An Analysis of Logistic Models:
Exponential Family Connections and Online Performance

Arindam Banerjee*

Abstract

Logistic models are arguably one of the most widely used data analysis techniques. In this paper, we present analyses focussing on two important aspects of logistic models—its relationship with exponential family based generative models, and its performance in online and potentially adversarial settings. In particular, we present two new theoretical results on logistic models focusing on the above two aspects. First, we establish an exact connection between logistic models and exponential family based generative models, resolving a long-standing ambiguity over their relationship. Second, we show that online Bayesian logistic models are competitive to the best batch models, even in potentially adversarial settings. Further, we discuss relevant connections of our analysis to the literature on integral transforms, and also present a new optimality result for Bayesian models. The analysis makes a strong case for using logistic models and partly explains the success of such models for a wide range of practical problems.

1   Introduction

Logistic models are arguably one of the most widely used tools in data analysis techniques [22]. Two of the most popular logistic models are logistic regression (LR) [22, 32, 34], used for classification, and conditional random fields (CRFs) [27], used for structured prediction. Both types of models have been successfully applied to prediction problems in a wide range of problem domains. The key unifying aspect of logistic models is that they are discriminative models where the class posterior distributions are assumed to be Gibbs distributions [22, 27] over the features. In practice, one often prefers a Bayesian logistic regression approach [29, 26] that provides regularization, which can be valuable for high-dimensional problems [33].

Discriminative probabilistic models for classification are often more desirable than their generative counterparts as discriminative models do not try to model the joint distribution $p(\mathbf{x}, y)$ over features $\mathbf{x} \in \mathbb{R}^d$ and labels $y \in \{-1, +1\}$. Instead, they explicitly model the posterior $p(y|\mathbf{x})$. In order to do the modeling, discrimi-

native models make appropriate assumptions about the parametric form of $p(y|\mathbf{x})$ [25, 27], and use training data to learn the parameters. Logistic models are one of the most widely studied discriminative models [35, 24]. The parametric assumption of logistic regression is that the log-odds ratio of the class posteriors $p(y|\mathbf{x})$ is an affine function of the features $\mathbf{x}$. This assumption has appropriate extensions to the multi-class [22] and structured prediction [27] settings.

Generative models, on the other hand, make explicit assumptions about the class conditional distributions $p(\mathbf{x}|y)$. Such assumptions translate to assumptions about the marginal $p(\mathbf{x})$ as well as the joint distribution $p(\mathbf{x}, y)$. Modeling the joint distribution for classification purposes is considered an over-kill in that one is solving a more general problem than is necessary for solving the classification problem. Further, generative model typically have higher "bias" [17], which typically translate to faster training but higher asymptotic errors [34]. Nevertheless, generative models perform quite well for some applications such as speech recognition [37] and text analysis [10]. Popular generative models often use mixtures of exponential family distributions employing either point estimation [37] or Bayesian estimation [10].

While it is well known [22, 9, 24] that certain specific assumptions about the class conditional distributions lead to log-odds ratio of the class posterior distributions being affine, the literature does not have a characterization of exactly what family of conditional models lead to affine log-odds ratio of the posterior [24, 32, 3, 22]. In particular, it is known that if the conditional distributions all belong to the same exponential family, then the log-odds of the posteriors is affine. However, it is not known if the converse of the above is true, or, in case it is false, what other conditionals lead to affine log-odds of the posterior. In this paper, we give an exact characterization of the family of conditional models that lead to affine log-odds of the posterior. For logistic regression models, we show that the log-odds ratio of the posteriors will be affine *if and only if* the class conditional distributions belong to the same exponential family. We establish a similar result for structured prediction problems

---

by showing that the label sequence posterior will be that of a conditional random field (CRF) *if and only if* the conditionals belong to structured exponential families. We also show that hidden Markov models (HMMs) assume label sequence conditionals to be in a fixed structured exponential family, and hence have significantly higher bias than CRFs.

A wide range of real-life prediction problems have to be solved in an online setting, where data is coming one at a time or in small groups. Predicting events in the stock market [23] or online models for novelty/anomaly detection [41] are practical examples of such settings. Further, unlike certain other predictive modeling settings, one cannot make assumptions about the data since the data may potentially be coming from an adversary, e.g., in the anomaly detection setting. What can be said about the predictive performance of logistic models in such online and potentially adversarial settings? From a more practical consideration, we ask: Do we need to train a logistic regression classifier on the entire historical data every time a new (group of) point(s) comes in, or is there a way to incrementally update the current classifier and still get predictive performance as good as that of retraining the classifier on the entire historical data? The second main result of this paper shows that an incrementally updated Bayesian logistic regression model will have performance comparable to that of the single best logistic regression classifier that can be obtained by training on the entire historic data. The theoretical guarantee holds without any assumption about the data, and hence strongly supports the use of Bayesian models for real-life online setting.

The rest of the paper is organized as follows. We review some background material on exponential families, linear discriminant models, and logistic regression in Section 2. In Section 3, we present and prove our first main result connecting logistic regression and exponential families, and its generalization to structured prediction models connecting conditional random fields and structured exponential families. In Section 4, we present our second main result on relative performance of Bayesian logistic regression in an online and potentially adversarial setting. Section 5 presents a discussion connecting exponential families and integral transforms, in particular Fourier and Laplace transforms, which play an important role in our analysis. We also present a new perspective on the optimality of Bayesian models using Bregman divergences. We conclude in Section 6.

2 Background

In this section, we review some background material on exponential families and logistic regression.

2.1 Exponential Families.
A multivariate parametric family $\mathcal{F}_\psi$ of distributions $\{p_{(\psi,\boldsymbol{\theta})}|\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$ is called an exponential family if each probability density is of the form

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(\langle\mathbf{x}, \boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}))p_0(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d ,$$

where $\mathbf{x}$ is a sufficient statistic for the family and $p_0(\mathbf{x})$ is a non-negative base measure independent of the parameters θ. The function $\psi(\boldsymbol{\theta})$ is known as the *log-partition function* or the *cumulant function* corresponding to the exponential family. If the sufficient statistic $\mathbf{x}$ is minimal, i.e., $\nexists$ non-zero $\mathbf{a} \in \mathbb{R}^d$ such that $\langle\mathbf{a}, \mathbf{x}\rangle = c$ (a constant) $\forall \mathbf{x}$, and $\Theta = \text{dom}(\psi)$ is open, the family is called regular. For a regular exponential family $\mathcal{F}_\psi$, the log-partition function ψ is uniquely determined up to a constant additive term. It can be shown [6] that Θ is a non-empty convex set in $\mathbb{R}^d$ and ψ is a convex function of Legendre type [38].

