Machine Learning, 43, 211–246, 2001

Relative Loss Bounds for On-Line Density Estimation with the Exponential Family of Distributions*

KATY S. AZOURY kazoury@sfsu.edu
College of Business, San Francisco State University, San Francisco, CA 94132, USA

M. K. WARMUTH manfred@cse.ucsc.edu
Computer Science Department, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

Editor: Yoram Singer

Received October 25, 1999; Revised October 25, 1999; Accepted March 6, 2000

Abstract. We consider on-line density estimation with a parameterized density from the exponential family. The on-line algorithm receives one example at a time and maintains a parameter that is essentially an average of the past examples. After receiving an example the algorithm incurs a loss, which is the negative log-likelihood of the example with respect to the current parameter of the algorithm. An off-line algorithm can choose the best parameter based on all the examples. We prove bounds on the additional total loss of the on-line algorithm over the total loss of the best off-line parameter. These relative loss bounds hold for an arbitrary sequence of examples. The goal is to design algorithms with the best possible relative loss bounds. We use a Bregman divergence to derive and analyze each algorithm. These divergences are relative entropies between two exponential distributions. We also use our methods to prove relative loss bounds for linear regression.

Keywords: relative loss bounds, worst-case loss bounds, on-line algorithms, exponential family of distributions, linear regression, Bregman divergences

1. Introduction

A main focus of statistical decision theory consists of the following: After receiving statistical information in the form of sampling data, the goal is to make decisions that minimize a loss or an expected loss with respect to an underlying distribution that is assumed to model the data. This distribution is often defined in terms of certain parameters. The statistical decisions depend on the specific values chosen for the parameters. Thus, for statistical decision theory, there are three important elements: parameters and the values they can take, decisions, and loss functions that evaluate the decisions.

In Bayesian statistical decision theory, a prior distribution on the parameters of the data distribution is an additional important element of information that is needed to assess decision performance. As a simple case, suppose we are given a sample of T data points

*An extended abstract appeared in UAI*99 (Azoury & Warmuth, 1999).

$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ (also referred to as T examples), and assume that the examples were independently generated by a Gaussian with an unknown mean and a known variance. One wants to find a parameter setting (here the mean) that minimizes the expected loss on a new example that is drawn from the same data distribution. In the Bayesian framework, a prior distribution on the mean would also enter into the decision making process.

In the context of learning theory, this setup, Bayesian or not, would be described as a batch or off-line learning model, since all the examples are given to the learner ahead of time and the decisions are made based on the information from the entire data set.

In this paper, we focus on on-line learning models. As with off-line learning, we have the same fundamental elements: parameters, decisions, and loss functions. The difference, however, is that the examples are given to the learner one at a time. Thus, on-line learning is naturally partitioned into trials, where in each trial one example is processed. A trial proceeds as follows. It begins with a current parameter setting (hypothesis). Then the next example is presented and the learner or the on-line algorithm incurs a loss. The loss is a function of this most recent example and the current parameter setting. Finally, the algorithm updates its parameter setting and a new trial begins.

In the context of on-line learning, the decisions are the parameter updates of the learner. The goal is to design on-line learning algorithms with good bounds on their total loss. Clearly, there cannot be meaningful bounds on the total loss of an on-line algorithm that stand alone and also hold for an arbitrary sequence of examples. However, in the on-line learning literature, a certain type of "relative" loss bound is desirable and has been used successfully. The term "relative" means that there is a comparison to the best parameter chosen off-line after seeing the whole batch of T examples. The parameter space is called the comparison class. The *relative loss bounds* quantify the additional total loss of the on-line algorithm over the total loss of the best off-line parameter (comparator). Since the on-line learner does not see the sequence of examples in advance, the additional loss (sometimes called regret) is the price of hiding the future examples from the learner.

In this paper we design and motivate on-line algorithms and their parameter updates so that they utilize the available information in the best possible manner and lead to good relative loss bounds. We focus on two types of learning problems: on-line density estimation and on-line regression.

For density estimation, the on-line algorithm receives a sequence of T unlabeled examples or data vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$. At the start of trial $t = 1, 2, 3, \ldots, T$, the learner has a current parameter setting $\boldsymbol{\theta}_t$ which is used to predict the next example $\boldsymbol{x}_t$. After making a prediction, the algorithm receives the example $\boldsymbol{x}_t$ and incurs a loss $L(\boldsymbol{\theta}_t, \boldsymbol{x}_t)$. Then the algorithm updates its parameter setting to $\boldsymbol{\theta}_{t+1}$. In contrast, on-line regression problems receive a labeled sequence of examples $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_T, y_T)$, where the $\boldsymbol{x}_t$ are called instances and the y_t are the labels. In trial t, the learner starts with the current parameter $\boldsymbol{\theta}_t$ and receives the instance $\boldsymbol{x}_t$. The learner then makes a prediction $\hat{y}_t$ for the label y_t. This prediction depends on $\boldsymbol{\theta}_t$ and $\boldsymbol{x}_t$. A loss $L(\boldsymbol{\theta}_t, (\boldsymbol{x}_t, y_t))$ is incurred and the parameter is updated to $\boldsymbol{\theta}_{t+1}$. For both types of problems, density estimation and regression, we use the abbreviated notation $L_t(\boldsymbol{\theta}_t)$ to denote the loss incurred in trial t. The subscript t indicates the dependence on the example presented in trial t. The total loss of the on-line algorithm is $\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t)$ and the total loss of the best off-line (batch) parameter $\boldsymbol{\theta}_B$ is

$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_B)$ plus a charge for the "size" of $\boldsymbol{\theta}_B$. Relative loss bounds give upper bounds on the difference between the two total losses.

We prove relative loss bounds for density estimation when the underlying model is a member of the exponential family of distributions and also for on-line linear regression. We consider two algorithms. The first algorithm is called the Incremental Off-line Algorithm. It predicts (i.e., chooses its parameter) as the best off-line algorithm would have predicted based on the examples seen so far. The second algorithm is called the Forward Algorithm. This algorithm is called "forward" because it uses a guess of a future example and the corresponding future loss when forming its prediction. It is motivated by Vovk's work (1997) on linear regression.

Our relative loss bounds for both algorithms grow logarithmically with the number of trials T. The motivation for the parameter updates has a Bayesian probabilistic interpretation. However, the relative loss bounds we prove hold for an arbitrary (or worst-case) sequence of examples. A key element in the design and analysis of the on-line learning algorithms is a generalized notion of distance called the Bregman divergence. This divergence can be interpreted as a relative entropy between two exponential distributions.

Outline

The rest of this paper is organized as follows. In Section 2, we present a brief overview of previous work. In Section 3, we define Bregman divergences and give the relevant background for the exponential family of distributions. We show that relative entropies between two exponential distributions (Amari, 1985) are special Bregman divergences. We conclude this section by listing some basic properties of Bregman divergences.

In Section 4, we introduce the Incremental Off-line Algorithm in a general setting and then apply this algorithm to the problem of density estimation with the exponential family and to linear regression. We give a number of relative loss bounds for specific examples.

In Section 5, we define and motivate the Forward Algorithm, which can be seen as a generalization of the Incremental Off-line Algorithm. Again, we apply this algorithm to density estimation with the exponential family. For the case of linear regression, we reprove the relative loss bounds obtained by Vovk (1997). Our proofs are more concise. An alternate simple proof for the "forward" linear regression algorithm is given in Forster (1999).

In Section 6, we briefly discuss an alternate method developed by Vovk for proving relative loss bounds that uses integration over a generalized posterior and discuss the advantages of our methods. Finally, we conclude (Section 7) with a discussion of a number of open problems.

2. Overview of previous work

The method of proving bounds on the additional total loss of an on-line algorithm over the total loss of the the best parameter in a comparison class essentially goes back to the work of the Blackwell (1956) and Hannnan (1957). They investigated such bounds in the context of game theory where the comparison class consists of all mixture strategies. Later

Cover (1991) proved such bounds in the context of mathematical finance. He used the comparison class of all constant rebalanced portfolios.

The research of this paper is rooted in the study of relative loss bounds for on-line learning algorithms of the computational learning theory community. Even though these bounds may underestimate the performance on natural data, they have been used as a powerful yardstick for analyzing and comparing on-line algorithms. In the computational learning theory community this line of research was initiated by Littlestone with the discovery of the Winnow algorithm (Littlestone, 1988). Littlestone also pioneered a style of amortized analysis for proving relative loss bounds which use certain divergence functions as potential functions. Winnow is designed for disjunctions as the comparison class and the total number of mistakes is used as the loss. The next wave of on-line algorithms were designed for a finite set of experts as a comparison class and a wide range of loss functions (Littlestone & Warmuth, 1994; Vovk, 1990; Cesa-Bianchi et al., 1997; Haussler, Kivinen & Warmuth, 1998). After that, algorithms were developed for the on-line linear least squares regression, i.e., when the comparison class consists of linear neurons (linear combination of experts) (Littlestone, Long & Warmuth, 1995; Cesa-Bianchi, Long & Warmuth, 1996; Kivinen & Warmuth, 1997). This work has been generalized to the case where the comparison class is the set of sigmoided linear neurons (Helmbold, Kivinen & Warmuth, 1995; Kivinen & Warmuth, 1998). Also starting with Littlestone's work, relative loss bounds for the comparison class of linear threshold functions have been investigated (Littlestone, 1988; Grove, Littlestone & Schuurmans, 1997).

All the on-line algorithms cited in the previous paragraph use fixed learning rates. In the simple settings the relative loss bounds do not grow with the number of trials. However, already for linear regression with the square loss, the relative loss bounds for algorithms with fixed learning rates grow with the square root of the loss of the best linear predictor (Cesa-Bianchi, Long & Warmuth, 1996; Kivinen & Warmuth, 1997) and the best loss is often linear in the number of trials T.

In contrast, our algorithms use a variable learning rate and our relative loss bounds grow logarithmically with T. Such bounds have been proven for a generalization of Bayes' Algorithm (Freund, 1996; Vovk, 1997; Xie & Barron, 1997; Yamanishi, 1998) which maintains a posterior on all parameters of the comparison class. We outline this method for proving relative loss bounds in Section 6. Bounds that grow logarithmically with T have also been proven previously for on-line linear regression (Foster, 1991; Vovk, 1997). An important insight we gained from this research is that $O(\log T)$ relative loss bounds seem to require the use of variable learning rates. In this paper, the learning rate applied in trial t is $O(1/t)$. The use of $O(1/t)$ learning rates for the exponential family was also suggested by Gordon (1999) as a possible strategy for leading to better bounds. However, no specific examples were worked out. In the case of linear regression, the $O(1/t)$ learning rates become inverses of the covariance matrix of the past examples.

General frameworks of on-line learning algorithms were developed in Grove, Littlestone and Schuurmans, (1997), Kivinen and Warmuth (1997; 1998), and Gordon (1999). We follow the philosophy of Kivinen and Warmuth (1997) of starting with a divergence function. From the divergence function we derive the on-line update and then use the same divergence as a potential in the amortized analysis. A similar method was developed in Grove,

Littlestone and Schuurmans (1997) for the case when the comparison class consists of linear threshold functions. They start with an update and construct the appropriate divergence that is used in the analysis.

Recently we have learned that the divergences used in on-line learning have been employed extensively in convex optimization and are called Bregman-distances (Bregman, 1967; Censor & Lent, 1981; Csiszar, 1991; Jones & Byrne, 1990). Bregman's method is to pick from a set of allowable models the one of minimal distance to the current model. In other words the current hypothesis is projected onto a convex set of allowable models. Some mild additional assumptions assure the uniqueness of the projections. With these assumptions a generalized Pythagorean Theorem can be proven for Bregman divergences (Bregman, 1967; Censor & Lent, 1981; Csiszar, 1991; Jones & Byrne, 1990; Herbster & Warmuth, 1998). The latter theorem often contradicts the triangular inequality and this is the reason why we use the term "divergence" instead of "distance".

Projections with respect to Bregman divergences have recently been applied in Herbster and Warmuth (1998) for the case when the off-line comparator is allowed to shift over time. The projections are used to keep the parameters of the algorithm in reasonable regions. This aids the recovery process when the underlying model shifts.

3. Bregman divergences and the exponential family

In this paper, we use a notion of divergence due to Bregman (1967) for deriving and analyzing on-line learning algorithms. These divergences can be interpreted as relative entropies between distributions from an exponential family. We begin this section by defining Bregman divergences and then review some important features of the exponential family of distributions that are relevant to this paper. We conclude this section with some properties of the divergences.

For an arbitrary real-valued convex and differentiable function $G(\boldsymbol{\theta})$ on the parameter space $\boldsymbol{\Theta} \subseteq \mathbf{R}^d$, the Bregman divergence between two parameters $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$ is defined as

$$\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) := G(\tilde{\boldsymbol{\theta}}) - G(\boldsymbol{\theta}) - (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}). \tag{3.1}$$

Here $\nabla_{\boldsymbol{\theta}}$ denotes the gradient with respect to $\boldsymbol{\theta}$. Throughout the paper, all vectors are column vectors and we use "$\cdot$" to denote the dot product between vectors.

Note that the Bregman divergence $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ equals $G(\tilde{\boldsymbol{\theta}})$ minus the first two terms of the Taylor expansion of $G(\tilde{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$. In other words, $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ is the tail of the Taylor expansion of $G(\tilde{\boldsymbol{\theta}})$ beyond the linear term. Since $G(\boldsymbol{\theta})$ is convex, $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \geq 0$. More properties will be listed in Section 3.4.

For example, let the parameter space be $\boldsymbol{\Theta} = \mathbf{R}^d$ and let $G(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta} \cdot \boldsymbol{\theta}$. In this case the Bregman divergence becomes the squared Euclidean distance, i.e.,

$$\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \tfrac{1}{2}\tilde{\boldsymbol{\theta}} \cdot \tilde{\boldsymbol{\theta}} - \tfrac{1}{2}\boldsymbol{\theta} \cdot \boldsymbol{\theta} - (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \cdot \boldsymbol{\theta} = \tfrac{1}{2}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2.$$

Also if $\Theta = \mathbf{R}$ and $G(\theta) = \ln(1 + e^{\theta})$, then

$$\Delta_G(\tilde{\theta}, \theta) = e^{\tilde{\theta}} - e^{\theta} - (\tilde{\theta} - \theta)\frac{e^{\theta}}{1 + e^{\theta}}.$$

3.1. *The exponential family*

The features of the exponential family that are used throughout this paper include a measure of divergence between two members of the family and an intrinsic duality relationship. See Barndorff-Nielsen (1978) and Amari (1985) for a more comprehensive treatment of the exponential family.

A multivariate parametric family $\mathcal{F}_G$ of distributions is said to be an exponential family when its members have a density function of the form

$$P_G(\boldsymbol{x} \mid \boldsymbol{\theta}) = \exp(\boldsymbol{\theta} \cdot \boldsymbol{x} - G(\boldsymbol{\theta}))P_0(\boldsymbol{x}), \tag{3.2}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are vectors in R^d, and $P_0(\boldsymbol{x})$ represents any factor of the density which does not depend on $\boldsymbol{\theta}$.

