating functions $\psi(t)$. Then*

(a) $\frac{d}{dt}\Big|_{t=0} \psi(t) = \dfrac{H_\Phi^{-1}(q) f \cdot \bar{1}}{H_\Phi^{-1}(q)\, \bar{1} \cdot \bar{1}} = E_{H_\Phi^{-1}(q)}[f]$

(b) $\frac{d}{dt}\Big|_{t=0} D_\Phi(p,\, q(t)) = f \cdot (q - p) = q[f] - p[f]$

(c) $\frac{d^2}{dt^2}\Big|_{t=0} D_\Phi(p,\, q(t)) = (f - \psi'(0)\,\bar{1})^\top\, H_\Phi^{-1}(q)\, (f - \psi'(0)\,\bar{1})$

where $H_\Phi(q)$ is the Hessian $\nabla^2 \Phi(q)$.

Proof. From the proof of Lemma 3.2 we have that

$$\begin{aligned} H_\Phi(q)\, \frac{d}{dt}\Big|_{t=0} & (tf) \circ_\alpha q = \\ &= \frac{d}{dt}\Big|_{t=0} \nabla\Phi((tf) \circ_\alpha q) \\ &= \frac{d}{dt}\Big|_{t=0} (\nabla\Phi(q) + tf - \psi(t)\,\bar{1}) \\ &= f - \psi'(0)\,\bar{1}. \end{aligned}$$

Equality a) now follows from $\frac{d}{dt}\left(q(t) \cdot \bar{1}\right) = \left(\frac{d}{dt} q(t)\right) \cdot \bar{1} = 0$. To derive the second equality, note that

$$\begin{aligned} \frac{d}{dt} & D_\Phi(p,\, q(t)) = \\ &= \frac{d}{dt}\left[\Phi(p) - \Phi(q(t)) - \nabla\Phi(q(t)) \cdot (p - q(t))\right] \\ &= -\left(\frac{d}{dt}\nabla\Phi(q(t))\right) \cdot (p - q(t)) \\ &= (f - \psi'(t)\,\bar{1}) \cdot (q(t) - p) \\ &= f \cdot (q(t) - p) \end{aligned}$$

where the last equality is a consequence of $q(t) \cdot \bar{1} = p \cdot \bar{1}$. Finally, using the above calculation we have that

$$\begin{aligned} \frac{d^2}{dt^2}\Big|_{t=0} & D_\Phi(p,\, q(t)) = \\ &= \frac{d}{dt}\Big|_{t=0} (f - \psi'(t)\,\bar{1}) \cdot (q(t) - p) \\ &= (f - \psi'(0)\,\bar{1}) \cdot \frac{d}{dt}\Big|_{t=0} q(t) \\ &= (f - \psi'(0)\,\bar{1})^\top\, H_\Phi^{-1}(q)\, (f - \psi'(0)\,\bar{1}) \end{aligned}$$

where the second equality follows from property a). ∎

The remainder of the proof of Theorem 4.1 follows the same steps as the proof for the Kullback-Leibler divergence given in [12].

As a very special case of duality, we can reinterpret the explanation of boosting as adjusting the weight of the current weak learner f_t so that it is uncorrelated with the labels y_i [23]. That is, at the t-th round AdaBoost adds the feature f_t with weight α to the additive model, so that the probability assigned to the i-th sample is given by

$$q_{t+1}(i) = \frac{1}{Z_t(\alpha)} e^{-\alpha y_i f_t(x_i)} q_t(i)$$

The weight α is chosen to minimize the exponential loss criterion $\sum_i e^{-\alpha y_i f_t(x_i)} q_t(i) = Z_t(\alpha)$. Using the notation $yf_t(x) = h_t(x,y)$, let $\tilde{p}(x,y)$ be any distribution satisfying $\tilde{p}[h_t] = 0$ which assigns nonzero probability to each of the training samples. Then by Theorem 4.1, we have that

$$\min_{p:\, p[h_t]=0} D(p \,\|\, q_t) = \min_\alpha D(\tilde{p} \,\|\, q_{t+1}) = \min \log Z_t(\alpha)$$

This relationship is further developed in [20].

6. Weighted Regression

The lemma from the previous section allows us to approximate the loss in Bregman divergence due to adding a feature to an additive model, using a quadratic approximation that extends the analysis given in Section 2. We will assume that the distributions are normalized, but similar results apply for positive measures that are not necessarily probability distributions.