2.2 Linear Discriminant Analysis.
Linear discriminant analysis (LDA) is one of the traditional approaches to (multiclass) classification [22]. In its general form, one assumes that the class conditional densities all belong an exponential family with each class having different (natural) parameters. If $\boldsymbol{\theta}_h$ is the natural parameters of the conditional density for class h, we have

$$p(\mathbf{x}|h) = \exp(\langle\mathbf{x}, \boldsymbol{\theta}_h\rangle - \psi(\boldsymbol{\theta}_h))p_0(\mathbf{x}) .$$

For classification purposes, one considers the log-odds ratio of the posteriors of the classes under consideration. For classes h, k, with prior probabilities π_h, π_k respectively, the log-odds ratio is

$$\begin{aligned}
\log\frac{P(h|\mathbf{x})}{P(k|\mathbf{x})} &= \log\frac{p(\mathbf{x}|h)}{p(\mathbf{x}|k)} + \log\frac{\pi_h}{\pi_k} \\
&= \langle\mathbf{x}, \boldsymbol{\theta}_h - \boldsymbol{\theta}_k\rangle - (\psi(\boldsymbol{\theta}_h) - \psi(\boldsymbol{\theta}_k)) + \log\frac{\pi_h}{\pi_h} ,
\end{aligned}$$

which is linear in $\mathbf{x}$, thereby justifying the name. Two of the most popular special cases of LDA are Gaussian LDA, which assumes class conditionals to be multivariate Gaussians with different means but equal covariances, and naive-Bayes models [36, 34], which assume factored exponential families typically applied for multinomial conditionals [36]. Some generalizations such as quadratic discriminant analysis (QDA) can also be incorporated in this framework by appropriate choice of sufficient statistics. In particular, QDA reduces to LDA using sufficient statistics based on x_i and $x_i x_j$, $i, j = 1, \ldots, d$.

2.3 Logistic Regression.
In the basic setting of logistic regression, one assumes the existence of a distribution over $X \times \{1, \ldots, k\}$ such that the log-odds ratio

of the class posteriors is affine in $\mathbf{x}$, i.e., $\forall \mathbf{x}$, for any class h and the (arbitrary) reference class k,

$$(2.1) \quad \log\left(\frac{P(h|\mathbf{x})}{P(k|\mathbf{x})}\right) = \langle \mathbf{a}_h, \mathbf{x} \rangle + b_h , \quad h = 1, \ldots, (k-1) ,$$

where $\mathbf{a}_h \in \mathbb{R}^d$ and $b_h \in \mathbb{R}$. For the purposes of this article, we make the definition more precise by introducing *logistic families*. First, note that $P(y|\mathbf{x}) = P(y)p(\mathbf{x}|y)/p(\mathbf{x})$. Further, assuming that $P(h) = \frac{1}{k}$, $\forall h \in \{1, \ldots, k\}$, we have[1]

$$\log\left(\frac{P(\mathbf{x}|h)}{P(\mathbf{x}|k)}\right) = \langle \mathbf{a}_h, \mathbf{x} \rangle + b_h , \quad h = 1, \ldots, (k-1) .$$

We say a family of distributions belongs to a logistic family $\mathcal{F}_{\log}$ if for any two distributions $p_1, p_2 \in \mathcal{F}_{\log}$, their log-odds ratio is affine, i.e., $\log\left(\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}\right) = \langle \mathbf{a}, \mathbf{x} \rangle + b$, for some $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$. It is important to note that the logistic regression assumption implies that the conditional distributions corresponding to the two classes belong to the same logistic family. However, unlike an exponential family, we do not have an explicit characterization of the members of the logistic family.

3 Logistic Models and Exponential Families

Since logistic models make an explicit assumption about the posterior $p(y|\mathbf{x})$, a natural question to ask is: what kind of assumptions about the conditional distributions $p(\mathbf{x}|y)$ lead to the assumed form of the posterior? In the literature, there are several examples of conditional distributions that lead to the desired posterior [22, 9, 24]. In particular, it is known that if the conditional distributions all belong to the same exponential family, then the log-odds of the posteriors is affine. What is not known can be posed as follows: Are exponential families the only conditional distributions that can lead to a log-odds of the posterior that is affine? Several authors have commented on the relationship between exponential family distributions and logistic regression. For example, Anderson [3] and McLachlan [32, page 256] discuss the wide applicability of logistic regression, as a wide variety of families of conditional distributions and perturbations of them satisfy the requirement that the log-odds ratio of the posterior is affine, and go on to give specific examples. Jordan [24], Bishop [9, pages 233-234], and Hastie et. al. [22, pages 103-105] have presented discussions on the relationship as well. Interestingly, there is no precise quantification of exactly

what conditional distributions lead to affine log-odds ratio of the posteriors.

In this section, we give an exact characterization of the class of conditional models that lead to affine log-odds of the posterior. In particular, for logistic regression models, we show that the log-odds ratio of the class posteriors $p(y|\mathbf{x})$ will be affine *if and only if* the class conditional distributions $p(\mathbf{x}|y)$ belong to the same exponential family. Since linear discriminant models are obtained from exponential family conditionals,[2] our result shows that logistic regression models have a significantly lower "bias" [17] than linear discriminant models [22, 34]. We establish a similar result for structured prediction problems by showing that the label sequence posterior distribution will be that of a conditional random field (CRF) *if and only if* the conditionals belong to structured exponential families. We also show that hidden Markov models (HMMs) assume label sequence conditionals to be in a fixed structured exponential family, and hence have significantly higher bias compared to CRFs.

3.1 Laplace Transform of Non-negative Functions. Our analysis hinges on viewing exponentially families as being generated from generalized Laplace transforms [8] of bounded non-negative measures on $\mathbb{R}^d$. Let $P_0(\mathbf{x})$ be a bounded non-negative measure on $\mathbb{R}^d$ with density $p_0(\mathbf{x})$. Then, the generalized Laplace transform of $p_0(\mathbf{x})$ is given by

$$(3.2) \quad L(\boldsymbol{\theta}) = \int_{\mathbf{x}} \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle) p_0(\mathbf{x}) d\mathbf{x} .$$

Since $p_0(\mathbf{x})$ is non-negative, $L(\boldsymbol{\theta})$ is non-negative and can be expressed as $L(\boldsymbol{\theta}) = \exp(\psi(\boldsymbol{\theta}))$ for some function $\psi(\boldsymbol{\theta})$. Plugging this back into (3.2) and rearranging terms, we get

$$(3.3) \quad \exp(\psi(\boldsymbol{\theta})) = \int_{\mathbf{x}} \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle) p_0(\mathbf{x}) d\mathbf{x}$$

$$(3.4) \quad 1 = \int_{\mathbf{x}} \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})) p_0(\mathbf{x}) d\mathbf{x}$$

showing that $\exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta})) p_0(\mathbf{x})$ is a probability density function of an exponential family distribution. Since $P_0(\mathbf{x})$ is a bounded measure, let $k_0 = \int_{\mathbb{R}^d} p_0(\mathbf{x}) d\mathbf{x}$. But $L(\mathbf{0}) = k_0$ so that $\psi(\mathbf{0}) = \log k_0$. Hence, the density function $p_0(\mathbf{x})/k_0$ is a member of this exponential family for $\boldsymbol{\theta} = \mathbf{0}$. In general, an appropriately scaled version of the base measure is always a member the family. In other words, one member of every exponential family

---

is an arbitrary integrable non-negative density function over $\mathbb{R}^d$ and the other members are generated from it. This observation is key to the rest of our analysis.