The d-dimensional parameter $\boldsymbol{\theta}$ is usually called the *natural* (or *canonical*) parameter. Many common parametric distributions are members of this family, including the Gaussian. The function $G(\boldsymbol{\theta})$ is a normalization factor defined by

$$G(\boldsymbol{\theta}) = \ln \int \exp(\boldsymbol{\theta} \cdot \boldsymbol{x}) P_0(\boldsymbol{x})\, d\boldsymbol{x}.$$

The space $\Theta \subseteq R^d$, for which the integral above is finite, is called the *natural* parameter space. The exponential family is called *regular* if Θ is an open subset of R^d. It is well known (Barndorff-Nielsen, 1978; Amari, 1985) that Θ is a convex set, and that $G(\boldsymbol{\theta})$ is a strictly convex function on Θ. The function $G(\boldsymbol{\theta})$ is called the *cumulant function*, and it plays a fundamental role in characterizing members of this family of distributions.

We use $g(\boldsymbol{\theta})$ to denote the gradient $\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$ and $\nabla^2_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$ to denote the Hessian of $G(\boldsymbol{\theta})$. Let

$$\lambda_G(\boldsymbol{\theta}; \boldsymbol{x}) = \ln P_G(\boldsymbol{x} \mid \boldsymbol{\theta})$$

represent the log-likelihood which is viewed as a function of $\boldsymbol{\theta}$. Under some standard regularity conditions, log-likelihood functions satisfy well-known moment identities (McCullagh & Nelder, 1989). Applying these identities to the exponential family reveals the special role played by the cumulant function $G(\boldsymbol{\theta})$.

Let $E_{\boldsymbol{\theta}}(.)$ be the expectation with respect to the distribution $P_G(\boldsymbol{x} \mid \boldsymbol{\theta})$. The first moment identity of log-likelihood functions is

$$E_{\boldsymbol{\theta}}(\nabla_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x})) = \mathbf{0}. \tag{3.3}$$

The gradient of the log-likelihood in (3.2) is linear in $\boldsymbol{x}$:

$$\nabla_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x}) = \boldsymbol{x} - \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}). \tag{3.4}$$

Applying (3.3), we get

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) = g(\boldsymbol{\theta}).$$

This shows that the mean of $\boldsymbol{x}$ is equal to the gradient of $G(\boldsymbol{\theta})$. We let $g(\boldsymbol{\theta}) = \boldsymbol{\mu}$, and call $\boldsymbol{\mu}$ the expectation parameter.

Since the cumulant function G is strictly convex, the map $g(\boldsymbol{\theta}) = \boldsymbol{\mu}$ has an inverse. We denote the image of $\boldsymbol{\Theta}$ under the map $g(.)$ by $\mathbf{M}$ and the inverse map from $\mathbf{M}$ to $\boldsymbol{\Theta}$ by $g^{-1}(\boldsymbol{\mu}) = \boldsymbol{\theta}$. The set $\mathbf{M}$ is called the expectation space, which may not necessarily be a convex set.

The second moment identity for log-likelihood functions is

$$E_{\boldsymbol{\theta}}((\nabla_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x}))(\nabla_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x}))') = -E_{\boldsymbol{\theta}}\big(\nabla^2_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x})\big), \tag{3.5}$$

where $'$ denotes the transpose. For the exponential family we have

$$E_{\boldsymbol{\theta}}((\nabla_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x}))(\nabla_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x}))') = E_{\boldsymbol{\theta}}((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})')$$
$$\text{and} - \nabla^2_{\boldsymbol{\theta}} \lambda_G(\boldsymbol{\theta}; \boldsymbol{x}) = \nabla^2_{\boldsymbol{\theta}} G(\boldsymbol{\theta}).$$

Thus, by the second moment identity, the variance-covariance matrix $\mathrm{Var}(\boldsymbol{x} \mid \boldsymbol{\theta})$ for $\boldsymbol{x}$ is the Hessian of the cumulant function $G(\boldsymbol{\theta})$ (also called the *Fisher Information Matrix*). Since $G(\boldsymbol{\theta})$ is strictly convex, this Hessian is symmetric positive definite.

3.2. Duality between the natural parameters and the expectation parameters

Sometimes, it is more convenient to parameterize a distribution in the exponential family by using its expectation parameter $\boldsymbol{\mu}$ instead of its natural parameter $\boldsymbol{\theta}$. This pair of parameterizations have a dual relationship. We provide the aspects of the duality that are relevant to this paper. First, define a second function on the range $\mathbf{M}$ of g as follows:

$$F(\boldsymbol{\mu}) := \boldsymbol{\theta} \cdot \boldsymbol{\mu} - G(\boldsymbol{\theta}) \tag{3.6}$$

Let $f(\boldsymbol{\mu}) := \nabla_{\mu} F(\boldsymbol{\mu})$ denote the gradient of $F(\boldsymbol{\mu})$.

Note that by taking the gradient of $F(\boldsymbol{\mu})$ in (3.6) with respect to $\boldsymbol{\mu}$ and treating $\boldsymbol{\theta}$ as a function of $\boldsymbol{\mu}$, we get

$$\nabla_{\mu} F(\boldsymbol{\mu}) = \boldsymbol{\theta} + (\nabla_{\mu} \boldsymbol{\theta})\boldsymbol{\mu} - (\nabla_{\mu} \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) = \boldsymbol{\theta}, \tag{3.7}$$

where $\nabla_{\mu} \boldsymbol{\theta}$ is the Jacobian of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\mu}$.

Thus, $f(\boldsymbol{\mu})$ is the inverse map $g^{-1}(\boldsymbol{\mu})$ and the two parameterizations $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ are related by the following transformations

$$g(\boldsymbol{\theta}) = \boldsymbol{\mu} \quad \text{and} \quad f(\boldsymbol{\mu}) = \boldsymbol{\theta}. \tag{3.8}$$

Since $g(\boldsymbol{\theta})$ has a positive definite Jacobian for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $g^{-1}(\boldsymbol{\mu})$ has a positive definite Jacobian for all $\boldsymbol{\mu} \in \mathbf{M}$. Thus, the second function $F(\boldsymbol{\mu})$ is strictly convex as well. This

function is called the *dual* of $G(\boldsymbol{\theta})$ (Amari, 1985). Furthermore, $F(\boldsymbol{\mu})$ is the negative entropy of $P_G(\boldsymbol{x} \mid \boldsymbol{\theta})$ with respect to the reference measure $P_0(\boldsymbol{x})$, i.e.,

$$F(\boldsymbol{\mu}) = -E_{\boldsymbol{\theta}}\left(- \ln \frac{P_G(\boldsymbol{x} \mid \boldsymbol{\theta})}{P_0(\boldsymbol{x})} \right). \tag{3.9}$$

It follows from (3.8) that the Hessian of $F(\boldsymbol{\mu})$ is the inverse of the Fisher Information Matrix, i.e., $\nabla^2_{\boldsymbol{\mu}} F(\boldsymbol{\mu}) = (\nabla^2_{\boldsymbol{\theta}} G(\boldsymbol{\theta}))^{-1}$. Now consider the function $V(\boldsymbol{\mu}) = \nabla^2_{\boldsymbol{\theta}} G(f(\boldsymbol{\mu}))$, which is the Fisher Information Matrix expressed in terms of the expectation parameter. This function is defined on the expectation space $\mathbf{M}$, takes values in the space of symmetric $d \times d$ matrices, and is called the *variance function*. The variance function plays an important role in characterizing members in the exponential family (Morris, 1982; Gutiérrez-Peña & Smith, 1995). The matrix $V(\boldsymbol{\mu})$ is positive definite for all $\boldsymbol{\mu} \in \mathbf{M}$, and $V(\boldsymbol{\mu}) = (\nabla^2_{\boldsymbol{\mu}} F(\boldsymbol{\mu}))^{-1}$. Thus, in the context of exponential families the functions F and G are not arbitrary convex functions but must have positive definite Hessians.

3.3. *Divergence between two exponential distributions*

Consider two distributions $P_G(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}})$ with an old parameter setting $\tilde{\boldsymbol{\theta}}$, and $P_G(\boldsymbol{x} \mid \boldsymbol{\theta})$ with a new parameter setting $\boldsymbol{\theta}$. Following Amari (1985) one may see the exponential family $\mathcal{F}_G$ as a manifold. The parameters $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ represent two points on this manifold. Several measures of distance (divergence) between these two points have been proposed in the literature. Amari introduced α divergences (Amari, 1985), and other related "distances" were introduced by Csiszar (1991) known as f divergences (we use the letter h below). Also Chernoff distances and Renyi's α information are related (Amari, 1985). These divergences all have the following general form:

$$E_{\tilde{\theta}}\, h\left(\frac{P_G(\boldsymbol{x} \mid \boldsymbol{\theta})}{P_G(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}})} \right),$$

where $h(\cdot)$ is some continuous convex function.

Our main choice for h is $h(z) = -\ln z$, which gives the relative entropy

$$E_{\tilde{\theta}}\, \ln \frac{P_G(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}})}{P_G(\boldsymbol{x} \mid \boldsymbol{\theta})} = \Delta_G(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}).$$

Another interesting choice is $h(z) = z \ln z$, which gives the "opposite" relative entropy:

$$E_{\boldsymbol{\theta}}\, \ln \frac{P_G(\boldsymbol{x} \mid \boldsymbol{\theta})}{P_G(\boldsymbol{x} \mid \tilde{\boldsymbol{\theta}})} = \Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}).$$

These two entropies are, respectively, called -1 and $+1$ divergences by Amari (1985).

3.4. *Properties of divergences*

In this section we give some simple properties of the divergences. For these properties we do not need that $G(\boldsymbol{\theta})$ has a positive definite Hessian. Hence the properties hold for the more general definition of Bregman divergence (3.1) where we allow $G(\boldsymbol{\theta})$ to be an arbitrary real-valued differentiable convex function $G(\boldsymbol{\theta})$ on the parameter space $\boldsymbol{\Theta}$.

Throughout the paper we use $\boldsymbol{\mu}$ to represent $\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$.

1. $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ is convex in its first argument since $G(\tilde{\boldsymbol{\theta}})$ is convex.
2. $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \geq 0$ and if $G(\boldsymbol{\theta})$ is strictly convex then equality holds iff $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$.
3. The gradient of the divergence with respect to the first argument has the following simple form

$$\nabla_{\tilde{\boldsymbol{\theta}}} \Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}.$$

4. Divergences are usually not symmetric, i.e., $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \neq \Delta_G(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$.
5. The divergence is a linear operator, i.e.,

$$\Delta_{G+H}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) + \Delta_H(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$$
$$\text{and } \Delta_{aG}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = a\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}), \quad \text{for } a \geq 0.$$

6. The divergence is not affected by adding a linear term to $G(\boldsymbol{\theta})$:

$$\text{if } G(\boldsymbol{\theta}) - H(\boldsymbol{\theta}) = a\boldsymbol{\theta} + \boldsymbol{b}, \quad \text{for } a \in \mathbf{R} \quad \text{and} \quad \boldsymbol{b} \in \mathbf{R}^d,$$
$$\text{then } \Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \Delta_H(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}).$$

7. For any $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$,

$$\Delta_G(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \Delta_G(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3) = \Delta_G(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3) + (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \cdot (\boldsymbol{\mu}_3 - \boldsymbol{\mu}_2).$$

   The dot product can usually have any sign. When it is negative then the above contradicts the triangular inequality. The case when the dot product is zero is exploited in the proof of the generalization of the Pythagorean Theorem to Bregman divergences (See for example Censor & Lent, 1981; Herbster, 1998).
8. If $G(\boldsymbol{\theta})$ is strictly convex, then the definition of the dual convex function $F(\boldsymbol{\mu})$ and the parameter transformations still hold (3.6 – 3.8) and Bregman divergences have the following duality property:

$$\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = G(\tilde{\boldsymbol{\theta}}) + F(\boldsymbol{\mu}) - \tilde{\boldsymbol{\theta}} \cdot \boldsymbol{\mu} = \Delta_F(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}).$$

The first six properties are immediate. Property 7 is proven in the appendix. This property was first used in Warmuth and Jagota (1998) for proving relative loss bounds. The last property follows from the definition of the dual function $F(\boldsymbol{\mu})$ (also called convex conjugate (Rockafellar, 1970). Note that the order of the arguments in $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ and $\Delta_F(\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}})$ is switched. In this paper, we only need Property 8 for the case when $G(\boldsymbol{\theta})$ is strictly convex.

However, for any real-valued differentiable convex function $G(\boldsymbol{\theta})$ one can define the dual function as $F(\boldsymbol{\mu}) = \boldsymbol{\theta} \cdot \boldsymbol{\mu} - G(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is any parameter in $\boldsymbol{\Theta}$ such that $\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) = \boldsymbol{\mu}$ (Rockafellar, 1970). With this definition Property 8 still holds.

We note that Gordon (1999) gives an elegant generalization of Bregman divergences to the case when the convex function $G(\boldsymbol{\theta})$ is not necessarily differentiable. For the sake of simplicity we restrict ourselves to the differentiable case in this paper.

Finally, when $G(\boldsymbol{\theta})$ is differentiable, the Bregman divergence can also be written as a path integral:

$$\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \int_{\boldsymbol{\theta}}^{\tilde{\boldsymbol{\theta}}} (g(\mathbf{r}) - g(\boldsymbol{\theta})) \cdot d\mathbf{r}.$$

This integral version of the divergence has been used to define a notion of a convex loss "matching" the increasing transfer function $g(\boldsymbol{\theta})$ of an artificial neuron (Auer, Herbster & Warmuth, 1995; Helmbold, Kivinen & Warmuth, 1995; Kivinen & Warmuth, 1998).

4. The incremental off-line algorithm

In this section we give our most basic algorithm and show how to prove relative loss bounds in a general setting. Learning proceeds in trials $t = 1, \ldots, T$. In each trial t an example is processed. For density estimation, the examples are data vectors $\boldsymbol{x}_t$ from some domain $\mathcal{X}$. In the regression setting, the t-th example consists of an instance $\boldsymbol{x}_t$ from some *instance domain* $\mathcal{X}$ and a label y_t from some *label domain* $\mathcal{Y}$.

The setup for a learning problem is defined by three parts. A parameter space $\boldsymbol{\Theta} \subseteq \mathbf{R}^d$, a real-valued loss function and a divergence function that is a measure of "distance" to an initial parameter setting. The parameter space $\boldsymbol{\Theta}$ represents the models to which the algorithms are compared. The loss of parameter vector $\boldsymbol{\theta}$ on the $t-$th example is denoted by $L_t(\boldsymbol{\theta})$ and $L_{1..t}(\boldsymbol{\theta})$ is shorthand for $\sum_{q=1}^{t} L_q(\boldsymbol{\theta})$. Usually losses are non-negative. The third component of the setup is an initial parmeter $\boldsymbol{\theta}_0$ and a Bregman divergence $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ to the initial parameter. The initial parameter $\boldsymbol{\theta}_0$ may be interpreted as a summary of any prior learning and the divergence $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ represents a measure of "distance" to the initial parameter.