Proposition 4.2. *For a Bregman divergence D_Φ, the quadratic approximation of the gain*

$$\mathcal{G}_\Phi(\tilde{p}; f, q) = \max_\lambda D_\Phi(p, q) - D_\Phi(p, (\lambda f) \circ_\Phi q)$$

is given by

$$\begin{aligned} \mathcal{G}_\Phi(\tilde{p}; f, q) &\approx \mathcal{G}^2_\Phi(\tilde{p}; f, q) \\ &\stackrel{\text{def}}{=} \frac{1}{2} \frac{(\tilde{p}[f] - q[f])^2}{(f - \psi'(0)\,\bar{1})^\top H_\Phi^{-1}(q)\,(f - \psi'(0)\,\bar{1})}\,. \end{aligned}$$

Proof. From Lemma 4.2 we have that to second order

$$\begin{aligned} D_\Phi(p, q) - D_\Phi(p, (\lambda f) \circ_\Phi q) &\approx \lambda(\tilde{p}[f] - q[f]) \\ &+ \frac{1}{2}\lambda^2 (f - \psi'(0)\,\bar{1})^\top H_\Phi^{-1}(q)\,(f - \psi'(0)\,\bar{1}) + O(\lambda^2) \end{aligned}$$

Maximizing with respect to λ yields the statement of the proposition. ▌

While the quadratic approximation gives the most direct correspondence to the results of [16], we note that a practical alternative for feature selection is to compute the *full* gain $\mathcal{G}_\Phi(\tilde{p}; f, q)$ using a generalized iterative scaling algorithm [8, 12, 14]. While this is more computationally demanding, it gives a more accurate assessment of the value of a feature. From the duality theorem, it also has an interpretation in terms of constrained optimization.

Now, for the Bregman-Csiszár divergences D_α, we have that $\phi''(x) = x^{\alpha-2}$. Thus $H^{-1}_{\phi_\alpha}(q)$ is a diagonal matrix with diagonal entries $q(y \mid x)^{2-\alpha}$. Applying Proposition 4.2 to the two-class case, a simple calculation shows that the Newton updates to select the features $f(x)$ result in the weighted least-squares criterion

$$\widehat{f}(x) = \arg\min_{f(x)} E_{w_\alpha(x)} \left(\frac{1}{2} \frac{y^* - q(x)}{w_\alpha(x)} - f(x) \right)^2$$

with weights

$$w_\alpha(x) = \frac{q(x)^{2-\alpha}\,(1 - q(x))^{2-\alpha}}{q(x)^{2-\alpha} + (1 - q(x))^{2-\alpha}}.$$

When $\alpha = 1$ this results in weights $w_1(x) = q(x)(1 - q(x))$, which are the same as those used in the LogitBoost algorithm. As α decreases, the weights place relatively more emphasis on the events for which the current model is less certain. A subset of this family of weights, for $0 < \alpha < 1$ is shown in Figure 3.

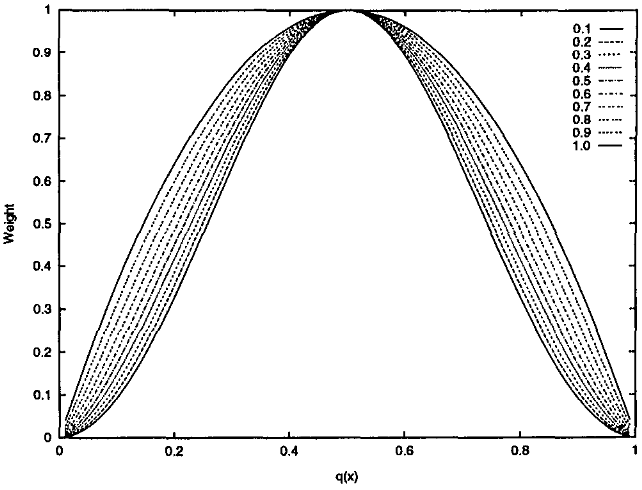

Figure 3. *Weights used in weighted least-squares regression derived from Bregman-Csiszár divergences, for several values of* $0 < \alpha < 1$, *chosen in increments of* $\Delta\alpha = \frac{1}{10}$.

7. Conclusions and Discussion

We have presented a statistical framework for building incremental or stepwise classification algorithms using generalized divergences. Our approach is based upon the use of Bregman divergences as a similarity measure between probability distributions, and uses the Legendre transform to define a class of additive models. The Bregman-Csiszár divergences D_α were shown to yield a one-parameter family of loss functions that includes the log-likelihood loss of logistic regression as a special case, and that closely approximates the exponential loss criterion used in the AdaBoost algorithms with confidence-rated predictions. We considered the gain due to adding a single feature to an additive model, and showed how this gain can be approximated to second order. This leads to a class of weighted stepwise regression procedures that includes the LogitBoost algorithm of Friedman, Hastie, and Tibshirani as a special case.

The class of Bregman divergences enjoys a number of very useful qualities, and this recent addition to the machine learning literature may have many further applications. As exploited by Warmuth *et al.* and Della Pietra *et al.*, these similarity measures have convexity properties that allow bounds and "auxiliary functions" to be easily derived [19, 17, 14]. Their use can often be given an interpretation in terms of a generalized maximum entropy principle [11], and the projection operators that are defined for Bregman divergences can be useful for proving convergence of various learning algorithms and constrained optimization procedures.