3.2 Relationship for Logistic Regression.

For simplicity, we first focus on the 2-class classification setting. Let $EXP_2(d)$ denote the class of joint distributions $p(\mathbf{x}, y)$ on $\mathcal{X} \times \{-1, 1\}$ that can be represented as a mixture of two distributions from any fixed exponential family with $\mathcal{X} \subseteq \mathbb{R}^d$ and equal priors, i.e., if $p(\mathbf{x}, y) \in EXP_2(d)$, then $p(\mathbf{x}|y) = p_{(\psi, \theta_y)}(\mathbf{x})$ and $P(-1) = P(+1) = \frac{1}{2}$, where $p_{(\psi, \theta_y)}(\mathbf{x}) \in \mathcal{F}_\psi$. Let $LR_2(d)$ denote the class of joint distributions $p(\mathbf{x}, y)$ on $\mathcal{X} \times \{-1, +1\}$ such that $p(\mathbf{x}|+1)$ and $p(\mathbf{x}|-1)$ belong to the same logistic family $\mathcal{F}_{\log}$ and $P(-1) = P(+1) = \frac{1}{2}$. Note that $p(\mathbf{x}, y) \in LR_2(d)$ is equivalent to the assumptions of logistic regression, i.e., the log-odds ratio is affine in $\mathbf{x}$. In this section, we establish a connection between the class of distributions in $LR_2(d)$ and the class of distributions $EXP_2(d)$. In particular, with a rather simple argument, we show that $LR_2(d) = EXP_2(d)$, so that the assumption made by logistic regression is exactly the same as assuming that the generative model for the observed data is a mixture of any two distributions from any fixed exponential family.

Theorem 1 *For a 2-class classification problem with equal priors, the log-odds ratio of the class posteriors is affine if and only if the class conditional distributions belong to any fixed exponential family. Hence, $LR_2(d) = EXP_2(d)$.*

Proof. First, we prove that $p(\mathbf{x}, y) \in EXP_2(d) \Rightarrow p(\mathbf{x}, y) \in LR_2(d)$ so that $EXP_2(d) \subseteq LR_2(d)$.[3] In other words, if the class conditionals belong any fixed exponential family, then the log-odds ratio of the posterior is affine. Let $p(\mathbf{x}, y)$ be a mixture of two distributions from the exponential family with equal priors and with cumulant function $\psi : \mathbb{R}^d \mapsto \mathbb{R}$. Let $\boldsymbol{\theta}_+, \boldsymbol{\theta}_- \in \mathbb{R}^d$ be the natural parameters of the two distributions. Then,

$$
\begin{aligned}
\log\left(\frac{P(+1|\mathbf{x})}{P(-1|\mathbf{x})}\right) &= \log p_{(\psi, \boldsymbol{\theta}_+)}(\mathbf{x}) - \log p_{(\psi, \boldsymbol{\theta}_-)}(\mathbf{x}) \\
&= \langle \mathbf{a}, \mathbf{x} \rangle + b,
\end{aligned}
$$

where $\mathbf{a} = \boldsymbol{\theta}_+ - \boldsymbol{\theta}_-$, and $b = \psi(\boldsymbol{\theta}_-) - \psi(\boldsymbol{\theta}_+)$. Hence $p(\mathbf{x}, y) \in LR_2(d)$.

Now, we prove the converse that $p(\mathbf{x}, y) \in LR_2(d) \Rightarrow p(\mathbf{x}, y) \in EXP_2(d)$ so that $LR_2(d) \subseteq EXP_2(d)$. In other words, if the log-odds ratio of the class posteriors is affine, then the class conditionals both

---

[3]This part of the result has appeared in the literature [9, 24]. We provide the proof in our notation for completeness.

belong to any fixed exponential family. By definition, $\forall \mathbf{x}$ we have

$$
\begin{aligned}
\log\left(\frac{P(+1|\mathbf{x})}{P(-1|\mathbf{x})}\right) &= \langle \mathbf{a}, \mathbf{x} \rangle + b \\
(3.5) \quad \Rightarrow \quad p(\mathbf{x}|+1) &= \exp(\langle \mathbf{a}, \mathbf{x} \rangle + b)\, p(\mathbf{x}|-1) \,.
\end{aligned}
$$

Now, $p(\mathbf{x}|-1)$ can be either an arbitrary density function or a member of an exponential family. We consider both cases. When $p(\mathbf{x}|-1)$ is an arbitrary density function, let $p(\mathbf{x}|-1) = f(\mathbf{x})$. Then, from (3.5), we have

$$
p(\mathbf{x}|+1) = \exp(\langle \mathbf{a}, \mathbf{x} \rangle + b)f(\mathbf{x}) \,.
$$

Since $p(\mathbf{x}|+1)$ is a probability density function, we have

$$
\begin{aligned}
1 &= \int_\mathbf{x} p(\mathbf{x}|+1)d\mathbf{x} \\
&= \int_\mathbf{x} \exp(\langle \mathbf{a}, \mathbf{x} \rangle + b)f(\mathbf{x})d\mathbf{x} \\
\Rightarrow \quad \exp(-b) &= \int_\mathbf{x} \exp(\langle \mathbf{a}, \mathbf{x} \rangle)f(\mathbf{x})d\mathbf{x} \,.
\end{aligned}
$$

From (3.3) it follows that if ψ is the cumulant function of the exponential family with base measure $f(\mathbf{x})$, then $-b = \psi(\mathbf{a})$. Hence, $p(\mathbf{x}|+1)$ is the exponential family distribution with density $p_{(\psi, \mathbf{a})} \in \mathcal{F}_\psi$. Furthermore, since $p(\mathbf{x}|-1) = p_{(\psi, \mathbf{0})} \in \mathcal{F}_\psi$, the mixture distribution $p(\mathbf{x}, y) \in EXP_2(d)$.