The off-line (or batch) algorithm sees all T examples at once and it sets its parameter to

$$\boldsymbol{\theta}_B = \operatorname{argmin}_{\boldsymbol{\theta}} U_{T+1}(\boldsymbol{\theta}),$$
$$\text{where } U_{T+1}(\boldsymbol{\theta}) = \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..T}(\boldsymbol{\theta}).$$

Assumptions: The losses $L_t(\boldsymbol{\theta})$ (for $1 \leq t \leq T$) and $U_0(\boldsymbol{\theta})$ are differentiable and convex functions from the parameter space $\boldsymbol{\Theta}$ to the reals. Furthermore, we assume that $\operatorname{argmin}_{\boldsymbol{\theta}} U_{T+1}(\boldsymbol{\theta})$ always has a solution in $\boldsymbol{\Theta}$.

Note that this off-line algorithm trades the total loss on the examples against closeness to the original parameter. Alternatively the divergence $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ may be interpreted as the "size" of parameter $\boldsymbol{\theta}$. With this interpretation, the off-line algorithm finds a parameter that minimizes the sum of size and total loss.

The on-line algorithm sees one example at a time according to the following protocol:

On-line protocol of the Incremental Off-line Algorithm
Initial hypothesis is $\boldsymbol{\theta}_0$.
Set $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$.
For $t = 1$ to T do
  Predict with $\boldsymbol{\theta}_t$.
  Get t-th example.
  Incur loss $L_t(\boldsymbol{\theta}_t)$.
  Update hypothesis $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$.

The goal of the on-line algorithm is to incur a loss that is never too much larger than the loss of the off-line algorithm which sees all examples at once. At the end of trial t the on-line algorithm knows the first t examples and expects to see the next example. One reasonable and desirable setup for the parameter update at this point is to make the on-line algorithm do exactly what an off-line algorithm would have done after seeing t examples. We use the name *Incremental Off-line* for the on-line algorithm with this property.

The Incremental Off-line Algorithm

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta}), \text{ for } 0 \le t \le T,$$
$$\text{where } U_{t+1}(\boldsymbol{\theta}) = \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..t}(\boldsymbol{\theta}). \tag{4.1}$$

Additional assumptions: Here we assume that the $\operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta})$ (for $1 \le t \le T$) always have a solution in $\boldsymbol{\Theta}$.

If there is more than one solution for $\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta})$, then this is interpreted as $\boldsymbol{\theta}_{t+1} \in \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta})$. In the learning problems that we use as examples in this paper, $U_{t+1}(\boldsymbol{\theta})$ is typically strictly convex and so there is only one solution. Note that the final parameter $\boldsymbol{\theta}_{T+1}$ of the Incremental Off-line Algorithm coincides with the parameter $\boldsymbol{\theta}_B$ chosen by the batch algorithm.

When $t = 0$, then by (4.1) $\boldsymbol{\theta}_1 = \operatorname{argmin}_{\boldsymbol{\theta}} \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \boldsymbol{\theta}_0$, which is consistent with the protocol given above. While not necessary here, we begin the indexing of $\boldsymbol{\theta}_t$ at $t = 0$ to parallel the indexing of a second on-line algorithm given in the next section. This second algorithm is called the Forward Algorithm because it uses a guess of the next loss when updating the parameter.

The setup for the update in (4.1) does not seem truly on-line since it needs all the previous t examples. A truly on-line, yet equivalent setup, is given by the following lemma:

Lemma 4.1. *For the Incremental Off-line Algorithm and* $1 \le t \le T$,

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} \left(\Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta})\right).$$

Proof: Note that since $\nabla U_t(\boldsymbol{\theta}_t) = \mathbf{0}$ and $U_{t+1}(\boldsymbol{\theta}) = U_t(\boldsymbol{\theta}) + L_t(\boldsymbol{\theta})$,

$$\begin{aligned}\Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta}) &= U_t(\boldsymbol{\theta}) - U_t(\boldsymbol{\theta}_t) - (\boldsymbol{\theta} - \boldsymbol{\theta}_t) \cdot \nabla U_t(\boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta}) \\ &= U_{t+1}(\boldsymbol{\theta}) - U_t(\boldsymbol{\theta}_t).\end{aligned}$$

Thus, since $U_t(\boldsymbol{\theta}_t)$ is a constant, the argmin for $\boldsymbol{\theta}_{t+1}$ used in the definition (4.1) is the same as the argmin of the lemma. □

We are now ready to show the key lemma for the Incremental Off-line Algorithm. In this lemma we compare the total loss of the on-line algorithm to the total loss of any comparator $\boldsymbol{\theta}$, where the total loss of the comparator includes the divergence term $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$.

Lemma 4.2. *For the Incremental Off-line Algorithm, any sequence of T examples, and any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - (\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..T}(\boldsymbol{\theta})) = \sum_{t=1}^{T} \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \Delta_{U_{T+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{T+1}).$$

Proof: For $0 \le t \le T$, we expand the divergence $\Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1})$ and use $\nabla U_{t+1}(\boldsymbol{\theta}_{t+1}) = \mathbf{0}$. This gives us

$$\Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1}) = U_{t+1}(\boldsymbol{\theta}) - U_{t+1}(\boldsymbol{\theta}_{t+1}). \tag{4.2}$$

Since $U_{t+1}(\boldsymbol{\theta}) = U_t(\boldsymbol{\theta}) + L_t(\boldsymbol{\theta})$, for $1 \le t \le T$, we obtain

$$L_t(\boldsymbol{\theta}) = \Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1}) + U_{t+1}(\boldsymbol{\theta}_{t+1}) - U_t(\boldsymbol{\theta}). \tag{4.3}$$

For the special case of $\boldsymbol{\theta} = \boldsymbol{\theta}_t$, we get:

$$L_t(\boldsymbol{\theta}_t) = \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) + U_{t+1}(\boldsymbol{\theta}_{t+1}) - U_t(\boldsymbol{\theta}_t). \tag{4.4}$$

Subtracting (4.3) from (4.4) and applying $U_t(\boldsymbol{\theta}) - U_t(\boldsymbol{\theta}_t) = \Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ (a version of (4.2)) gives

$$L_t(\boldsymbol{\theta}_t) - L_t(\boldsymbol{\theta}) = \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1}) + \Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t).$$

By summing the above over all T trials we obtain:

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - L_{1..T}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \Delta_{U_{T+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{T+1}) + \Delta_{U_1}(\boldsymbol{\theta}, \boldsymbol{\theta}_1).$$

The lemma now follows from the equality $\Delta_{U_1}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) = \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$. This equality follows from Property 6, because $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$, and because $U_1(\boldsymbol{\theta}) - U_0(\boldsymbol{\theta}) = \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - U_0(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$. □

To obtain relative loss bounds we choose the best off-line parameter $\boldsymbol{\theta}_B$ as the comparator $\boldsymbol{\theta}$ and bound the right-hand-side of the equation of the lemma. Note that in case of the Incremental Off-line Algorithm, $\boldsymbol{\theta}_{T+1} = \boldsymbol{\theta}_B$, and thus the last divergence on the right-hand-side is zero. The divergence $\Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$ represents the cost of the update of $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$ incurred by the on-line algorithm. Relative loss bounds are bounds on the total cost $\sum_{t=1}^{T} \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})$ of the on-line updates.

4.1. *Incremental off-line algorithm for the exponential family*

We now apply the Incremental Off-line Algorithm to the problem of density estimation for the exponential family of distributions. We first give a general treatment and then prove relative loss bounds for specific members of the family in the subsections that follow.

We make the most obvious choice for a loss function, namely, the negative log-likelihood. So using the general form of the log-likelihood (3.2), the loss of parameter $\boldsymbol{\theta}$ on the example $\boldsymbol{x}_t$ is

$$L_t(\boldsymbol{\theta}) = -\ln P_G(\boldsymbol{x}_t \mid \boldsymbol{\theta}) = G(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \boldsymbol{x}_t - \ln P_0(\boldsymbol{x}_t).$$

For the purpose of the relative loss bounds (see Lemma 4.2) changing the loss by a constant that does not depend on $\boldsymbol{\theta}$ is inconsequential. Thus, the form of the reference measure $P_0(\boldsymbol{x})$ is immaterial.

As before, we allow the algorithm to have an initial parameter value at $\boldsymbol{\theta}_0$ and choose $U_0(\boldsymbol{\theta})$ as a multiple of the cumulant function, i.e., $U_0(\boldsymbol{\theta}) = \eta_0^{-1} G(\boldsymbol{\theta})$, where $\eta_0^{-1} \geq 0$. Thus, in the context of density estimation with the exponential family, the Incremental Off-line Algorithm (4.1) becomes

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta}), \quad \text{where } U_{t+1}(\boldsymbol{\theta}) = \eta_0^{-1} \Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..t}(\boldsymbol{\theta}). \tag{4.5}$$

Throughout the paper we use the notation η^{-1} to denote trade-off parameters. This has two reasons. First, the inverse of the trade-off parameters will become the learning rates of the algorithms and learning rates are commonly denoted by η. Also, we use η^{-1} instead of $1/\eta$, because in linear regression the η parameters are generalized to matrices.

The setup (4.5) can be interpreted as finding a maximum a-posteriori (MAP) parameter where the divergence term corresponds to the conjugate prior and $\eta_0^{-1} > 0$ is a hyper parameter. When $\eta_0^{-1} = 0$, then the divergence term disappears and we have maximum likelihood estimation. Alternatively one can think of $\boldsymbol{\theta}_0$ an initial parameter estimate based on some hypothetical examples seen before the first real example $\boldsymbol{x}_1$ and η_0^{-1} as the number of those examples. Also one can interpret the parameter η_0^{-1} as a trade-off parameter between staying close to the initial parameter $\boldsymbol{\theta}_0$ and minimizing the loss on the t examples seen by the end of trial t.

Yet another interpretation of (4.5) follows from rewriting $U_{t+1}(\boldsymbol{\theta})$ as

$$\left(\eta_0^{-1} + t\right) G(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \left(\eta_0^{-1} \boldsymbol{\mu}_0 + \sum_{q=1}^{t} \boldsymbol{x}_q \right) + \text{const.}$$

Thus, $U_{t+1}(\boldsymbol{\theta})$ corresponds to the negative log-likelihood of an exponential density with the cumulant function $(\eta_0^{-1} + t) G(\boldsymbol{\theta})$ and the example is $\eta_0^{-1} \boldsymbol{\mu}_0 + \sum_{q=1}^{t} \boldsymbol{x}_q$.

We now develop the alternate on-line motivation given in Lemma 4.1. Let $\eta_{t-1}^{-1} = \eta_0^{-1} + t - 1$. Since $U_t(\boldsymbol{\theta}) - \eta_{t-1}^{-1} G(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, it follows from the properties of the divergences that $\Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \Delta_{\eta_{t-1}^{-1} G}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \eta_{t-1}^{-1} \Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$. Thus, the on-line motivation of

Lemma 4.1 becomes

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}}\big(\eta_{t-1}^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta})\big). \tag{4.6}$$

Now the divergence measures the distance to the last parameter and the trade-off parameter is η_{t-1}^{-1}.

The updated parameter $\boldsymbol{\theta}_{t+1}$ can be obtained by minimizing $U_{t+1}(\boldsymbol{\theta})$ (as defined in (4.5)) which is a strictly convex function in $\boldsymbol{\theta}$. The gradient of this function in terms of the expectation parameters is

$$\eta_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \sum_{q=1}^{t}(\boldsymbol{\mu} - \boldsymbol{x}_q).$$

Setting the above to zero, for $0 \le t \le T$, gives the update of the expectation parameter of the Incremental Off-line Algorithm:

$$\boldsymbol{\mu}_{t+1} = \eta_t\left(\eta_0^{-1}\boldsymbol{\mu}_0 + \sum_{q=1}^{t}\boldsymbol{x}_q\right), \quad \text{where } \eta_t = \big(\eta_0^{-1} + t\big)^{-1}. \tag{4.7}$$

For $1 \le t \le T$, we can also express $\boldsymbol{\mu}_{t+1}$ as a convex combination of $\boldsymbol{\mu}_t$ and the last instance $\boldsymbol{x}_t$:

$$\boldsymbol{\mu}_{t+1} = \eta_t\eta_{t-1}^{-1}\boldsymbol{\mu}_t + \eta_t\boldsymbol{x}_t. \tag{4.8}$$

Note that $\eta_t\eta_{t-1}^{-1} + \eta_t = 1$. Alternate recursive forms of the update, that are used later on, are (for $1 \le t \le T$):

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \eta_{t-1}(\boldsymbol{\mu}_{t+1} - \boldsymbol{x}_t), \tag{4.9}$$

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \eta_t(\boldsymbol{\mu}_t - \boldsymbol{x}_t). \tag{4.10}$$

Thus, the on-line update may be seen as gradient descent with different learning rates. The update (4.9) uses the gradient of the loss at $\boldsymbol{\mu}_{t+1}$, while (4.10) uses the gradient of the loss evaluated at $\boldsymbol{\mu}_t$.

In the special case when $\eta_0^{-1} = 0$ (maximum likelihood), then (4.9) is not valid for $t = 1$ since $\eta_1 = \infty$. Now $\boldsymbol{\mu}_2 = \boldsymbol{x}_1$, which is consistent with updates (4.8) and (4.10).

The relative loss bounds are proven by using Lemma 4.2. Thus, for density estimation this equality simplifies to:

$$\begin{aligned}\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \big(\eta_0^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..T}(\boldsymbol{\theta})\big)\\ = \sum_{t=1}^{T}\eta_t^{-1}\Delta_G(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \eta_T^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_{T+1}).\end{aligned} \tag{4.11}$$

The following lemma gives a concise expression for the minimum of $U_{t+1}(\boldsymbol{\theta})$ (see (4.5)) in terms of the dual of the cumulant function. This lemma and the following discussion is

interesting in its own right. Although it is not essential for the main development of this paper, we will use it in the Bernoulli example discussed later. By combining (4.11) with this lemma one can also get an expression for the total loss $\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t)$ of the on-line algorithm (Lemma 3.1 of Azoury and Warmuth (1999)).