From a more practical and empirical standpoint, we believe that there are many ways of using these methods to design more effective and practical feature selection algorithms. We have worked extensively with a closely related set of techniques for incrementally adding features to a maximum entropy model, and have most recently applied these to the text segmentation problem [2]. When evaluated against the real AdaBoost algorithm using one- and three-level decision trees as weak learners, we found that the logistic regression (maximum entropy) methods result in significantly better performance [7]. One particular difference between our stepwise techniques and those used in the LogitBoost and AdaBoost algorithms is that after a new feature is included in the additive model, *all* of the model's feature weights are readjusted using a generalized iterative scaling algorithm [12, 14]. While this better accounts for the correlations between features, one might expect it to result in models that are more prone to overfitting than the corresponding algorithms that "freeze" each feature's weight after it is included in the model. In practice, however, we have observed that this approach is extremely robust to overfitting. We hope to gain a better understanding of this behavior through future research and experimentation.

The main result of the current paper is to show how Bregman divergences can be used to generalize and complement more standard statistical methods such as stepwise logistic regression. These observations build upon the correspondence between boosting and logistic regression that are established in [16], as well as on our experience using incremental maximum entropy and minimum divergence methods for a range of practical problems. The empirical success of boosting algorithms calls for a better understanding of its properties, and we believe that a statistical and information-theoretic perspective complements boosting's roots in error bounds and the PAC model of learning [22], and offers several advantages as well. While the classical techniques of logistic regression do not fully suffice to explain boosting, we have argued that Bregman divergences may enable the development of new variations on voting-style learning algorithms that may make such techniques even more effective in practice.

Appendix A: Statistics of the Gain

In this appendix we give proofs of the corollaries to Proposition 2.1, which describe the statistics of the quadratic approximation of the gain due to adding a single feature with small weight to the additive model. Similar results can be derived for general Bregman divergences.

Proposition 2.1. *To quadratic order, the gain $\mathcal{G}(f;\tilde{p},q)$ is approximated by*

$$\mathcal{G}(f;\tilde{p},q) \approx \mathcal{G}_2(f;\tilde{p},q) \stackrel{\text{def}}{=} \frac{1}{2}\frac{(q[f]-\tilde{p}[f])^2}{q\,[f^2-q[f\,|\,x]^2]}.$$

Corollary 2.2. *Let $\mathcal{Q}_{q,f}$ be the one-parameter family of exponential models given by*

$$\mathcal{Q}_{q,f} = \left\{ q_\lambda(y\,|\,x) = \frac{e^{\lambda f(x,y)}}{Z_\lambda(x)} q(y\,|\,x)\,;\ \lambda \in \mathbb{R} \right\}$$

and let $\pi(\lambda\,|\,\mathcal{Q}_{q,f})$ be the non-conjugate prior distribution on the parameter λ which is Gaussian with mean zero and variance $\frac{1}{\sqrt{N}} < \sigma^2 \leq \infty$. Then the log-probability of N events $d = \{(x_i,y_i)\}_{i=1}^N$ is given by

$$\begin{aligned} &\log p\,(d\,|\,\mathcal{Q}_{q,f}) = \\ &\quad = \log \int p(d\,|\,q_\lambda)\,\pi(\lambda\,|\,\mathcal{Q}_{q,f})\,d\lambda \\ &\quad = \log p(d\,|\,q) - N\mathcal{G}_2(f;\tilde{p},q) + O(\log N). \end{aligned}$$

Proof. From our definitions, we have that $p(d\,|\,q_\lambda) = e^{N\mathcal{L}(\tilde{p},q_\lambda)}$. Using the approximation in Proposition 2.1, this is approximately

$$p(d\,|\,q_\lambda) = p(d\,|\,q)e^{-a\,\lambda-\frac{1}{2}b\,\lambda^2}$$

where $a = N(q[f]-\tilde{p}[f])$ and $b = Nq[f^2-q[f\,|\,x]^2]$. Carrying out the integration, we find that

$$\begin{aligned} p(d\,|\,\mathcal{Q}_{q,f}) &= \frac{1}{\sqrt{2\pi}\sigma}\int p(d\,|\,q_\lambda)\,\pi(\lambda\,|\,\mathcal{Q}_{q,f})\,d\lambda \\ &= p(d\,|\,q)\frac{1}{\sqrt{2\pi}\sigma}\int e^{-a\lambda-\frac{1}{2}(b+\frac{1}{\sigma^2})\lambda^2}\,d\lambda \\ &= p(d\,|\,q)\,(b+\sigma^{-2})^{-\frac{1}{2}}\,exp\left(-\frac{a^2}{b+\sigma^{-2}}\right). \end{aligned}$$