The alternative possibility is that $p(\mathbf{x}|-1)$ is actually an exponential family distribution $p_{(\psi, \boldsymbol{\theta})}$, say with base measure $p_0(\mathbf{x})$, i.e., $p(\mathbf{x}|-1) = \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - \psi(\boldsymbol{\theta}))p_0(\mathbf{x})$. Then,

$$
(3.6) \quad p(\mathbf{x}|+1) = \exp(\langle \mathbf{x}, \mathbf{a} + \boldsymbol{\theta} \rangle + b - \psi(\boldsymbol{\theta}))p_0(\mathbf{x})d\mathbf{x} \,.
$$

Again, since $p(\mathbf{x}|+1)$ is a probability density function

$$
1 = \int_\mathbf{x} \exp(\langle \mathbf{x}, \mathbf{a} + \boldsymbol{\theta} \rangle + b - \psi(\boldsymbol{\theta}))p_0(\mathbf{x})d\mathbf{x} \,,
$$

so that

$$
\begin{aligned}
\exp(-(b - \psi(\boldsymbol{\theta}))) &= \int_\mathbf{x} \exp(\langle \mathbf{x}, \mathbf{a} + \boldsymbol{\theta} \rangle)p_0(\mathbf{x})d\mathbf{x} \\
&= \exp(\psi(\mathbf{a} + \boldsymbol{\theta})) \,.
\end{aligned}
$$

Hence, $-(b - \psi(\boldsymbol{\theta})) = \psi(\mathbf{a} + \boldsymbol{\theta})$ implying that $p(\mathbf{x}|+1)$ is the exponential family distribution with density $p_{(\psi, \mathbf{a}+\boldsymbol{\theta})} \in \mathcal{F}_\psi$. Since, $p(\mathbf{x}|-1) = p_{(\psi, \boldsymbol{\theta})} \in \mathcal{F}_\psi$, the mixture density $p(\mathbf{x}, y) \in EXP_2(d)$. Hence, $LR_2(d) \subseteq EXP_2(d)$. That completes the proof. ∎

The above argument can be easily extended to the case when the class priors are not equal—there is one additive terms in addition to b, but the main argument remains unchanged. The extension to the multi-class case

is also straightforward. For the uniform class prior case, with suitable extensions of our definitions to $EXP_k(d)$ (mixture of k exponential family distributions, one corresponding to each class) and $LR_k(d)$ (the pairwise log-odds ratios, and hence log-conditional ratios, are affine), we have the following result.

Theorem 2 *For a k-class classification problem with equal priors, the pairwise log-odds ratio of the class posteriors is affine if and only if the class conditional distributions belong to any fixed exponential family. Hence, $LR_k(d) = EXP_k(d)$.*

The proof is identical to that of 2 classes. As before, $EXP_k(d) \subseteq LR_k(d)$ essentially follows from the definition of exponential family. For the converse, i.e., $LR_k(d) \subseteq EXP_k(d)$, if one assumes that one of the conditional distributions $p(\mathbf{x}|h)$ belongs to an exponential family, then since the log-odds of all other conditionals are affine with respect to $p(\mathbf{x}|h)$, it follows that all conditionals belong to the same family. On the other hand, if one assumes that one of the conditionals $p(\mathbf{x}|h)$ is an arbitrary density function, then all conditionals belong to the exponential family generated by that density. The extension to the case of unequal priors in the multiclass setting is also straightforward.

3.3 Relationship for Conditional Random Fields. Conditional Random Fields (CRFs) [27] are logistic models on structures, such as sequences and graphs. Our discussion focuses on the case of sequences, where any labeled training exampled $(\mathbf{x}, \mathbf{y})$ is a sequence of labels applied to a sequence of observations, such as part-of-speech tagging of English text. The extension to the general case of structured prediction is straightforward.

In CRFs, the probability of a label sequence $\mathbf{y}$ given an observation sequence $\mathbf{x}$ is given by

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\lambda})} \exp\left(\sum_{j=1}^{d} \lambda_j F_j(\mathbf{y}, \mathbf{x})\right) ,$$

where, for a n-length sequence, $F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i)$, where each $f_j(y_{i-1}, y_i, \mathbf{x}, i)$ can be a state-potential function $s(y_i, \mathbf{x}, i)$ or a transition-potential function $t(y_i, y_{i-1}, \mathbf{x}, i)$, and $Z(\mathbf{x}, \boldsymbol{\lambda})$ is a normalization constant or cumulant given by

$$Z(\mathbf{x}, \boldsymbol{\lambda}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp\left(\sum_{j=1}^{d} \lambda_j F_j(\mathbf{y}', \mathbf{x})\right) ,$$

where $\mathcal{Y}$ is the set of all possible label sequences. Since $|\mathcal{Y}|$ is typically exponential in n, carefully designed algorithms [27, 40] are necessary for maximum-likelihood

estimation of $\boldsymbol{\lambda}$. Written in a similar form, a k-class logistic regression formulation only has k terms in the cumulant, making simple gradient-based solutions possible for maximum-likelihood estimation of the parameters.

Consider a class of generative exponential family models that use the feature vector $\Phi(\mathbf{x}, \mathbf{y}) = [F_1(\mathbf{x}, \mathbf{y}) \ldots F_d(\mathbf{x}, \mathbf{y})]^T$ as a sufficient statistic. Then, the joint probability of the observation sequence $\mathbf{x}$ and the label sequence $\mathbf{y}$ is given by

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = \exp(\langle\Phi(\mathbf{x}, \mathbf{y}), \boldsymbol{\theta}\rangle - \Psi(\boldsymbol{\theta}))q_0(\Phi(\mathbf{x}, \mathbf{y})) .$$

We focus on a special class of base measures for which $q_0(\Phi(\mathbf{x}, \mathbf{y})) = p_0(\mathbf{x})$, and call the corresponding joint distributions $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$ as belonging to *structured exponential families*. Then, we have the following result.

Theorem 3 *For a sequence classification problem, the posterior probability of a label sequence is that of a conditional random field if and only if the joint distributions belongs to any structured exponential family.*

Proof. First we prove the "if" part. If the joint distributions belong to a structured exponential family, then

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = \exp(\langle\Phi(\mathbf{x}, \mathbf{y}), \boldsymbol{\theta}\rangle - \Psi(\boldsymbol{\theta}))p_0(\mathbf{x}) .$$

Then, by Bayes theorem,

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) &= \frac{p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{\sum_{\mathbf{y}' \in \mathcal{Y}} p(\mathbf{x}|\mathbf{y}', \boldsymbol{\theta})p(\mathbf{y}'|\boldsymbol{\theta})} \\
&= \frac{p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})}{\sum_{\mathbf{y}' \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}'|\boldsymbol{\theta})} \\
&\overset{(a)}{=} \frac{1}{Z(\mathbf{x}, \boldsymbol{\theta})} \exp(\langle\Phi(\mathbf{x}, \mathbf{y}), \boldsymbol{\theta}\rangle) ,
\end{aligned}
$$

where $Z(\mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\langle\Phi(\mathbf{x}, \mathbf{y}'), \boldsymbol{\theta}\rangle)$, and (a) follows since the $\exp(-\Psi(\boldsymbol{\theta}))p_0(\mathbf{x})$ terms cancel out. Clearly, the last expression is the posterior corresponding to a CRF with $\lambda = \boldsymbol{\theta}$.