Lemma 4.3.

$$\min_{\boldsymbol{\theta}} \left(\eta_0^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..t}(\boldsymbol{\theta})\right) = \eta_0^{-1}F(\boldsymbol{\mu}_0) - \eta_t^{-1}F(\boldsymbol{\mu}_{t+1}) - \sum_{t=1}^{t} \ln P_0(\boldsymbol{x}_t).$$

Proof: We rewrite the right-hand-side of the equality of the lemma using the definition of the dual function $F(\boldsymbol{\mu})$ (3.6) and the expression (4.7) for $\boldsymbol{\mu}_{t+1}$:

$$\eta_0^{-1}(\boldsymbol{\theta}_0 \cdot \boldsymbol{\mu}_0 - G(\boldsymbol{\theta}_0)) - \boldsymbol{\theta}_{t+1} \cdot \left(\eta_0^{-1}\boldsymbol{\mu}_0 + \sum_{q=1}^{t} \boldsymbol{x}_q\right) + \eta_t^{-1}G(\boldsymbol{\theta}_{t+1}) - \sum_{q=1}^{t} \ln P_0(\boldsymbol{x}_q).$$

We rewrite the above using $\eta_t^{-1} = \eta_0^{-1} + t$ and the definitions of the loss and divergence:

$$\begin{aligned}
&\eta_0^{-1}(G(\boldsymbol{\theta}_{t+1}) - G(\boldsymbol{\theta}_0) - (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_0) \cdot \boldsymbol{\mu}_0) \\
&\quad + \sum_{q=1}^{t}(G(\boldsymbol{\theta}_{t+1}) - \boldsymbol{\theta}_{t+1} \cdot \boldsymbol{x}_q - \ln P_0(\boldsymbol{x}_q)) \\
&= \eta_0^{-1}\Delta_G(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_0) + L_{1..t}(\boldsymbol{\theta}_{t+1}).
\end{aligned}$$

The above is equal to $U_{t+1}(\boldsymbol{\theta}_{t+1})$. □

When $\eta_0^{-1} = 0$ (the case of maximum likelihood), the above can be rewritten as:

$$\inf_{\boldsymbol{\theta}} \left(\frac{1}{t}\sum_{q=1}^{t} -\ln \frac{P_G(\boldsymbol{x}_q \mid \boldsymbol{\theta})}{P_0(\boldsymbol{x}_q)}\right) = -F(\boldsymbol{\mu}_{t+1}) = E_{\boldsymbol{\theta}_{t+1}}\left(-\ln \frac{P_G(\boldsymbol{x} \mid \boldsymbol{\theta}_{t+1})}{P_0(\boldsymbol{x})}\right),$$

$$\text{where } \boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}}\left(\frac{1}{t}\sum_{q=1}^{t} -\ln \frac{P_G(\boldsymbol{x}_q \mid \boldsymbol{\theta})}{P_0(\boldsymbol{x}_q)}\right).$$

Thus, essentially the infimum of the average loss on the data equals the expected loss at the parameter that minimizes the average loss, i.e., the maximum likelihood parameter. The above relationship was used in Gruenwald (1998).

In the remaining subsections we discuss specific examples and give their relative loss bounds.

4.1.1. Density estimation with a Gaussian. Here we derive relative loss bounds for the Gaussian density estimation problem. Consider a Gaussian density over $\mathbf{R}^d$ with a known

and fixed variance-covariance matrix Σ:

$$P(\boldsymbol{x} \mid \boldsymbol{\theta}, \Sigma) \sim \exp\left(\boldsymbol{x}'\Sigma^{-1}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}'\Sigma^{-1}\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}\right).$$

Without loss of generality, we will develop the bounds for the special case when Σ is the identity matrix. Similar bounds immediately follow for the general case of fixed but arbitrary variance-covariance matrix by a linear transformation argument.

The Gaussian density with the identity matrix as the variance-covariance matrix is

$$P(\boldsymbol{x} \mid \boldsymbol{\theta}) \sim \exp\left(\boldsymbol{x} \cdot \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^2 + \frac{1}{2}\boldsymbol{x}^2\right).$$

(Here, $\boldsymbol{x}^2$ is shorthand for $\boldsymbol{x} \cdot \boldsymbol{x}$.) This density is a member of the exponential family with natural parameter $\boldsymbol{\theta}$:

Cumulant function: $G(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^2$.
Parameter transformations: $g(\boldsymbol{\theta}) = \boldsymbol{\theta} = \boldsymbol{\mu}$ and $f(\boldsymbol{\mu}) = \boldsymbol{\mu} = \boldsymbol{\theta}$ (identity functions).
Dual convex function: $F(\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}^2$.
Loss: $L_t(\boldsymbol{\theta}) = \boldsymbol{x}_t \cdot \boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^2 + \frac{1}{2}\boldsymbol{x}_t^2 + \text{const.}$

Note that the constant in the loss is immaterial for the bounds and therefore, we set it to zero.

For the sake of simplicity, set $\boldsymbol{\theta}_0 = \mathbf{0}$ and assume $T \geq 2$. Recall that for the Incremental Off-line Update, $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_B = \boldsymbol{\theta}_{T+1}$. By (4.11) with $\boldsymbol{\theta} = \boldsymbol{\theta}_B$, the difference between the total loss of the Incremental Off-line Algorithm and the off-line algorithm is

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - U_{T+1}(\boldsymbol{\theta}_B) = \sum_{t=1}^{T} \eta_t^{-1} \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})^2. \tag{4.12}$$

We use the on-line updates (e.g. (4.9), (4.10)) to rewrite the divergences on the right-hand-side:

$$\begin{aligned}
&2\eta_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})^2 \\
&\quad \overset{(4.10)}{=} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}) \cdot (\boldsymbol{\theta}_t - \boldsymbol{x}_t) \\
&\quad = (\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}) \cdot (-\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t) + (\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}) \cdot (\boldsymbol{\theta}_t + \boldsymbol{\theta}_{t+1}) \\
&\quad \overset{(4.9)}{=} \eta_{t-1}(\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t) \cdot (-\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t) + \boldsymbol{\theta}_t^2 - \boldsymbol{\theta}_{t+1}^2 \\
&\quad = \eta_{t-1}\left(\boldsymbol{x}_t^2 - \boldsymbol{\theta}_{t+1}^2\right) + \boldsymbol{\theta}_t^2 - \boldsymbol{\theta}_{t+1}^2.
\end{aligned} \tag{4.13}$$

We want to allow $\eta_0^{-1} = 0$. In this case, update (4.9) cannot be applied for $t = 1$. However, by update (4.8), we have that for any $\eta_0^{-1} \geq 0$, $\boldsymbol{\theta}_2 = \eta_1 \boldsymbol{x}_1$. Thus, $\frac{1}{2}\eta_1^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^2 = \frac{1}{2}\eta_1 \boldsymbol{x}_1^2$,

for any $\eta_0^{-1} \geq 0$. Now (4.12) can rewritten as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - U_{T+1}(\boldsymbol{\theta}_B) \\
&\quad = \frac{1}{2}\eta_1^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^2 + \sum_{t=2}^{T} \eta_t^{-1}\frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})^2 \\
&\quad \overset{(4.13)}{=} \frac{1}{2}\eta_1 \boldsymbol{x}_1^2 + \sum_{t=2}^{T} \frac{1}{2}\eta_{t-1}\boldsymbol{x}_t^2 - \sum_{t=2}^{T} \frac{1}{2}\eta_{t-1}\boldsymbol{\theta}_{t+1}^2 + \frac{1}{2}\boldsymbol{\theta}_2^2 - \frac{1}{2}\boldsymbol{\theta}_{T+1}^2. \qquad (4.14)
\end{aligned}$$

It is easy to find two examples $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for which the difference (4.14) depends on the order in which the two examples are presented. We now develop an upper bound that is *permutation invariant*, i.e., it does not depend on order in which the examples are presented. We drop the negative terms from (4.14), use $\boldsymbol{\theta}_2 = \eta_1 \boldsymbol{x}_1$, $\eta_1 \leq 1$ and assume $\boldsymbol{x}_t^2 \leq X^2$:

$$\begin{aligned}
\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - U_{T+1}(\boldsymbol{\theta}_B) &\leq \frac{1}{2}X^2\left(\eta_1 + \eta_1^2 + \sum_{t=2}^{T}\eta_{t-1}\right) \\
&\leq \frac{1}{2}X^2\left(3 + \sum_{t=3}^{T}\eta_{t-1}\right).
\end{aligned}$$

Since

$$\begin{aligned}
\sum_{t=3}^{T}\eta_{t-1} = \sum_{t=3}^{T} 1/\left(\eta_0^{-1} + t - 1\right) &\leq \int_{\eta_0^{-1}+1}^{\eta_0^{-1}+T-1} (1/x)\, dx \\
&= \ln\left(\left(\eta_0^{-1} + T - 1\right)/\left(\eta_0^{-1} + 1\right)\right),
\end{aligned}$$

we obtain the following relative loss bound:

Theorem 4.4. *For Gaussian density estimation with the Incremental Off-line Algorithm and* $\boldsymbol{\theta}_0 = \mathbf{0}$, $\eta_0^{-1} \geq 0$,

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}}\left(\frac{1}{2}\eta_0^{-1}\boldsymbol{\theta}^2 - \sum_{t=1}^{T} L_t(\boldsymbol{\theta})\right) \leq \frac{1}{2}X^2\left(3 + \ln\left(1 + \frac{T-2}{\eta_0^{-1}+1}\right)\right),$$

where $X = \max_{t=1}^{T} \boldsymbol{x}_t^2$.

Note that in the special case when $\eta_0^{-1} = 0$, then the off-line algorithm chooses a maximum likelihood parameter and the above bound simplifies to $\frac{1}{2}X^2(3 + \ln(T-1))$.

4.1.2. Density estimation with a Gamma. Here we give the relative loss bounds for the Gamma distribution. The Gamma density with shape parameter α and inverse scale parameter β is

$$P(x \mid \alpha, \beta) = \frac{e^{-x\beta}x^{\alpha-1}\beta^{\alpha}}{\Gamma(\alpha)}, \quad \text{for } x, \alpha, \beta > 0.$$

This is a member of the exponential family with natural parameter $\theta = -\alpha/\beta$. So the density above in terms of θ can be written as

$$P(x \mid \theta, \alpha) = e^{\alpha(x\theta + \ln(-\theta))} \frac{\alpha^\alpha x^{\alpha-1}}{\Gamma(\alpha)}.$$

We assume α is known and fixed. The parameter α scales the loss and the divergences. The inverse of α is called the dispersion parameter. So for the sake of simplicity we drop α just as we ignored the fixed variance in the case of Gaussian density estimation.

Cumulant function: $G(\theta) = -\ln(-\theta)$.
Parameter transformations: $g(\theta) = -1/\theta = \mu$ and $f(\mu) = -1/\mu = \theta$.
Dual convex function: $F(\mu) = -1 - \ln \mu$.
Loss: $L_t(\boldsymbol{\theta}) = x_t\theta + \ln(-\theta) + \text{const}$.

We bound the divergence between θ_t and θ_{t+1}, which leads to a relative loss bound for the Incremental Off-line Algorithm (see (4.11)):

$$\begin{aligned}\eta_t^{-1}\Delta_G(\theta_t, \theta_{t+1}) &= \eta_t^{-1}\Delta_F(\mu_{t+1}, \mu_t) \\ &= \eta_t^{-1}\left(\ln\frac{\mu_t}{\mu_{t+1}} + \frac{\mu_{t+1}}{\mu_t} - 1\right). \end{aligned} \tag{4.15}$$

Using the update (4.10) and the notation $r_t = x_t/\mu_t > 0$:

$$\begin{aligned}\eta_t^{-1}\Delta_F(\mu_{t+1}, \mu_t) &= \eta_t^{-1}\left(\ln\left(\frac{\mu_t}{\mu_t - \eta_t(\mu_t - x_t)}\right) - \frac{\eta_t(\mu_t - x_t)}{\mu_t}\right) \\ &= \eta_t^{-1}\ln\left(1 + \frac{1 - r_t}{\eta_{t-1}^{-1} + r_t}\right) + r_t - 1 \\ &\le \eta_t^{-1}\frac{1 - r_t}{\eta_{t-1}^{-1} + r_t} + r_t - 1 \\ &= \frac{(1 - r_t)^2}{\eta_{t-1}^{-1} + r_t} \le \eta_{t-1}(1 - r_t)^2\end{aligned}$$

If $x_t \le \mu_t$, then $(1 - r_t)^2 \le 1$; if $x_t > \mu_t$, then $(1 - r_t)^2 \le (\frac{x_t}{\mu_t})^2$. Let $X = \max_{t=1}^T x_t$ and $Z = \min(\{x_t : 1 \le t \le T\} \cup \{\mu_0\})$. Then $(1 - r_t)^2 \le X^2/Z^2$, because the μ_t are convex combinations of the elements of $\{x_t : 1 \le t \le T\} \cup \{\mu_0\}$. If $\eta_0^{-1} > 0$, then the sum of the divergences on the right-hand-side of (4.11) can be bounded by

$$\frac{X^2}{Z^2}\sum_{t=1}^T \eta_{t-1} = \frac{X^2}{Z^2} O(\log T).$$

In summary we have the following relative loss bound:

Theorem 4.5. *For density estimation with a Gamma distribution using the Incremental Off-line Algorithm and* $\eta_0^{-1} > 0$,

$$\sum_{t=1}^{T} L_t(\theta_t) - \min_{\theta} \left(\eta_0^{-1} \Delta_G(\theta, \theta_0) + \sum_{t=1}^{T} L_t(\theta) \right) = \frac{X^2}{Z^2} O(\log T),$$

where $X = \max_{t=1}^{T} x_t$ *and* $Z = \min(\{x_t : 1 \le t \le T\} \cup \{\mu_0\})$.

Better relative loss bounds that include the case when $\eta_0^{-1} = 0$ might be possible by bounding (4.15) more carefully.

4.1.3. Density estimation for the general exponential family. Here we give a brief discussion of the form the bounds take for any member of the exponential family. We rewrite the divergence between $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$:

$$\begin{aligned} \eta_t^{-1} \Delta_G(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) &= \eta_t^{-1} \Delta_F(\boldsymbol{\mu}_{t+1}, \boldsymbol{\mu}_t) \\ &= \eta_t^{-1} (F(\boldsymbol{\mu}_{t+1}) - F(\boldsymbol{\mu}_t) - (\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t) \cdot f(\boldsymbol{\mu}_t)). \end{aligned}$$

After doing a second order Taylor expansion of $F(\boldsymbol{\mu}_{t+1})$ at $\boldsymbol{\mu}_t$, this last term equals

$$\begin{aligned} &\frac{1}{2} \eta_t^{-1} (\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t)' \nabla_{\mu}^2 F(\boldsymbol{\mu})|_{\mu=\tilde{\mu}_t} (\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t) \\ &\quad = \frac{1}{2} \eta_t (\boldsymbol{\mu}_t - \boldsymbol{x}_t)' \nabla_{\mu}^2 F(\boldsymbol{\mu})|_{\mu=\tilde{\mu}_t} (\boldsymbol{\mu}_t - \boldsymbol{x}_t), \end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_t$ is a convex combination of $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{t+1}$. If $\nabla_{\mu}^2 F(\boldsymbol{\mu})$ is constant then we essentially have a Gaussian. The general case may be seen as a local Gaussian with the time-varying curvature. Any reasonable methods have to proceed on a case-by-case basis (Morris, 1982; Gutiérrez-Peña & Smith, 1995) based on the form of $\nabla_{\mu}^2 F(\boldsymbol{\mu})$, which is the inverse of the variance function. Recall that $\eta_t = O(1/t)$. So summing the last term should always give a $\log(T)$-style bound. Sometimes the range of the $\boldsymbol{x}_t$ needs to be restricted as done in the previous subsections for density estimation with Gaussian and Gamma distributions.