Taking logarithms, we find that

$$\begin{aligned} &\log p(d\,|\,\mathcal{Q}_{q,f}) = \\ &\quad = \log p(d\,|\,q) \\ &\qquad - N\mathcal{G}(f;\tilde{p},q)\left(1+\frac{1}{N\sigma^2 q[f^2-q[f\,|\,x]^2]}\right) \\ &\qquad + O(\log N) \\ &\quad = \log p(d\,|\,q) - N\mathcal{G}(f;\tilde{p},q) + O(\log N) \end{aligned}$$

under the assumption that $\frac{1}{\sqrt{N}} < \sigma^2 \leq \infty$. ▌

Corollary 2.3. *Fixing $q(y\,|\,x)$, consider $\tilde{p}(x,y)$ to be the empirical distribution of a sample of size N drawn from the distribution $q(x,y) = \tilde{p}(x)\,q(y\,|\,x)$. Then the quadratic approximation is distributed as*

$$\mathcal{G}_2(f;\tilde{p},q) \sim \frac{\gamma}{2N}\chi^2$$

where

$$\gamma = \frac{q\,[(f-q[f])^2]}{q\,[(f-q[f\,|\,x])^2]}$$

and χ^2 is the chi-squared distribution with one degree of freedom.

Proof. Let $f_i = f(x_i,y_i)$ be the value of f on the i-th labelled example, and let

$$y_i = \frac{f_i - q[f]}{\sqrt{q\,[(f-q[f])^2]}}.$$

Then

$$\mathcal{G}_2(f;\tilde{p},q) = \frac{\gamma}{2N}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N y_i\right)^2.$$

Since the y_i are independent and identically distributed with mean zero and variance one, we have that $\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N y_i\right)^2 \sim \chi^2$ for large N. ▌

Acknowledgements

The author would like to thank Manfred Warmuth for several interesting discussions on Bregman divergences, and also Fernando Pereira, for inquiring several years ago about the relationship between multiplicative update algorithms and the incremental maximum entropy method. Stephen and Vincent Della Pietra made many essential contributions to this work.

References

[1] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory,* Wiley, New York, 1978.

[2] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, special issue on Natural Language Learning, **34**(1-3), pp. 177–210, 1999.

[3] A. Berger, V. Della Pietra, and S. Della Pietra, "A maximum entropy approach to natural language processing," Computational Linguistics, **22**, No. 1, pp. 39–71, 1996.

[4] L. Brown, *Fundamentals of Statistical Exponential Families*, Institute of Mathematical Statistics Lecture Notes–Monograph Series, Volume 9, Hayward, California, 1986.

[5] L.M. Bregman, "The relaxation method to find the common point of convex sets and its applications to the solution of problems in convex programming," USSR Computational Mathematics and Mathematical Physics, **7**, pp. 200–217, 1967.

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, MA, 1984.

[7] J. Carbonell, Y. Yang, J. Lafferty, R. Brown, T. Pierce, X. Liu, "CMU Report on TDT-2: Segmentation, detection and tracking," in Proceedings of the 1999 DARPA Broadcast News Conference.

[8] Y. Censor and A. Lent, "An iterative row-action method for interval convex programming," *J. Optim. Theory Appl.* **34**, pp. 321–353, 1981.

[9] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, **19**(4), pp. 2032–2066, 1991.

[10] I. Csiszár, "Generalized projections for non-negative functions," *Acta Math. Hungar.*, **68**(1-2), pp. 161–185, 1995.

[11] I. Csiszár, "Maxent, mathematics, and information theory," In *Maximum Entropy and Bayesian Methods*, K. Hanson and R. Silver, eds., Kluwer Academic Publishers, 1996.

[12] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. Pattern Analysis and Machine Intell.*, **19**(4), April, 1997, pp. 380–393.

[13] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Bregman distances, iterative scaling, and auxiliary functions," unpublished manuscript, 1995.

[14] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Statistical learning algorithms based on Bregman distances,"in *Proceedings of the Canadian Workshop on Information Theory*, Toronto, 1997.

[15] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," technical report, Department of Statistics, Stanford University, August 20, 1998.

[17] M. Herbster and M. Warmuth, "Tracking the best regressor," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, Madison, WI, 1998.

[18] E.T. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, R. Rosenkrantz, ed., D. Reidel Publishing Co., Dordrecht–Holland, 1983.

[19] J. Kivinen and M. Warmuth, "Additive versus exponentiated gradient updates for linear prediction," *Information and Computation,* **132**(1), pp. 1–64, 1997.

[20] J. Kivinen and M. Warmuth, "Boosting as entropy projection," in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, Santa Cruz, CA, 1999.

[21] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[22] R. Schapire, "The strength of weak learnability," *Machine Learning* **5**(2), pp. 197–227.

[23] R. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, Madison, WI, 1998.