Now, we prove the "only if" part. Since the posterior probability of any label sequence is that of a CRF, for any pair of label-sequences $\mathbf{y}, \mathbf{y}_0$, we have

$$\log\left(\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda})}{p(\mathbf{y}_0|\mathbf{x}, \boldsymbol{\lambda})}\right) = \langle\Phi(\mathbf{x}, \mathbf{y}) - \Phi(\mathbf{x}, \mathbf{y}_0), \boldsymbol{\lambda}\rangle .$$

Hence, we have

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda}) = \exp(\langle\Phi(\mathbf{x}, \mathbf{y}) - \Phi(\mathbf{x}, \mathbf{y}_0), \boldsymbol{\lambda}\rangle)p(\mathbf{x}, \mathbf{y}_0|\boldsymbol{\lambda}) .$$

If $p(\mathbf{x}, \mathbf{y}_0|\boldsymbol{\lambda})$ is a structured exponential family distribution, then

$$p(\mathbf{x}, \mathbf{y}_0|\boldsymbol{\lambda}) = \exp(\langle\Phi(\mathbf{x}, \mathbf{y}_0), \boldsymbol{\lambda}\rangle - \Psi(\boldsymbol{\lambda}))p_0(\mathbf{x}) .$$

Plugging this back in the expression for $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda})$, we get

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda}) = \exp(\langle \Phi(\mathbf{x}, \mathbf{y}), \boldsymbol{\lambda} \rangle - \Psi(\boldsymbol{\lambda}))p_0(\mathbf{x}) \ .$$

Thus, if one of the joint distributions is a structured exponential family distribution, then all of them must be structured exponential family distributions. Next, we consider the case where $p(\mathbf{x}, \mathbf{y}_0|\boldsymbol{\lambda})$ is an arbitrary distribution of the form $f_0(\mathbf{x})\exp(b)$, where $f_0(\mathbf{x}) \geq 0$ and b is a real constant, so that $\int_{\mathbf{x}} f_0(\mathbf{x})\exp(b)d\mathbf{x} = 1$. Note that the assumed form is general, since $p(\mathbf{x}, \mathbf{y}_0|\boldsymbol{\lambda})$ is a probability distribution, and $\mathbf{y}_0$ is fixed. Now, we define new features $\Phi_0(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}, \mathbf{y}) - \Phi(\mathbf{x}, \mathbf{y}_0)$. Then, we have

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda}) = \exp(\langle \Phi_0(\mathbf{x}, \mathbf{y}), \boldsymbol{\lambda} \rangle + b)f_0(\mathbf{x}) \ .$$

Since $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda})$ is a probability distribution,

$$
\begin{aligned}
1 &= \int_{\mathbf{x}, \mathbf{y}} \exp(\langle \Phi_0(\mathbf{x}, \mathbf{y}), \boldsymbol{\lambda} \rangle + b)f_0(\mathbf{x})d\mathbf{x} \\
\exp(-b) &= \int_{\mathbf{x}, \mathbf{y}} \exp(\langle \Phi_0(\mathbf{x}, \mathbf{y}), \boldsymbol{\lambda} \rangle)f_0(\mathbf{x})d\mathbf{x} \ .
\end{aligned}
$$

If $\Psi_0(\boldsymbol{\lambda})$ is the cumulant function of an exponential family with sufficient statistics $\Phi_0(\mathbf{x}, \mathbf{y})$ and base measure $f_0(\mathbf{x})$, then $b = -\Psi_0(\boldsymbol{\lambda})$. Hence,

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda}) = \exp(\langle \Phi_0(\mathbf{x}, \mathbf{y}), \boldsymbol{\lambda} \rangle - \Psi_0(\boldsymbol{\lambda}))f_0(\mathbf{x}) \ .$$

Since $\mathbf{y}$ is arbitrary, all joint distributions $p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda})$, are structured exponential family distributions. That completes the proof. ∎

Theorem 3 exactly characterizes the "bias" of CRFs as compared to specific generative models that belong to structured exponential families. We wrap up this section by showing that the widely used Hidden Markov Model (HMM) [37] is a specific instance of a structured exponential family distribution. Recall that according to a HMM, the joint probability of an observation and label sequence is given by

$$p(\mathbf{x}, \mathbf{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} p(y_i|y_{i-1})p(x_i|y_i) \ .$$

For k states, and n observables, in an over-complete representation the parameters $\boldsymbol{\lambda}$ consist of the $(k + k^2 + nk)$ components, of which k are based on state initiation probabilities $p(h), h = 1, \ldots, k$; k^2 are based on state-transition probabilities $p(h|\ell), h, \ell = 1, \ldots, k$; and nk are based on emission probabilities $p(i|h), h = 1, \ldots, k, i = 1, \ldots, n$. In particular, we have

$$\lambda_h = \log p(h), \quad \lambda_{h\ell} = \log p(h|k), \quad \lambda_{ih} = \log p(i|h) \ .$$

The first set of k sufficient statistics are boolean, indicating whether h was the start state; the set of k^2 statistics are counts of state transitions; and the set of nk statistics are again counts of emissions. With $q_0(\mathbf{x}, \mathbf{y}) = p_0(\mathbf{x})$ being the counting measure, and corresponding $\Psi(\boldsymbol{\lambda}) = 0$, it is straightforward to see that a HMM is a specific example of structured exponential family distribution. From Theorem 3 it follows that the posterior distribution of label sequences has the form of that of a CRF. However, CRFs are significantly more robust since several other joint distributions lead to the same posterior.

4 Online Performance of Logistic Models

In several real life applications, additional data becomes available over time. Such a situation can arise because the data actually gets created over time, e.g., text categorization applications on the web, or additional measurements are being made, e.g., satellite images of earths surface. In such a setting, a natural question to ask is: Do we have to re-train the classifier on the entire dataset, or can we make incremental updates to the existing classifier based on the new data and still get good performance? There is yet another reason that makes the capability of incremental training desirable. Consider training a classifier such as logistic regression or support vector machine on a very large dataset, consisting of several million points. The computation and main memory storage necessary for training can be prohibitive on several commonly available computation infrastructures. Instead, if one does incremental training using one-point or, a small group of points at a time, the approach will be feasible both in terms of computation and storage. However, the crucial question becomes: Will the incrementally trained classifier be comparable to the best classifier trained in batch? In this section, we give an answer to this question for logistic regression models. Without making any statistical assumptions about the data, we show that incrementally trained Bayesian logistic regression models have performance comparable to the best logistic regression model that can be trained in batch, looking back at the entire dataset.

In some situations, Bayesian models are preferred since they can naturally incorporate prior belief about the parameter values [21]. The regularization provided by Bayesian models are critical, particularly for high-dimensional problems [33]. Bayesian logistic regression has gained popularity in the recent years for a wide variety of applications[26, 29]. We consider a Bayesian prior over all logistic regression models, where for any input $\mathbf{x}$, the predictor corresponding to $\mathbf{w}$ assigns

probability

$$p(y|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-y\langle \mathbf{w}, \mathbf{x}\rangle)} \ ,$$

where $y \in \{-1, +1\}$. [4] Then, we assume a Gaussian prior distribution P_0 with density $p_0(\mathbf{w}) = N(\mathbf{w}; \mathbf{0}, \mathbb{I})$ over all predictors $\mathbf{w} \in \mathbb{R}^d$.