4.1.4. Linear regression. In this subsection the bounds for linear regression are developed. Here the instance domain $\mathcal{X}$ is $\mathbf{R}^d$ and the label domain $\mathcal{Y}$ is $\mathbf{R}$. The parameter domain $\boldsymbol{\Theta}$ is also $\mathbf{R}^d$ and the d components of the parameter vectors $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ are the d linear weights. For a given example $(\boldsymbol{x}_t, y_t)$ and parameter vector $\boldsymbol{\theta}$, the linear model predicts with $\boldsymbol{x}_t \cdot \boldsymbol{\theta}$. The square loss $L_t(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{x}_t \cdot \boldsymbol{\theta} - y_t)^2$ is used to measure the discrepancy between the prediction and the label for that example. Note that $L_t(\boldsymbol{\theta})$ is not strictly convex in $\boldsymbol{\theta}$. Thus, we make $U_0(\boldsymbol{\theta})$ strictly convex so that the initial divergence $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ becomes strictly convex and our updates always have a unique solution. We use $U_0(\boldsymbol{\theta}) = \frac{1}{2} \boldsymbol{\theta}' \boldsymbol{\eta}_0^{-1} \boldsymbol{\theta}$, where $\boldsymbol{\eta}_0^{-1}$ is a $d \times d$ symmetric positive definite matrix. Now the divergence to the initial parameter becomes $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{\eta}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. Thus, for linear regression, the update (4.1)

of the Incremental Off-line Update becomes:

$$\boldsymbol{\theta}_{t+1} = \text{argmin}_{\boldsymbol{\theta}}\, U_{t+1}(\boldsymbol{\theta}), \quad \text{for } 0 \le t \le T,$$
$$\text{where } U_{t+1}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\boldsymbol{\eta}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \sum_{q=1}^{t}\frac{1}{2}(\boldsymbol{x}_q'\boldsymbol{\theta}-y_q)^2. \tag{4.16}$$

Note that we use the transpose notation $\boldsymbol{x}_q'\boldsymbol{\theta}$ instead of the dot product because the subsequent derivations will use matrix algebra. The above setup for linear regression is usually interpreted as a conditional density estimation problem for a Gaussian label y_t given $\boldsymbol{x}_t$, where the cumulant function in trial t is $\frac{1}{2}(\boldsymbol{\theta}'\boldsymbol{x}_t)^2$ and the divergence corresponds to a Gaussian prior on $\boldsymbol{\theta}$.

Again we develop the alternate on-line version of (4.16) as done in general in Lemma 4.1. Since $U_t(\boldsymbol{\theta})$ equals $\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\eta}_{t-1}^{-1}\boldsymbol{\theta}$ except for linear terms, the on-line version becomes

$$\boldsymbol{\theta}_{t+1} = \text{argmin}_{\boldsymbol{\theta}}\left(\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_t)'\boldsymbol{\eta}_{t-1}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}-y_t)^2\right).$$

By differentiating (4.16) for $0 \le t \le T$, we obtain the Incremental Off-line Update for linear regression:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\eta}_t\left(\boldsymbol{\eta}_0^{-1}\boldsymbol{\theta}_0 + \sum_{q=1}^{t}\boldsymbol{x}_q y_q\right), \quad \text{where } \boldsymbol{\eta}_t = \left(\boldsymbol{\eta}_0^{-1} + \sum_{q=1}^{t}\boldsymbol{x}_q\boldsymbol{x}_q'\right)^{-1}. \tag{4.17}$$

This is the standard linear least squares update. It is easy to derive the following recursive versions (for $1 \le t \le T$):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\eta}_t\boldsymbol{\eta}_{t-1}^{-1}\boldsymbol{\theta}_t + \boldsymbol{\eta}_t\boldsymbol{x}_t y_t$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\eta}_{t-1}(\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t y_t) \tag{4.18}$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\eta}_t(\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{\theta}_t - \boldsymbol{x}_t y_t) \tag{4.19}$$

Note the correspondence of the above updates to the updates (4.7–4.10) for density estimation. Also Lemma 4.2 becomes the following quadratic equation (see (4.11) for the corresponding equation in density estimation):

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \left(\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\boldsymbol{\eta}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + L_{1..T}(\boldsymbol{\theta})\right)$$
$$= \sum_{t=1}^{T}\frac{1}{2}(\boldsymbol{\theta}_t-\boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_t^{-1}(\boldsymbol{\theta}_t-\boldsymbol{\theta}_{t+1}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{T+1})'\boldsymbol{\eta}_T^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_{T+1}). \tag{4.20}$$

We now reprove a bound obtained by Vovk (1997) for the Incremental Off-line Algorithm. (For the sake of simplicity we choose $\boldsymbol{\eta}_0^{-1}$ as a multiple of the identity matrix $\boldsymbol{I}$.) A similar bound was proven by Foster for the same algorithm. However, he assumes that the comparator $\boldsymbol{\theta}$ is a probability vector (Foster, 1991).

Theorem 4.6. *For linear regression with the Incremental Off-line Algorithm and* $\boldsymbol{\eta}_0^{-1} = a\boldsymbol{I}$ *with* $a > 0$,

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}} \left(\frac{1}{2} a(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2 - \sum_{t=1}^{T} L_t(\boldsymbol{\theta}) \right) \leq 2Y^2 d \ln\left(\frac{TX^2}{a} + 1 \right),$$

where $X = \max\{|x_{t,i}| \,:\, 1 \leq t \leq T,\ 1 \leq i \leq d\}$ *and* $Y = \max\{|y_t|, |\boldsymbol{\theta}_t \cdot \boldsymbol{x}_t| \,:\, 1 \leq t \leq T\}$.

Note that this theorem assumes that the predictions $\boldsymbol{x}_t'\boldsymbol{\theta}_t$ of the labels y_t at trial t lie in $[-Y, Y]$. If this assumption is not satisfied, we might use clipping, i.e., the algorithm predicts with the number in [−Y,Y] that is closest to $\boldsymbol{x}_t'\boldsymbol{\theta}_t$. However, clipping requires the algorithm to know Y. There is little incentive to work out the details for the Incremental Off-line Algorithm because for the algorithm of the next section we can prove a better relative loss bound and the predictions don't need to lie in $[-Y, Y]$.

Proof: We apply the Update (4.19) twice to the divergence in the sum of the right-hand-side of (4.20). This give the first two equalities below. The third equality follows from Lemma A.1 of the appendix.

$$\begin{aligned}
\frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})) &= \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{x}_t(\boldsymbol{x}_t'\boldsymbol{\theta}_t - y_t) \\
&= \frac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t - y_t)^2 \boldsymbol{x}_t'\boldsymbol{\eta}_t\boldsymbol{x}_t \\
&= \frac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t - y_t)^2 \left(1 - \frac{\left|\boldsymbol{\eta}_{t-1}^{-1}\right|}{\left|\boldsymbol{\eta}_t^{-1}\right|} \right) \\
&\leq 2Y^2 \ln \frac{\left|\boldsymbol{\eta}_t^{-1}\right|}{\left|\boldsymbol{\eta}_{t-1}^{-1}\right|}. \qquad (4.21)
\end{aligned}$$

In the last inequality we used the assumption that $\boldsymbol{x}_t'\boldsymbol{\theta}_t \in [-Y, Y]$ and the fact that $z - 1 \geq \ln z$. Note that the last inequality may not hold without the assumption $\boldsymbol{x}_t'\boldsymbol{\theta}_t \in [-Y, Y]$.

The theorem now follows from applying these crude approximations to the equality of Lemma (4.2):

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})) &\leq \sum_{t=1}^{T} 2Y^2 \ln \frac{\left|\boldsymbol{\eta}_t^{-1}\right|}{\left|\boldsymbol{\eta}_{t-1}^{-1}\right|} \\
&= 2Y^2 \ln \frac{|\boldsymbol{\eta}_T^{-1}|}{|\boldsymbol{\eta}_0^{-1}|} \\
&= 2Y^2 \ln \frac{\left|\boldsymbol{\eta}_0^{-1} + \sum_{q=1}^{T} \boldsymbol{x}_q\boldsymbol{x}_q'\right|}{\left|\boldsymbol{\eta}_0^{-1}\right|}
\end{aligned}$$

$$\leq 2Y^2 \sum_{i=1}^{d} \ln\left(1 + \frac{\sum_{q=1}^{T} x_{q,i}^2}{a}\right)$$
$$\leq 2Y^2 d \ln\left(1 + \frac{T\ X^2}{a}\right).$$

The last inequality follows from the assumption that $x_{q,i} \leq X^2$. The second to last inequality follows from the fact that $\boldsymbol{\eta}_0^{-1} = a\boldsymbol{I}$ and that the determinant of a symmetric matrix is at most the product of the diagonal elements (see Beckenbach & Bellman, 1965), Chapter 2, Theorem 7). □

Ideally we don't want to use this crude bounding method. The goal is to rewrite the sum of divergences so that further telescoping occurs. For the Incremental On-line Update we have not been able to do that. Below is a partial attempt that follows what we did for Gaussian density estimation (4.13).

$$\begin{aligned}
&\tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_t^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})) \\
&\overset{(4.19)}{=} \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{x}_t(\boldsymbol{x}_t'\boldsymbol{\theta}_t - y_t) \\
&= \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{x}_t(-\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1} - y_t) + \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{x}_t(\boldsymbol{x}_t'\boldsymbol{\theta}_t + \boldsymbol{x}_t'\boldsymbol{\theta}_{t+1}) \\
&\overset{(4.18)}{=} \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1} - y_t)'\boldsymbol{x}_t'\boldsymbol{\eta}_{t-1}'\boldsymbol{x}_t(-\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1} - y_t) + \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t)^2 - \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1})^2 \\
&= \tfrac{1}{2}\boldsymbol{x}_t'\boldsymbol{\eta}_{t-1}\boldsymbol{x}_t\left(y_t^2 - (\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1})^2\right) + \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t)^2 - \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_{t+1})^2. \qquad (4.22)
\end{aligned}$$

In the last equality we use the fact that $\boldsymbol{\eta}_{t-1}$ is symmetric. Note that the last two terms in the final expression do not telescope as they did for the Gaussian case (4.13). Surprisingly, for the Forward Algorithm that will be introduced in the next section, the corresponding two terms do telescope. Thus, for Forward Algorithm, one can prove a bound as the one given in Theorem 4.6 except that the last term in the bound is now $\frac{1}{2}Y^2 d \ln(\frac{TX^2}{a} + 1)$, a quarter of what it was in Theorem 4.6. In the related problem for density estimation with a Gaussian, the corresponding improved bound (with factor $\frac{1}{2}$) already holds for the Incremental Off-line Algorithm.

5. Estimating the future loss—the Forward Algorithm

In this section we present our second algorithm called the Forward Algorithm and give some lemmas that are used for proving relative loss bounds. In trial t, the Forward Algorithm expects to see the next example and we allow it to incorporate an estimate of the loss on this next example when choosing its parameter.

In regression, "part" of the example, namely the instance $\boldsymbol{x}_t$, is available at trial t before the algorithm must commit itself to a parameter $\boldsymbol{\theta}_t$. So the algorithm can use the instance $\boldsymbol{x}_t$ to form an estimate $\hat{L}_t(\boldsymbol{\theta})$ of the loss at trial t. As we shall see in linear regression, incorporating such an estimate in the motivation can be used to include the current instance into the learning rate of the algorithm and this leads to better relative loss bounds. In density

estimation, however, there are no instances, yet the algorithm still uses an estimate of the future loss.

On-line protocol of the Forward Algorithm

Regression:	**Density estimation:**
Initial hypothesis is $\boldsymbol{\theta}_0$.	Initial hypothesis is $\boldsymbol{\theta}_0$.
For $t = 1$ to T do	For $t = 1$ to T do
Get instance $\boldsymbol{x}_t$.	.....
Guess loss on t-th example.	Guess loss on t-th example.
Update hypothesis $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_t$.	Update hypothesis $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_t$.
Predict with $\boldsymbol{\theta}_t$.	Predict with $\boldsymbol{\theta}_t$.
Get label y_t of t-th example.	Get example $\boldsymbol{x}_t$.
Incur loss $L_t(\boldsymbol{\theta}_t)$.	Incur loss $L_t(\boldsymbol{\theta}_t)$.

We now define the update analogous to the previous section by minimizing a sum of a divergence plus the losses in the past t trials and an estimate of the loss $\hat{L}_{t+1}(\boldsymbol{\theta})$ in the next trial.

The Forward Algorithm

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta}), \quad \text{for } 0 \le t \le T,$$
$$\text{where } U_{t+1}(\boldsymbol{\theta}) = \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..t}(\boldsymbol{\theta}) + \hat{L}_{t+1}(\boldsymbol{\theta}). \tag{5.1}$$

Assumption: The losses $L_t(\boldsymbol{\theta})$ (for $1 \le t \le T$), the estimated losses $\hat{L}_t(\boldsymbol{\theta})$ (for $1 \le t \le T+1$), and $U_0(\boldsymbol{\theta})$ are differentiable and convex functions from the parameter space $\boldsymbol{\Theta}$ to the reals. Furthermore, we assume that the $\operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta})$ (for $1 \le t \le T$) always have a solution in $\boldsymbol{\Theta}$.

Note that the Incremental Off-line Algorithm is a special case of the Forward Algorithm where all the estimated losses $\hat{L}_t(\boldsymbol{\theta})$ are zero. As before there is an alternate on-line motivation of the update using a divergence to the last parameter vector.

Lemma 5.1. *For the Forward Algorithm and* $1 \le t \le T$,

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} \left(\Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta}) + \hat{L}_{t+1}(\boldsymbol{\theta}) - \hat{L}_t(\boldsymbol{\theta})\right).$$

Proof: Since $\nabla U_t(\boldsymbol{\theta}_t) = \mathbf{0}$ and

$$U_{t+1}(\boldsymbol{\theta}) = U_t(\boldsymbol{\theta}) + L_t(\boldsymbol{\theta}) + \hat{L}_{t+1}(\boldsymbol{\theta}) - \hat{L}_t(\boldsymbol{\theta}),$$

we can rewrite the argument of the argmin as:

$$U_t(\boldsymbol{\theta}) - U_t(\boldsymbol{\theta}_t) - (\boldsymbol{\theta} - \boldsymbol{\theta}_t) \cdot \nabla U_t(\boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta}) + \hat{L}_{t+1}(\boldsymbol{\theta}) - \hat{L}_t(\boldsymbol{\theta})$$
$$= U_{t+1}(\boldsymbol{\theta}) - U_t(\boldsymbol{\theta}_t).$$

Thus, since $U_t(\boldsymbol{\theta}_t)$ is a constant, the argmin for $\boldsymbol{\theta}_{t+1}$ used in the definition (5.1) is the same as the argmin of the lemma. □

The following key lemma is a generalization of Lemma 4.2 for the Incremental Off-line Algorithm.