Let $S_T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)\}$ be a sequence of examples presented to the Bayesian classifier. Prediction proceeds in iterations, where in iteration t, input $\mathbf{x}_t$ is presented, and each of the predictors generates a prediction. If y_t is the true label, then each predictor $\mathbf{w}$ incurs a loss of

$$\ell_t(\mathbf{w}) = -\log p(y_t|\mathbf{x}_t, \mathbf{w}) \ .$$

Further, the distribution over $\mathbf{w}$ is updated based on the loss incurred by individual predictors. If P_{t-1} is the distribution on the predictors after seeing the first t examples $S_{t-1} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t-1}, y_{t-1})\}$, by a direct application of Bayes rule, the updated distribution P_t after observing $(\mathbf{x}_t, y_t)$ has a density

$$\begin{aligned}
p_t(\mathbf{w}) &\equiv p(\mathbf{w}|S_t) = p(\mathbf{w}|S_{t-1}, \mathbf{x}_t, y_t) \\
&= \frac{p(y_t|\mathbf{x}_t, \mathbf{w}, S_{t-1})p(\mathbf{w}|\mathbf{x}_t, S_{t-1})}{p(S_t)} \\
&= \frac{p(y_t|\mathbf{x}_t, \mathbf{w})p(\mathbf{w}|S_{t-1})}{p(S_t)} = \frac{p(y_t|\mathbf{x}_t, \mathbf{w})p_{t-1}(\mathbf{w})}{p(S_t)} \ ,
\end{aligned}$$

since y_t is independent of S_{t-1} given $\mathbf{x}_t, \mathbf{w}$, i.e., individual predictors are memory-less, and $p(\mathbf{w}|\mathbf{x}_t, S_{t-1}) = p(\mathbf{w}|S_{t-1})$ as we update the distribution over $\mathbf{w}$ only after getting the label on the current input $\mathbf{x}_t$.

At iteration t, on receiving input $\mathbf{x}_t$, the Bayesian logistic regression model predicts

$$\begin{aligned}
p(y|\mathbf{x}_t, S_{t-1}) &= E_{\mathbf{w} \sim p_{t-1}}[p(y|\mathbf{x}_t, \mathbf{w})] \\
&= \int_{\mathbf{w}} \frac{p_{t-1}(\mathbf{w})}{1 + \exp(-y\langle \mathbf{w}, \mathbf{x}\rangle)} \ d\mathbf{w} \ .
\end{aligned}$$

If y_t is the true label, the Bayesian model incurs a loss of $-\log p(y_t|\mathbf{x}_t, S_{t-1})$. Hence, after such iterative prediction on S_T, the total loss incurred by Bayesian logistic regression is

$$L_{BLR}(S_T) = \sum_{t=1}^{T} -\log p(y_t|\mathbf{x}_t, S_{t-1}) \ .$$

Let Q be any fixed distribution over the predictors $\mathbf{w} \in \mathbb{R}^d$. Consider a Gibbs prediction scheme that

---

samples $\mathbf{w} \sim Q$, and then predicts on $\mathbf{x}_t$ based on $\mathbf{w}$. The expected loss incurred by Q at iteration t is simply

$$\ell_t(Q) = E_{\mathbf{w} \sim Q}[-\log p(y_t|\mathbf{x}_t, \mathbf{w})] \ ,$$

and the total loss

$$L_Q(S_T) = \sum_{t=1}^{T} \ell_t(Q) = E_{\mathbf{w} \sim Q}\left[\sum_{t=1}^{T} -\log p(y_t|\mathbf{x}_t, \mathbf{w})\right] \ .$$

Then, by a convexity argument (for example, see [4, 25, 20] for details) one can establish the following result, where we denote S_T by S for simplicity.

Theorem 4 ([4, 25]) *For any distribution Q on $\mathbf{w} \in \mathbb{R}^d$,*

$$L_{BLR}(S) \le L_Q(S) + KL(Q\|P_0) \ .$$

We use Theorem 4 to prove the following dimension independent and data size independent relative loss bound for Bayesian logistic regression. For simplicity, we assume that $\|\mathbf{x}_t\| = 1, \forall t$.

Theorem 5 *Let $\mathbf{w}^*$ be any weight vector used for logistic regression with cumulative loss $L_{\mathbf{w}^*}(S) = \sum_{t=1}^{T} \ell_t(\mathbf{w}^*) = \sum_{t=1}^{T} -\log p(y_t|\mathbf{x}_t, \mathbf{w}^*)$. Then, for any sequence of examples S,*

$$L_{BLR}(S) \le 2L_{\mathbf{w}^*}(S) + \frac{\|\mathbf{w}^*\|^2}{2} \ .$$

Proof. We prove the result by a derandomization argument applied to Theorem 4. Consider a distribution Q on the predictors $\mathbf{w}$ with density $q(\mathbf{w}) = N(\mathbf{w}; \mathbf{w}^*, \mathbb{I})$. Then, since $p_0(\mathbf{w}) = N(\mathbf{w}; \mathbf{0}, \mathbb{I})$, we have

$$\begin{aligned}
KL(Q\|P_0) &= E_{\mathbf{w} \sim Q}\left[-\frac{\|\mathbf{w} - \mathbf{w}^*\|^2}{2} + \frac{\|\mathbf{w}\|^2}{2}\right] \\
&= E_{\mathbf{w} \sim Q}[\langle \mathbf{w}, \mathbf{w}^*\rangle] - \frac{\|\mathbf{w}^*\|^2}{2} \\
&= \frac{\|\mathbf{w}^*\|^2}{2} \ .
\end{aligned}$$

We complete our proof by showing that $L_Q(S) \le 2L_{\mathbf{w}^*}(S)$. In fact, we show that the instantaneous losses satisfy the desired relationship, i.e., $\ell_t(Q) \le 2\ell_t(\mathbf{w}^*)$. Note that

$$\begin{aligned}
\ell_t(Q) &= E_{\mathbf{w} \sim Q}[-\log p(y_t|\mathbf{x}_t, \mathbf{w})] \\
&= E_{\mathbf{w} \sim Q}[\log(1 + \exp(-y_t\langle \mathbf{x}_t, \mathbf{w}\rangle))] \\
&\le \log(1 + E_{\mathbf{w} \sim Q}[\exp(-y_t\langle \mathbf{x}_t, \mathbf{w}\rangle)]) \ ,
\end{aligned}$$

by an application of Jensen's inequality since log is a concave function. Let $z = \langle \mathbf{x}_t, \mathbf{w}\rangle$. Since $\mathbf{w} \sim Q$ with

---

[4]Note that $\langle \mathbf{a}, \mathbf{x}\rangle + b$ can always be written as $\langle \mathbf{w}, \mathbf{x}\rangle$ by augmenting $\mathbf{x}$ by a constant element.

density $q(\mathbf{w}) = N(\mathbf{w}; \mathbf{w}^*, \mathbb{I})$, $z \sim N(z; z^*, 1)$, where $z^* = \langle \mathbf{x}_t, \mathbf{w}^* \rangle$, because the projection of multivariate identity covariance normal random variable along any unit vector has a univariate unit variance normal distribution. Then,

$$\ell_t(Q) \leq \log(1 + E_{z \sim N(z^*, 1)}[\exp(-y_t z)])$$

Recall that the moment generating function of the univariate unit variance normal distribution is given by

$$M(t) = E_{z \sim N(\mu, 1)}[\exp(tz)] = \exp(t\mu + t^2/2) .$$

Replacing $\mu = z^*$ and $t = -y_t$, we obtain

$$
\begin{aligned}
\ell_t(Q) &\leq \log(1 + \exp(-y_t z^* + 1/2)) \\
&\leq \log(1 + 2\exp(-y_t z^*)) \\
&\leq \log(1 + 2\exp(-y_t z^*) + \exp(-2y_t z^*)) \\
&= 2\log(1 + \exp(-y_t z^*)) \\
&= 2\log(1 + \exp(-y_t \langle \mathbf{x}_t, \mathbf{w}^* \rangle)) \\
&= 2\ell_t(\mathbf{w}^*) .
\end{aligned}
$$

Adding both sides over $t = 1, \ldots, T$, completes the proof. ∎

Theorem 5 shows that the loss incurred by the incrementally maintained Bayesian logistic regression classifier will be within twice the loss of the best logistic regression classifier that can be trained on the entire data, plus an additional constant that is independent of data properties such as number of data points or the dimensionality of the data. In other words, the incremental Bayesian model "tracks" the best logistic classifier, even though the best classifier can change as more points are added.

As an application of Lemma 4, [25, Theorem 2.2] recently proved a relative loss bound for generalized linear models, and applied the bound to the special case of logistic regression [25, Theorem 3.3]. In our notation, their bound for Bayesian logistic regression can be written as
(4.7)
$$L_{BLR}(S) \leq L_{\mathbf{w}^*}(S) + \frac{\|\mathbf{w}^*\|^2}{2\nu^2} + \frac{d}{2}\log\left(1 + \frac{T\nu^2}{d}\right) ,$$

where $\nu \leq 0.5$. The bound in (4.7), derived using a very different argument from ours, has a constant of 1 on $L_{\mathbf{w}^*}(S)$, instead of 2 as in Theorem 5. However, their bound has a dependency on the number of data points T as well as the data dimensionality d. In particular, for a fixed d, the bound degrades approximately according to $\log(T)$. Further, for very high-dimensional data, i.e., as $d \to \infty$, $(1 + a/d)^d \to \exp(a), \forall a \in \mathbb{R}$, and hence

$$L_{BLR} \leq L_{\mathbf{w}^*}(S) + \frac{\|\mathbf{w}^*\|^2}{2\nu^2} + \frac{T\nu^2}{2} ,$$

so the bound degrades linearly as the number of samples. Our bound is independent of the data dimensionality as well as the number of data points, and demonstrates that Bayesian logistic regression is capable of tracking the logistic best classifier irrespective of the dimensionality of the data and the number of samples. We believe that both bounds are useful and bring out different aspects of Bayesian logistic regression. For quantitative purposes, one can always use the minimum of the two.

5 Discussion

Our first set of results make use of a Laplace transform perspective of exponential family distributions. In Section 5.1, we take a more detailed look at the integral transform viewpoint, including both Laplace and Fourier transforms, and review connections between positive definite functions and probability measures. Our second main result makes use of a Bayesian model applied to logistic regression. In Section 5.2, we present a new perspective on the optimality of Bayesian models in terms of Bregman divergences [5].

5.1 The Integral Transform Viewpoint. In this subsection, we review some important results from harmonic analysis that are relevant to our current analysis of exponential and logistic family distributions. Further, the results help in connecting exponential family distributions to positive definite functions, which are currently widely studied in data mining.

5.1.1 Laplace Transform of Non-negative Functions. In Section 3.1, we presented a Laplace transform viewpoint of exponential family distributions. In particular, we saw that the cumulant function is the generalized Laplace transform of the base measure, which is an arbitrary non-negative function. Now, we review an important characterization of Laplace transform of non-negative functions in terms of exponentially convex functions, which are defined below.

Definition 1 A function $f : \Theta \mapsto \mathbb{R}_{++}$, $\Theta \subseteq \mathbb{R}^d$ is called exponentially convex if the kernel $K_f(\alpha, \beta) = f(\alpha + \beta)$, with $\alpha + \beta \in \Theta$, satisfies

$$\sum_{i=1}^{n} \sum_{j=1}^{n} K_f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) u_i \bar{u}_j \geq 0 ,$$

for any set $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n\} \subseteq \Theta$ with $\boldsymbol{\theta}_i + \boldsymbol{\theta}_j \in \Theta$, $\forall i, j$, and $\{u_1, \cdots, u_n\} \subset \mathbb{C}$ ($\bar{u}_j$ denotes the complex conjugate of u_j), i.e, the kernel K_f is positive semi-definite.

It is well known that the logarithm of an exponentially convex function is a convex function [2]. The following

result, due to [16], relates exponentially convex functions to Laplace transforms of bounded non-negative measures, and hence to exponential family distributions.

Theorem 6 ([16]) *Let* $\Theta \subseteq \mathbb{R}^d$ *be an open convex set. A necessary and sufficient condition that there exists a unique, bounded, non-negative measure* ν *such that* $L : \Theta \mapsto \mathbb{R}_{++}$ *can be represented as*

$$(5.8) \qquad L(\boldsymbol{\theta}) = \int_{\mathbf{x} \in \mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle) \, d\nu(\mathbf{x}) \ ,$$

is that L *is continuous and exponentially convex.*

From Theorem 6, it is easy to see that in the context of exponential family distributions, the exponentially convex function is the partition function, and the logarithm of the exponentially convex function is indeed the cumulant function $\psi(\boldsymbol{\theta})$. Also, it is important to note that exponentiation of arbitrary convex functions do not give exponentially convex functions.[5]

5.1.2 Fourier Transform of Non-negative Functions. Since Laplace transform of non-negative functions give rise to exponential family distributions, a natural question to ask is: Is there a good way to characterize Fourier transforms of non-negative functions? The answer is quite simple: Fourier transforms of non-negative functions give rise to positive definite functions, which are currently extensively used in data mining. In fact, the original result, due to Bochner [11], goes one step further to show that the only way to obtain a positive definitive function is by taking Fourier transform of a non-negative function. We state the result in the form it appears in [14, 18, 19].