Lemma 5.2. *For the Forward Algorithm, any sequence of T examples and any $\boldsymbol{\theta} \in \Theta$,*

$$\sum_{t=1}^{T} L_t\boldsymbol{\theta}_t) - \left(\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \sum_{t=1}^{T} L_t(\boldsymbol{\theta})\right)$$
$$= \sum_{t=1}^{T}(\Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \hat{L}_{t+1}(\boldsymbol{\theta}_t) + \hat{L}_t(\boldsymbol{\theta}_t))$$
$$- \Delta_{U_{T+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{T+1}) + \hat{L}_{T+1}(\boldsymbol{\theta}) - \hat{L}_1(\boldsymbol{\theta}) + \Delta_{U_1}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) - \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0).$$

Proof: For $0 \le t \le T$, we expand the divergence $\Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1})$ and use $\nabla U_{t+1}(\boldsymbol{\theta}_{t+1}) = \mathbf{0}$. As in the proof of Lemma 4.2, this gives us

$$\Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1}) = U_{t+1}(\boldsymbol{\theta}) - U_{t+1}(\boldsymbol{\theta}_{t+1}). \tag{5.2}$$

Since $U_{t+1}(\boldsymbol{\theta}) = U_t(\boldsymbol{\theta}) + L_t(\boldsymbol{\theta}) + \hat{L}_{t+1}(\boldsymbol{\theta}) - \hat{L}_t(\boldsymbol{\theta})$, for $1 \le t \le T$, we obtain

$$L_t(\boldsymbol{\theta}) = \Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1}) - \hat{L}_{t+1}(\boldsymbol{\theta}) + \hat{L}_t(\boldsymbol{\theta}) + U_{t+1}(\boldsymbol{\theta}_{t+1}) - U_t(\boldsymbol{\theta}). \tag{5.3}$$

For the special case of $\boldsymbol{\theta} = \boldsymbol{\theta}_t$, we get:

$$L_t(\boldsymbol{\theta}_t) = \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \hat{L}_{t+1}(\boldsymbol{\theta}_t) + \hat{L}_t(\boldsymbol{\theta}_t) + U_{t+1}(\boldsymbol{\theta}_{t+1}) - U_t(\boldsymbol{\theta}_t). \tag{5.4}$$

Subtracting (5.3) from (5.4) and applying $U_t(\boldsymbol{\theta}) - U_t(\boldsymbol{\theta}_t) = \Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ (a version of (5.2)), we obtain

$$L_t(\boldsymbol{\theta}_t) - L_t(\boldsymbol{\theta}) = \Delta_{U_{t+1}}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \hat{L}_{t+1}(\boldsymbol{\theta}_t) + \hat{L}_t(\boldsymbol{\theta}_t) - \Delta_{U_{t+1}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1})$$
$$+ \hat{L}_{t+1}(\boldsymbol{\theta}) - \hat{L}_t(\boldsymbol{\theta}) + \Delta_{U_t}(\boldsymbol{\theta}, \boldsymbol{\theta}_t).$$

The equation of the lemma follows by summing the above over all T trials and subtracting $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ from both sides. □

Any relative loss bound for the Forward Algorithm must be based on bounding the right-hand-side of this lemma.

5.1. Density estimation with the exponential family

Here we apply the Forward Algorithm to the problem of density estimation with the exponential family of distributions. We choose $U_0(\boldsymbol{\theta}) = \eta_0^{-1} G(\boldsymbol{\theta})$ as done for the Incremental Off-line Algorithm. Thus, the initial divergence becomes $U_0(\boldsymbol{\theta}) = \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \eta_0^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$.

For the estimated future loss we use $\hat{L}_{t+1}(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \boldsymbol{\mu}_0 + const$. This may be seen as the average loss of a number of examples $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ for which $(\sum_{q=1}^{k} \boldsymbol{z}_q)/k = \boldsymbol{\mu}_0$. Also, if $\boldsymbol{\mu}_0$ lies in the instance domain, then $\boldsymbol{\mu}_0$ can be seen as a guess for the future instance with the corresponding loss being $\hat{L}_{t+1}(\boldsymbol{\theta})$. The estimated loss $\hat{L}_{t+1}(\boldsymbol{\theta})$ can we rewritten as $\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + const$. Thus, with the above choices, the Forward Algorithm (5.1) becomes the following (for $0 \le t \le T$):

$$\begin{aligned}&\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta}),\\ &\quad \text{where } U_{t+1}(\boldsymbol{\theta}) = \left(\eta_0^{-1} + 1\right)\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..t}(\boldsymbol{\theta}).\end{aligned}$$

This is the same as the Incremental Off-line Algorithm (4.5) except that the trade-off parameter is now $\eta_0^{-1} + 1$ instead of η_0^{-1}. For $1 \le t \le T$, the on-line motivation becomes

$$\begin{aligned}&\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}}\left(\eta_t^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + L_t(\boldsymbol{\theta})\right),\\ &\quad \text{where } \eta_t^{-1} = \eta_0^{-1} + t.\end{aligned}$$

This is the same as the on-line motivation of the Incremental Off-line Algorithm (4.6) except that η_{t-1}^{-1} is increased by one to η_t^{-1}. The updates (4.7–4.10) and Lemma 4.3 remain the same but the learning rates are shifted:

$$\boldsymbol{\mu}_{t+1} = \eta_{t+1}\left(\eta_1^{-1}\boldsymbol{\mu}_0 + \sum_{t=1}^{t} \boldsymbol{x}_t\right), \tag{5.5}$$

$$\boldsymbol{\mu}_{t+1} = \eta_{t+1}\eta_{t-1}\boldsymbol{\mu}_t + \eta_{t+1}\boldsymbol{x}_t, \tag{5.6}$$

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \eta_t(\boldsymbol{\mu}_{t+1} - \boldsymbol{x}_t), \tag{5.7}$$

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \eta_{t+1}(\boldsymbol{\mu}_t - \boldsymbol{x}_t). \tag{5.8}$$

The above updates hold for $1 \le t \le T$. The first one holds for $t = 0$ as well which shows that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$ for our choice of the estimate $\hat{L}_{t+1}(\boldsymbol{\theta})$.

Since the estimated loss is independent of the trial, the estimated losses in the last equality of Lemma 5.2 cancel and we get:

$$\begin{aligned}&\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \left(\eta_0^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + L_{1..T}(\boldsymbol{\theta})\right)\\ &\quad = \sum_{t=1}^{T} \eta_{t+1}^{-1}\Delta_G(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) - \eta_{T+1}^{-1}\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_{T+1}) + \Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_0).\end{aligned} \tag{5.9}$$

5.2. Density estimation with a Gaussian

In this section, we give a bound for the Forward Algorithm that is better than the corresponding bound for the Incremental Off-line Algorithm. Following the same steps as in (4.13),

we simplify the following divergence:

$$2\eta_{t+1}^{-1}\Delta_G(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) = \eta_t\big(\boldsymbol{x}_t^2 - \boldsymbol{\theta}_{t+1}^2\big) + \boldsymbol{\theta}_t^2 - \boldsymbol{\theta}_{t+1}^2.$$

Using this, (5.9) becomes

$$\begin{aligned}\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \left(\frac{1}{2}\eta_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^2 + L_{1..T}(\boldsymbol{\theta})\right)\\ = \sum_{t=1}^{T}\frac{1}{2}\eta_t\big(\boldsymbol{x}_t^2 - \boldsymbol{\theta}_{t+1}^2\big) + \frac{1}{2}\boldsymbol{\theta}_1^2 - \frac{1}{2}\boldsymbol{\theta}_{T+1}^2 - \frac{1}{2}\eta_{T+1}^{-1}\big(\boldsymbol{\theta}-\boldsymbol{\theta}_{T+1}\big)^2 + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^2.\end{aligned}$$

We now set $\boldsymbol{\theta}_0$ (and thus $\boldsymbol{\theta}_1$) to zero and choose $\boldsymbol{\theta} = \boldsymbol{\theta}_B$.

$$\begin{aligned}\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \left(\frac{1}{2}\eta_0^{-1}\boldsymbol{\theta}_B^2 + L_{1..T}(\boldsymbol{\theta}_B)\right)\\ = \sum_{t=1}^{T}\frac{1}{2}\eta_t\boldsymbol{x}_t^2 - \sum_{t=1}^{T-1}\frac{1}{2}\eta_t\boldsymbol{\theta}_{t+1}^2\\ - \frac{1}{2}(\eta_T+1)\boldsymbol{\theta}_{T+1}^2 - \frac{1}{2}\eta_{T+1}^{-1}(\boldsymbol{\theta}_B-\boldsymbol{\theta}_{T+1})^2 + \frac{1}{2}\boldsymbol{\theta}_B^2.\end{aligned} \tag{5.10}$$

Since $\boldsymbol{\theta}_B = \eta_T\eta_{T+1}^{-1}\boldsymbol{\theta}_{T+1} = (1+\eta_T)\boldsymbol{\theta}_{T+1}$, $(\boldsymbol{\theta}_B - \boldsymbol{\theta}_{T+1})^2 = \eta_T^2\boldsymbol{\theta}_{T+1}^2$. Thus, the last three terms of the above equation can be rewritten as:

$$\frac{1}{2}\boldsymbol{\theta}_{T+1}^2(-1-\eta_T-\eta_T(1+\eta_T)+(1+\eta_T)^2) = 0. \tag{5.11}$$

If $\boldsymbol{x}_t^2 \le X^2$, then Equation (5.10) is bounded by

$$\frac{1}{2}X^2\sum_{t=1}^{T}\eta_t \le \frac{1}{2}X^2\left(\eta_1 + \ln\frac{\eta_0^{-1}+T}{\eta_0^{-1}+1}\right) \le \frac{1}{2}X^2\left(1+\ln\left(1+\frac{T-1}{\eta_0^{-1}+1}\right)\right).$$

Theorem 5.3. *For Gaussian density estimation with the Forward Algorithm and* $\boldsymbol{\theta}_0 = \mathbf{0}$, $\eta_0^{-1} \ge 0$,

$$\begin{aligned}\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}}\left(\frac{1}{2}\eta_0^{-1}\boldsymbol{\theta}^2 - \sum_{t=1}^{T} L_t(\boldsymbol{\theta})\right) &= \sum_{t=1}^{T}\frac{1}{2}\eta_t\boldsymbol{x}_t^2 - \sum_{t=1}^{T-1}\frac{1}{2}\eta_t\boldsymbol{\theta}_{t+1}^2\\ &\le \frac{1}{2}X^2\left(1+\ln\left(1+\frac{T-1}{\eta_0^{-1}+1}\right)\right),\end{aligned}$$

where $X = \max_{t=1}^{T}\boldsymbol{x}_t^2$.

Note that the above bound for the Forward Algorithm is better than the bound for the Incremental Off-line algorithm (see Theorem 4.4). The improvement is essentially $2X^2$.

5.3. *Density estimation with a Bernoulli*

In this subsection we give the relative loss bounds for the Bernoulli distribution. Here the examples x_t are coin flips in $\{0, 1\}$ and the distribution is typically expressed as $P(x \mid \mu) = \mu^x(1-\mu)^{1-x}$, where μ is the probability of 1. Let $\theta = \ln \frac{\mu}{1-\mu}$. So the distribution in terms of θ is

$$P(x \mid \theta) = \exp(\theta x - \ln(1 + e^\theta)).$$

This is a member of the exponential family with natural parameter θ.

Cumulant function: $G(\theta) = \ln(1 + e^\theta)$.
Parameter transformations: $\mu = g(\theta) = \frac{e^\theta}{1+e^\theta}$ and $\theta = f(\mu) = \ln \frac{\mu}{1-\mu}$.
Dual function: $F(\mu) = \mu \ln \mu + (1-\mu)\ln(1-\mu)$.
Loss: $L_t(\theta) = -x_t\theta + \ln(1 + e^\theta) = -x_t \ln \mu - (1 - x_t)\ln(1-\mu)$.

Consider the Forward Algorithm with $\eta_0^{-1} = 0$ and $\theta_0 = 0$ (i.e., $\mu_0 = \frac{1}{2}$). In this case $\mu_B = \frac{\sum_{t=1}^T x_t}{T}$ (maximum likelihood) and the Forward Algorithm uses $\mu_{t+1} = \frac{\frac{1}{2}+\sum_{q=1}^t x_q}{t+1}$. We first develop a concise expression for the total loss of the algorithm.

Lemma 5.4. *For Bernoulli density estimation with the Forward Algorithm and* $m_0 = \frac{1}{2}$, $\eta_0^{-1} = 0$,

$$\sum_{t=1}^T L_t(\theta_t) = \ln T! - \ln \prod_{q=1}^{T\mu_B}\left(q - \frac{1}{2}\right) - \ln \prod_{q=1}^{T-T\mu_B}\left(q - \frac{1}{2}\right).$$

Proof: We first rewrite the loss at trial t in various ways. Let s_t abbreviate $\sum_{q=1}^t x_q$.

$$\begin{aligned} L_t(\theta_t) &= -x_t \ln \frac{\frac{1}{2} + s_{t-1}}{t} - (1 - x_t)\ln\left(1 - \frac{\frac{1}{2} + s_{t-1}}{t}\right) \\ &= \ln t - x_t \ln\left(\frac{1}{2} + s_{t-1}\right) - (1 - x_t)\ln\left(t - \left(\frac{1}{2} + s_{t-1}\right)\right) \\ &= \ln t - \begin{cases} \ln\left(s_t - \frac{1}{2}\right) & \text{if } x_t = 1 \\ \ln\left(t - s_t - \frac{1}{2}\right) & \text{if } x_t = 0. \end{cases} \end{aligned}$$

We now develop a formula for $\sum_{t=1}^T L_t(\theta_t)$. Note that in all trials t in which $x_t = 1$ (i.e., when s_t increases by one), the loss $L_t(\theta_t)$ contains the term $-\ln(s_t - \frac{1}{2})$. Over all T trials, these terms contribute $\sum_{q=1}^{s_T} \ln(q - \frac{1}{2})$. Similarly, in all trials t in which $x_t = 0$ (i.e., when $t - s_t$ increases by one), the loss $L_t(\theta_t)$ contains the term $-\ln(t - s_t - \frac{1}{2})$. Over all T trials, these terms contribute $\sum_{q=1}^{T-s_T} \ln(q - \frac{1}{2})$. From this and the fact that $T\mu_B = s_T$, the lemma follows. □

Note that the right-hand-side of the expression of the lemma is independent of the order in which the examples were seen. Thus, for Bernoulli distribution, the total loss of the forward algorithm is permutation invariant. By Lemma 4.3, $L_{1..T}(\theta_B) = \min_\theta L_{1..T}(\theta) = -\eta_T^{-1} F(\mu_B)$ and thus

$$\sum_{t=1}^{T} L_t(\theta_t) - L_{1..T}(\theta_B) = \ln T! - \ln \prod_{q=1}^{T\mu_B} \left(q - \frac{1}{2}\right) - \ln \prod_{q=1}^{T-T\mu_B} \left(q - \frac{1}{2}\right) + T(\mu_B \ln \mu_B + (1-\mu_B)\ln(1-\mu_B)).$$

An equivalent expression using the Gamma function was first derived by Freund (1996) based on the Laplace method of integration, (see discussion in Section 6).

$$\sum_{t=1}^{T} L_t(\theta_t) - L_{1..T}(\theta_B) = \ln \Gamma(T+1) - \ln \Gamma\left(T\mu_B + \frac{1}{2}\right) - \ln \Gamma\left(T - T\mu_B + \frac{1}{2}\right) + \ln(\pi) + T(\mu_B \ln \mu_B + (1-\mu_B)\ln(1-\mu_B)).$$

Using the standard approximations of the Gamma function one can bound the right-hand-side of the above by $\frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\pi$.

Theorem 5.5 (Freund, 1996)**.** *For Bernoulli density estimation with the Forward Algorithm and* $\mu_0 = \frac{1}{2}$, $\eta_0^{-1} = 0$,

$$\sum_{t=1}^{T} L_t(\theta_t) - \min_\theta \sum_{t=1}^{T} L_t(\theta) \le \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\pi.$$

5.4. Linear regression

In this subsection we derive relative loss bounds for the Forward Algorithm when applied to linear regression. As for the Incremental Off-line Algorithm, we let $U_0(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\eta}_0^{-1}\boldsymbol{\theta}$, where $\boldsymbol{\eta}_0$ is symmetric and positive definite. The divergence to the initial $\boldsymbol{\theta}_0$ is again $\Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\boldsymbol{\eta}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$. We use the estimated future loss $\hat{L}_{t+1}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta} - \boldsymbol{x}_{t+1}'\boldsymbol{\theta}_0)^2$, i.e., the next label y_{t+1} is guessed as $\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_0$. Thus, the Forward Update (5.1) for linear regression becomes:

$$\boldsymbol{\theta}_{t+1} = \operatorname{argmin}_{\boldsymbol{\theta}} U_{t+1}(\boldsymbol{\theta}), \quad \text{for } 0 \le t \le T,$$

$$\text{where } U_{t+1}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{\eta}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \sum_{q=1}^{t} \frac{1}{2}(\boldsymbol{x}_q'\boldsymbol{\theta} - y_q)^2 + \frac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta} - \boldsymbol{x}_{t+1}'\boldsymbol{\theta}_0)^2.$$

With this definition, $U_1(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{\eta}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{x}_1'\boldsymbol{\theta} - \boldsymbol{x}_1'\boldsymbol{\theta}_0)^2$ is minimized at $\boldsymbol{\theta}_0$, and thus $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0$. By differentiating $U_{t+1}(\boldsymbol{\theta})$, we obtain the Forward Update for linear regression ($0 \le t \le T$):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\eta}_{t+1}\left((\boldsymbol{\eta}_0^{-1} + \boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}')\boldsymbol{\theta}_0 + \sum_{q=1}^{t} \boldsymbol{x}_q y_q\right),$$

$$\text{where } \boldsymbol{\eta}_{t+1} = \left(\boldsymbol{\eta}_0^{-1} + \sum_{q=1}^{t+1} \boldsymbol{x}_q \boldsymbol{x}_q'\right)^{-1}. \tag{5.12}$$

Note that $\boldsymbol{\theta}_{t+1}$ depends on $\boldsymbol{x}_{t+1}$. Thus, the Forward Algorithm is different from the Incremental Off-line Algorithm (4.17) in that it uses the current instance to form its prediction.

For the sake of simplicity we assume for the rest of this section that $\boldsymbol{\theta}_0 = \mathbf{0}$. Recursive versions of the above update are the following ($1 \le t \le T$).

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\eta}_{t+1}\boldsymbol{\eta}_t^{-1}\boldsymbol{\theta}_t + \boldsymbol{\eta}_{t+1}\boldsymbol{x}_t y_t$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\eta}_t(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t y_t) \tag{5.13}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\eta}_{t+1}(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_t - \boldsymbol{x}_t y_t). \tag{5.14}$$

We now rewrite the equality of Lemma 5.2 for linear regression. Since $\Delta_{U_1}(\boldsymbol{\theta}, \boldsymbol{\theta}_1) - \Delta_{U_0}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{x}_1\boldsymbol{x}_1'\boldsymbol{\theta} = \hat{L}_1(\boldsymbol{\theta})$, we get

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \left(\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\eta}_0^{-1}\boldsymbol{\theta} + L_{1..T}(\boldsymbol{\theta})\right)$$

$$= \sum_{t=1}^{T}\left(\frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_{t+1}^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}) - \frac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_t)^2 + \frac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t)^2\right)$$

$$- \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{T+1})'\boldsymbol{\eta}_{T+1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{T+1}) + \frac{1}{2}(\boldsymbol{x}_{T+1}'\boldsymbol{\theta})^2. \tag{5.15}$$

Following the same steps we used for the Incremental Off-line Algorithm (4.22) we get:

$$\begin{aligned}
&\tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_{t+1}^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}) \\
&\overset{(5.14)}{=} \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_t - \boldsymbol{x}_t y_t) \\
&= \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'(-\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t y_t) + \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'(\boldsymbol{\theta}_t + \boldsymbol{\theta}_{t+1}) \\
&\overset{(5.13)}{=} \tfrac{1}{2}(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t' y_t)\boldsymbol{\eta}_t'(-\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1} - \boldsymbol{x}_t y_t) \\
&\quad + \tfrac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_t)^2 - \tfrac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1})^2 \\
&= \tfrac{1}{2}\boldsymbol{x}_t'\boldsymbol{\eta}_t\boldsymbol{x}_t y_t^2 - \tfrac{1}{2}\boldsymbol{x}_{t+1}'\boldsymbol{\eta}_t\boldsymbol{x}_{t+1}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1})^2 + \tfrac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_t)^2 - \tfrac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1})^2.
\end{aligned}$$

Using the above, the argument of the sum on the right-hand-side of Equation (5.15) can be simplified as follows:

$$\begin{aligned}
&\tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1})'\boldsymbol{\eta}_{t+1}^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}) - \tfrac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_t)^2 + \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t)^2 \\
&\quad = \tfrac{1}{2}\boldsymbol{x}_t'\boldsymbol{\eta}_t\boldsymbol{x}_t y_t^2 - \tfrac{1}{2}\boldsymbol{x}_{t+1}'\boldsymbol{\eta}_t\boldsymbol{x}_{t+1}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1})^2 + \tfrac{1}{2}(\boldsymbol{x}_t'\boldsymbol{\theta}_t)^2 - \tfrac{1}{2}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1})^2.
\end{aligned}$$

Note that the last two terms telescope. In the corresponding derivation for the Incremental Off-line Update, the last two terms did not telescope (4.22). So now (5.15) with $\boldsymbol{\theta} = \boldsymbol{\theta}_B$ becomes:

$$\begin{aligned}
&\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \left(\frac{1}{2}\boldsymbol{\theta}_B'\boldsymbol{\eta}_0^{-1}\boldsymbol{\theta}_B + L_{1..T}(\boldsymbol{\theta}_B)\right) \\
&\quad = \sum_{t=1}^{T}\frac{1}{2}\boldsymbol{x}_t'\boldsymbol{\eta}_t\boldsymbol{x}_t y_t^2 - \sum_{t=1}^{T-1}\frac{1}{2}\boldsymbol{x}_{t+1}'\boldsymbol{\eta}_t\boldsymbol{x}_{t+1}(\boldsymbol{x}_{t+1}'\boldsymbol{\theta}_{t+1})^2 \\
&\qquad - \frac{1}{2}(1 + \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1})(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2 \\
&\qquad - \frac{1}{2}(\boldsymbol{\theta}_B - \boldsymbol{\theta}_{T+1})'\boldsymbol{\eta}_{T+1}^{-1}(\boldsymbol{\theta}_B - \boldsymbol{\theta}_{T+1}) + \frac{1}{2}(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_B)^2
\end{aligned} \tag{5.16}$$

As in Gaussian density estimation (5.10), we will now show that the last three terms of the above equation are zero. First we rewrite $\boldsymbol{\theta}_B$ as:

$$\boldsymbol{\theta}_B = \boldsymbol{\eta}_T\boldsymbol{\eta}_{T+1}^{-1}\boldsymbol{\theta}_{T+1} = (\boldsymbol{I} + \boldsymbol{\eta}_t\boldsymbol{x}_{T+1}\boldsymbol{x}_{T+1}')\boldsymbol{\theta}_{T+1},$$

where $\boldsymbol{I}$ is the identity matrix.

The last term becomes:

$$\tfrac{1}{2}(\boldsymbol{x}_{T+1}'(\boldsymbol{I} + \boldsymbol{\eta}_T\boldsymbol{x}_{T+1}\boldsymbol{x}_{T+1}')\boldsymbol{\theta}_{T+1})^2 = \tfrac{1}{2}(1 + \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1})^2(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2.$$

The second to last term simplifies to:

$$\begin{aligned}
&-\tfrac{1}{2}(\boldsymbol{\eta}_T\boldsymbol{x}_{T+1}\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})'\boldsymbol{\eta}_{T+1}^{-1}(\boldsymbol{\eta}_T\boldsymbol{x}_{T+1}\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1}) \\
&\quad = -\tfrac{1}{2}\boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{\eta}^{-1}1_{T+1}\boldsymbol{\eta}_T\boldsymbol{x}_{T+1}(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2 \\
&\quad = -\tfrac{1}{2}\boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T(\boldsymbol{I} + \boldsymbol{x}_{T+1}\boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T)\boldsymbol{x}_{T+1}(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2 \\
&\quad = -\tfrac{1}{2}\boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1}(1 + \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1})(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2.
\end{aligned}$$

We now sum the last three terms while pulling out the factor $\frac{1}{2}(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2$:

$$\begin{aligned}
&\tfrac{1}{2}(-1 - \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1} - \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1}(1 + \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1}) \\
&\quad + (1 + \boldsymbol{x}_{T+1}'\boldsymbol{\eta}_T\boldsymbol{x}_{T+1})^2)(\boldsymbol{x}_{T+1}'\boldsymbol{\theta}_{T+1})^2 = 0.
\end{aligned}$$

Thus, the last three terms are zero just as they were for Gaussian density estimation with the Forward algorithm (5.11).

Finally, we can use the upper bound (see Theorem 4.6)

$$\frac{1}{2}y_t^2 \boldsymbol{x}_t' \boldsymbol{\eta}_t \boldsymbol{x}_t = y_t^2 \left(1 - \frac{\left|\boldsymbol{\eta}_{t-1}^{-1}\right|}{\left|\boldsymbol{\eta}_t^{-1}\right|}\right) \leq \frac{1}{2} Y^2 \ln \frac{\left|\boldsymbol{\eta}_{t-1}^{-1}\right|}{\left|\boldsymbol{\eta}_t^{-1}\right|},$$

where $Y = \max_{t=1}^{T} y_t^2$. Now the sums on the right-hand-side of (5.16) can be bounded by $\frac{1}{2}Y^2 d \ln(1 + \frac{TX^2}{a})$.

We now summarize the relative loss bound we proved for the Forward Algorithm:

Theorem 5.6. *For linear regression with the Forward Algorithm and $\boldsymbol{\theta}_0 = 0$, $\boldsymbol{\eta}_0^{-1} = a\boldsymbol{I}$ with $a > 0$,*

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}} \left(\frac{1}{2} a\boldsymbol{\theta}^2 - \sum_{t=1}^{T} L_t(\boldsymbol{\theta})\right)$$

$$= \sum_{t=1}^{T} \frac{1}{2} \boldsymbol{x}_t' \boldsymbol{\eta}_t \boldsymbol{x}_t y_t^2 - \sum_{t=1}^{T-1} \frac{1}{2} \boldsymbol{x}_{t+1}' \boldsymbol{\eta}_t \boldsymbol{x}_{t+1} (\boldsymbol{x}_{t+1}' \boldsymbol{\theta}_{t+1})^2 \tag{5.17}$$

$$\leq \frac{1}{2} Y^2 d \ln\left(1 + \frac{TX^2}{a}\right). \tag{5.18}$$

where $Y = \max_{t=1}^{T} y_t^2$.

Note the above bound for the Forward algorithm is better than the corresponding bound for the Incremental Off-line Algorithm (see Theorem 4.6). The improvement is by a factor of four. The above bound (5.18) was first proven by Vovk (1997) using integration. An alternate proof for the exact expression (5.17) was given by Forster (1999).

6. Relationship to the Bayes' algorithm

There is an alternate method pioneered by Vovk (1997), Freund (1996) and Yamanishi (1998) for proving relative loss bounds. In this section we sketch this method and compare it to ours. A distribution is maintained on the set of parameters $\boldsymbol{\Theta}$. The parameters can be names of algorithms and are called experts in the on-line learning literature (Cesa-Bianchi et al., 1997). The initial work only considered the case when $\boldsymbol{\Theta}$ is finite (Vovk 1990; Littlestone & Warmuth, 1994; Haussler, Kivinen, & Warmuth, 1998). However, the method can also be applied when $\boldsymbol{\Theta}$ is continuous (Freund, 1996; Vovk, 1997; Yamanishi, 1998). At the beginning of trial t, the distribution on $\boldsymbol{\Theta}$ has the form

$$P_t(\boldsymbol{\theta}) = \frac{P_1(\boldsymbol{\theta})\ e^{-\eta L_{1..t-1}(\boldsymbol{\theta})}}{\int_{\tilde{\boldsymbol{\theta}}} P_1(\tilde{\boldsymbol{\theta}})\ e^{-\eta L_{1..t-1}(\tilde{\boldsymbol{\theta}})}}, \tag{6.1}$$

where P_1 is a prior and $\eta > 0$ a learning rate. The following type of inequality is the main part of the method:

$$\begin{aligned} L_A(t) &\leq -\frac{1}{\eta} \ln \int_{\boldsymbol{\theta}} P_t(\boldsymbol{\theta})\, e^{-\eta L_t(\boldsymbol{\theta})} \\ &= -\frac{1}{\eta} \ln \int_{\boldsymbol{\theta}} P_1(\boldsymbol{\theta})\, e^{-\eta L_{1..t}(\boldsymbol{\theta})} + \frac{1}{\eta} \ln \int_{\boldsymbol{\theta}} P_1(\boldsymbol{\theta})\, e^{-\eta L_{1..t-1}(\boldsymbol{\theta})}. \end{aligned} \tag{6.2}$$

Here $L_A(t)$ denotes the loss of the algorithm at trial t. An important special case occurs when when $\eta = 1$ and the loss $L_q(\boldsymbol{\theta})$ is a negative log-likelihood with respect to a parameterized density, i.e., $L_q(\boldsymbol{\theta}) = -\ln(P(\boldsymbol{x}_q \mid \boldsymbol{\theta}))$. We call this the *Bayes' Algorithm*. In this case, P_t as given in (6.1) is the posterior distribution after seeing the first $t-1$ examples. If the algorithm in trial t predicts with the predictive distribution $P(\boldsymbol{x} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1})$ (i.e., $L_A(t) = -\ln P(\boldsymbol{x}_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1})$), then (6.2) is an equality (DeSantis, Markowsky, & Wegman, 1988).

In the more general setting (when $\eta \neq 1$ and the losses are not necessarily negative log-likelihoods), the prediction of the algorithm is chosen so that inequality (6.2) holds no matter what the t-th example will be. Also, the larger the learning rate η, the better the resulting relative loss bounds. The learning rate η is chosen as large as possible so that a prediction is always guaranteed to exist for which inequality (6.2) holds. The same learning rate is used in all T trials. The inequality is often tight when the examples lie on the boundary of the set of possible examples. By summing the inequality (6.2) over all trials, one gets the bound

$$\sum_{t=1}^{T} L_A(t) \leq -\frac{1}{\eta} \ln \int_{\boldsymbol{\theta}} P_1(\boldsymbol{\theta}) e^{-\eta L_{1..T}(\boldsymbol{\theta})}. \tag{6.3}$$

If the above integral cannot be computed exactly, then it is bounded using Laplace's method of integration around the best off-line parameter (Freund, 1996; Yamanishi, 1998).

We would like to point out the following distinction between the above method and the algorithms presented in this paper. The prediction of the Bayes' Algorithm (i.e., the predictive distribution) is usually not represented by a parameter (or expert) in $\boldsymbol{\Theta}$. Instead, the algorithms analyzed in this paper chooses a MAP parameter in $\boldsymbol{\Theta}$ in each trial. In the case of the Bernoulli distribution, the Bayes' Algorithm based on Jeffrey's prior (Freund, 1996; Xie & Barron, 1997) coincides with the Forward Algorithm when $\eta_0^{-1} = 0$ and $\mu_0 = \frac{1}{2}$ (Section 5.3). Similarly, in the case of linear regression with the Gaussian prior, Vovk's algorithm (Vovk, 1997) coincides with the Forward Algorithm for linear regression (Section 5.4). However, it is not clear that for other density estimation problems in the exponential family the predictions of the Bayes' Algorithm (Takeuchi & Barron, 1997; Takeuchi & Barron, 1998) (or the algorithms produced by for the more general case when $\eta \neq 1$) are represented by parameters in the parameter space.

In contrast, our method of proving relative loss bounds avoids the use of the involved integration methods needed for (6.3) (Freund, 1996; Yamanishi, 1998; Cover, 1991; Cover & Ordentlich, 1996). The parameter we maintain is based on a simple average of the past examples, which is a sufficient statistic for the exponential family.

7. Conclusion

In this paper, we presented techniques for proving relative loss bound for density estimation with the exponential family. We gave a number of examples of how to apply our methods, including the case of linear regression. For an exponential density with cumulant function G, we use the Bregman divergence $\Delta_G(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ as the measure of distances between the parameters $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Thus, the loss $L_t(\boldsymbol{\theta}) = -\ln P_G(\boldsymbol{x}_t \mid \boldsymbol{\theta})$ and divergence are based on the same function G. However, lemmas 4.2 and 5.2 and the methodology for proving relative loss bounds are more general in that the initial divergence and the loss do not need to be related. These lemmas might be further extended to non-differentiable convex functions using the generalized notion of Bregman divergences that was introduced by Gordon (1999).

The parameters maintained by our algorithms are invariant to permuting the past examples. However, the total on-line loss of the algorithms is not permutation invariant. Therefore, an adversary could use this fact and present the examples in an order disadvantageous to the learning algorithm. This suggests that there are algorithms with better relative loss bounds.

The Incremental Off-line Algorithm and the Forward Algorithm are incomparable in the sense that either one may have a larger total loss. However, we believe that in some sense the Forward Algorithm is better and this needs to be formalized. This belief is inspired by the phenomenon of the Stein estimator in statistics (Stein, 1956), since like the Stein estimator, the Forward Update uses a shrinkage factor when compared to the Incremental Off-line Update.

We still need to explore how the bounds obtained in this paper relate to the large body of research from the Minimum Description Length literature (Rissanen, 1989; Rissanen, 1996). In this literature, lower and upper bounds on the relative loss of the form $\frac{d}{2}\ln(T) + O(1)$ have been explored extensively, where d is the number of parameters. However, in contrast to the setup used in this paper, the work in the Minimum Description Length literature does not require that the algorithm predicts with a parameter in the parameter space.

The methodology developed in this paper for proving relative loss bounds still needs to be worked out for more learning problems. For instance, is there always a $\log(T)$-style relative loss bound for any member of the exponential family (see Section 4.1.3). Another open problem is to prove $\log(T)$-style bounds for logistic regression?

Finally, lower bounds on the relative loss need to explored for the case when the algorithm is restricted to predict with a parameter in the parameter space. Such bounds have been shown for linear regression (Vovk, 1997). In particular, it was proven that the constant before $\ln(T)$ in the relative loss bound of the Forward Algorithm (Theorem 5.6) is tight. However, no general $\log(T)$-style lower bounds is known for an arbitrary member of the exponential family.

Appendix A

Proof of Property 7 of Section 3.3:

Since the divergence is convex in its first argument, we can use it to define another divergence $\Delta_J(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$, where $J(\boldsymbol{\theta}) = \Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_3)$. We now expand $\Delta_J(\boldsymbol{\theta}, \boldsymbol{\theta}_2)$ as follows:

$$\begin{aligned}\Delta_J(\boldsymbol{\theta}, \boldsymbol{\theta}_2) &= J(\boldsymbol{\theta}_1) - J(\boldsymbol{\theta}_2) - (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \cdot \nabla_{\boldsymbol{\theta}_2} J(\boldsymbol{\theta}_2) \\ &= \Delta_G(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3) - \Delta_G(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3) - (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \cdot (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3). \qquad \text{(A.1)}\end{aligned}$$

Since $G(\boldsymbol{\theta}) - J(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, we have by Property 6 that the divergences with respect to G and J are the same. Thus, (A.1) is equivalent to Property 7. □

Lemma A.1. *For $t = 1, \ldots, T$,*

$$\boldsymbol{x}_t' \eta_t \boldsymbol{x}_t = 1 - \frac{\left|\eta_{t-1}^{-1}\right|}{\left|\eta_t^{-1}\right|}.$$

Proof:

$$\begin{aligned}\left|\eta_{t-1}^{-1}\right| &= \left|\eta_t^{-1} - \boldsymbol{x}_t \boldsymbol{x}_t'\right| = \left|\eta_t^{-1}\right|\left|\boldsymbol{I} - \eta_t \boldsymbol{x}_t \boldsymbol{x}_t'\right| \\ &= \left|\eta_t^{-1}\right|\left|\eta_t^{-1/2}(\boldsymbol{I} - \eta_t \boldsymbol{x}_t \boldsymbol{x}_t')\eta_t^{1/2}\right| = \left|\eta_t^{-1}\right|\left|\boldsymbol{I} - \eta_t^{1/2} \boldsymbol{x}_t \boldsymbol{x}_t' \eta_t^{1/2}\right|.\end{aligned}$$

So we only have to show that $\left|\boldsymbol{I} - \eta_t^{1/2} \boldsymbol{x}_t \boldsymbol{x}_t' \eta_t^{1/2}\right| = 1 - \boldsymbol{x}_t' \eta_t \boldsymbol{x}_t$. With $\boldsymbol{z} = \eta_t^{1/2} \boldsymbol{x}_t$ this can be rewritten as $\left|\boldsymbol{I} - \boldsymbol{z}\boldsymbol{z}'\right| = 1 - \boldsymbol{z}'\boldsymbol{z}$. The determinant of $\boldsymbol{I} - \boldsymbol{z}\boldsymbol{z}'$ is the product of its d eigenvalues. Since

$$(\boldsymbol{I} - \boldsymbol{z}\boldsymbol{z}')\boldsymbol{z} = \boldsymbol{z} - \boldsymbol{z}\boldsymbol{z}'\boldsymbol{z} = \boldsymbol{z} - \boldsymbol{z}'\boldsymbol{z}\,\boldsymbol{z} = (1 - \boldsymbol{z}'\boldsymbol{z})\boldsymbol{z},$$

$\boldsymbol{z}$ is an eigenvector with eigenvalue $1 - \boldsymbol{z}'\boldsymbol{z}$. For any set of $d - 1$ vectors $\boldsymbol{u_i}$ orthogonal to $\boldsymbol{z}$ we have $(\boldsymbol{I} - \boldsymbol{z}\boldsymbol{z}')\boldsymbol{u_i} = 1\boldsymbol{u_i}$. Thus, the $\boldsymbol{u_i}$ are all eigenvectors with eigenvalue one and the product of all d eigenvectors is $1 - \boldsymbol{z}'\boldsymbol{z}$. □

Acknowledgments

Thanks to Leonid Gurvits for introducing us to Bregman divergences, and to Claudio Gentile, Peter Grünwald, Geoffrey Gordon, Eiji Takimoto, and two anonymous referees for many valuable comments. The second author was supported by NSF grants CCR 9700201 and CCR-9821087.

References

Auer, P. Herbster, M. & Warmuth, M. K. (1995). Exponentially many local minima for single neurons. In *Proc. 1995 Neural Information Processing Conference* (pp. 316–317). Cambridge, MA: MIT Press.

Amari, S. (1985). *Differential Geometrical Methods in Statistics*. Berlin: Springer Verlag.

Azoury, K. & Warmuth, M. K. (1999). Relative loss bounds for on-line density estimation with the exponential family of distributions. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 31–40) Morgan Kaufmann,San Francisco CA.

Beckenbach, E. F. & Bellman, R. (1965). *Inequalities*. Berlin: Springer. Second revised printing.

Blackwell, D. (1956). An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics, 6:1*, 1–8.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Chichester: Wiley.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics, 7*, 200–217.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., & Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM, 44:3*, 427–485.

Cesa-Bianchi, N., Long, P., & Warmuth, M. K. (1996). Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks, 7:2*, 604–619.

Censor, Y. & Lent, A. (1981). An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications, 34:3*, 321–353.

Cover, T. M. & Ordentlich, E. (1996). Universal portfolios with side information. *IEEE Transactions on Information Theory, 42:2*, 348–363.

Cover, T. M. (1991). Universal portfolios. *Mathematical Finance, 1:1*, 1–29.

Csiszar, I. (1991). Why least squares and maximum entropy? An axiomatic approach for linear inverse problems. *The Annals of Statistics, 19:4*, 2032–2066.

DeSantis, A., Markowsky, G., & Wegman, M. N. (1988). Learning probabilistic prediction functions. In *Proc. 29th Annu. IEEE Sympos. Found. Comput. Sci.* (pp. 110–119). Los Alamitos, CA: IEEE Computer Society Press.

Forster, J. (1999). On relative loss bounds in generalized linear regression. In *12th International Symposium on Fundamentals of Computation Theory (FCT '99)* (pp. 269–280). Springer.

Foster, D. P. (1991). Prediction in the worst case. *The Annals of Statistics, 19:2*, 1084–1090.

Freund, Y. (1996). Predicting a binary sequence almost as well as the optimal biased coin. In *Proc. 9th Annu. Conf. on Comput. Learning Theory* (pp. 89–98). New York, NYACM Press.

Grove, A. J., Littlestone, N., & Schuurmans, D. (1997). General convergence results for linear discriminant updates. In *Proc. 10th Annu. Conf. on Comput. Learning Theory* (pp. 171–183). ACM.

Gordon, G. J. (1999). Approximate solutions to Markov decision processes. Ph.D. Thesis, Department of Computer Science, Carnegie Mellon University, Pittsburgh. Technical report CMU-CS-99-143.

Gutiérrez-Peña & Smith, A. F. M. (1995). Conjugate parameterizations of natural exponential families. *Journal of the American Statistical Association, 90:432*, 1347–1356.

Gruenwald, P. D. (1998). The Minimum Discription Length Principle and reasoning under uncertainty. Ph.D. thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, October 1998.

Hannan, J. (1957). M. Dresher, A. W. Tucker & P. Wolfe (Eds.), *Approximation to Bayes risk in repeated play*, Vol. III. Princeton University Press.

Herbster, M. (1998). Learning algorithms for tracking changing concepts and an investigation into the error surfaces of single artificial neurons. Ph.D. Thesis, Department of Computer Science, University of California at Santa Cruz.

Helmbold, D. P., Kivinen, J., & Warmuth, M. K. (1999). Relative loss bounds for single neurons. *IEEE Transactions on Neural Networks, 10:6*, 1291–1304.

Haussler, D., Kivinen, J., & Warmuth, M. K. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory, 44:2*, 1906–1925.

Herbster, M. & Manfred, W. K. (1998). Tracking the best regressor. In *Proc. 11th Annu. Conf. on Comput. Learning Theory* (pp. 24–31). New York, NY: ACM Press.

Jones, L. & Byrne, C. (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. *IEEE Transactions on Information Theory, 36:1*, 23–30.

Kivinen, J. & Warmuth, M. K. (1997). Additive versus exponentiated gradient updates for linear prediction. *Information and Computation, 132:1*. 1–64.

Kivinen, J. & Warmuth, M. K. (2000). Relative loss bounds for multidimensional regression problems, to appear in Journal of Machine Learning.

Littlestone, N. (1988). Learning when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning, 2*, 285–318.

Littlestone, N., Long, P. M., & Warmuth, M. K. (1995). On-line learning of linear functions. *Journal of Computational Complexity, 5*, 1–23.

Littlestone N. & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation, 108:2*, 212–261.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics, 10:1*, 65–80.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42:1*, 40–47.

Rockafellar, R. (1997). *Convex Analysis*. Princeton University Press.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. of the Third Berkeley Symposium on Mathematical Statistics and Probability*, (Vol. 1) (pp 197–205). Berkeley: University of California Press.

Takeuchi, J. & Barron, A. (1997). Asymptotically minimax regret for exponential families. In *SITA '97* (pp. 665–668).

Takeuchi, J. & Barron, A. (1998). Asymptotically minimax regret by Bayes mixtures. In *IEEE ISIT '98*.

Vovk, V. (1990). Aggregating strategies. In *Proc. 3rd Annu. Workshop on Comput. Learning Theory* (pp. 371–383). Morgan Kaufmann.

Vovk, V. (1997). Competitive on-line linear regression. Technical Report CSD-TR-97-13, Department of Computer Science, Royal Holloway, University of London.

Warmuth, M. K. & Jagota, A. (1998). Continuous and discrete time nonlinear gradient descent: relative loss bounds and convergence. In R. Greiner & E. Boros (Eds.), *Electronic Proceedings of Fifth International Symposium on Artificial Intelligence and Mathematics*. Electronic, http://rutcor.rutgers.edu/˜amai.

Xie, Q. & Barron, A. R. (1997). Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory, 43:2*, 646–57.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Transaction on Information Theory, 44:4*, 1424–1439.

Received November 4, 1999
Revised November 4, 1999
Accepted September 11, 2000
Final manuscript September 11, 2000