Theorem 7 ([11]) *A necessary and sufficient condition that a function* $F : \mathbb{R}^k \mapsto \mathbb{C}$ *can be represented as*

$$(5.9) \qquad F(\mathbf{y}) = \int_{\mathbb{R}^d} \exp(i\langle \mathbf{x}, \mathbf{y} \rangle) \, d\nu(\mathbf{x}) \ ,$$

where ν *is a bounded non-negative function, is that* F *is continuous and positive definite.*

5.1.3 Laplace to Fourier and Back. The results of Theorems 6 and 7 leads to another tempting conclusion that Laplace and Fourier transforms of a non-negative function can be obtained from one another by a simple plug-in procedure, i.e., replacing $\boldsymbol{\theta}$ by $i\mathbf{y}$ to go from Laplace to Fourier, and replacing $\mathbf{y}$ by $-i\boldsymbol{\theta}$ to go from

Fourier to Laplace. Interestingly, as [19] showed, the plug-in procedure is correct given one extra condition is satisfied: the positive definite function $F(\mathbf{y})$ has to be entire. Recall that a complex valued function $F(\mathbf{y}), \mathbf{y} \in \mathbb{R}^d$, is *entire* if it can be uniquely extended to a necessarily unique analytic function $F(\mathbf{z}), \mathbf{z} \in \mathbb{C}^d$ [1]. With this extra condition, [19] recently established a bijection between entire positive definite functions and exponentially convex functions.

Theorem 8 ([19]) *If* $L(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^d$, *is an exponentially convex function, then it is entire, and* $F(\mathbf{y}) = L(i\mathbf{y})$ *is a positive definite function. Conversely, if* $F(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^d$, *is an entire positive definite function, then* $L(\boldsymbol{\theta}) = F(-i\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^d$, *is an exponentially convex function.*

5.1.4 Implications. The results of Theorems 6, 7 and 8 have important implications for data mining and predictive modeling, where exponential family distributions and positive definite functions are playing an increasingly important role. In particular, they can be used to design natural kernels from exponential family models of data, or conversely, obtain probabilistic models using kernels learned from data. For example, for identity covariance Gaussian distributions, $p_0(\mathbf{x}) = \exp(-\|\mathbf{x}\|^2/2)$ and we have

$$\int_{\mathbb{R}^d} \exp(\langle \mathbf{x}, \boldsymbol{\theta} \rangle) \exp(-\|\mathbf{x}\|^2/2) \, d\mathbf{x} = \exp(\|\boldsymbol{\theta}\|^2/2) \ .$$

From Theorem 6, $L(\boldsymbol{\theta}) = \exp(\|\boldsymbol{\theta}\|^2/2)$ is an exponentially convex function. Then, Theorem 8 implies that $L(\boldsymbol{\theta})$ is entire and $F(\mathbf{y}) = L(i\mathbf{y}) = \exp(-\|\mathbf{y}\|^2/2)$ is a positive definite function. Then, for a given set of data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, $F(\mathbf{x}_i - \mathbf{x}_j)$ forms a valid kernel that can be used by kernel methods. Of course, we clearly identify this particular kernel to be the widely used radial basis function (RBF) kernel. Note that such kernels can be constructed for every choice of non-negative function $p_0(\mathbf{x})$ on $\mathbb{R}^d$.

5.2 Bregman Optimality of Bayesian Models. Bayesian models have been extensively studied in statistics [21], and are getting increasing popular in large scale data mining and modeling applications [26, 29, 10]. Further, recent years have been seen the development of theoretical arguments in favor of Bayesian models [30, 4]. A derandomization of the Bayesian argument has led to the first margin-based dimension-independent error-rate bound for support vector machines and other margin-maximizing classifiers [31, 28].

In this section, we present a new perspective on the optimality of Bayesian models in terms of Bregman

---

[5]An argument establishing this fact is outside the scope of the current paper.

divergences [12, 13, 5], which are a broad class of divergence functions that include squared Euclidean distance, relative-entropy, etc., as special cases. Let $\phi : \mathcal{S} \mapsto \mathbb{R}$ be a strictly convex function defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ such that ϕ is differentiable on $\mathrm{ri}(\mathcal{S})$, the relative interior of $\mathcal{S}$ [38]. The Bregman divergence $d_\phi : \mathcal{S} \times \mathrm{ri}(\mathcal{S})$ is defined as

$$d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla \phi(\mathbf{y}) \rangle \ ,$$

where $\nabla \phi(\mathbf{y})$ is the gradient of ϕ at $\mathbf{y}$.

For the purposes of our analysis, we consider a multi-class classification settings and focus on a set of predictors $f_\alpha(\mathbf{x})$, parameterized by α. Note that α can be a parameter of a model such as logistic regression or support vector machine, or simply an index running over arbitrary classifiers [39]. For each choice of α, $f_\alpha(\mathbf{x})$ makes a deterministic or stochastic prediction. In either case, the prediction can be represented as a probability distribution $p_\alpha(\mathbf{x})$ over the k-classes, where, in the deterministic case, only one component is 1 and the rest are 0s. We assume $p_0(\alpha)$ to be a prior probability distribution over all possible α. Then, a Bayesian predictor will predict based on a weighted vote of all the predictors so that

$$(5.10) \qquad p_{\mathrm{Bayes}}(\mathbf{x}) = \int_\alpha p_\alpha(\mathbf{x}) p_0(\alpha) d\alpha \ .$$

Let $d_\phi(\cdot, \cdot)$ be any Bregman divergence well defined on Δ^{k-1}, the simplex embedded in $\mathbb{R}^k$. Then, we have the following optimality result.

Theorem 9 *Let $p^* \in \Delta^{k-1}$ be the distribution over the classes that minimize expected disagreement with all other predictors $p_\alpha(\mathbf{x})$, where disagreement is measured by any Bregman divergence, i.e.,*

$$p^*(\mathbf{x}) = \underset{p \in \Delta}{\mathrm{argmin}} \ E_{\alpha \sim p_0}[d_\phi(p_\alpha(\mathbf{x}), p(\mathbf{x}))] \ .$$

Then, $p^*(\mathbf{x})$ is unique and is given by

$$p^*(\mathbf{x}) = p_{\mathrm{Bayes}}(\mathbf{x}) \ .$$

Proof. Let $p'(\mathbf{x}) \in \Delta^{k-1}$ be an arbitrary distribution over the classes. Then,

$$\begin{aligned} &E_{\alpha \sim p_0}[d_\phi(p_\alpha(\mathbf{x}), p'(\mathbf{x}))] - E_{\alpha \sim p_0}[d_\phi(p_\alpha(\mathbf{x}), p_{\mathrm{Bayes}}(\mathbf{x}))] \\ &= \phi(p_{\mathrm{Bayes}}(\mathbf{x})) - \phi(p'(\mathbf{x})) \\ &\quad - E_{\alpha \sim p_0}[\langle p_\alpha(\mathbf{x}) - p'(\mathbf{x}), \nabla \phi(p'(\mathbf{x})) \rangle] \\ &\quad + E_{\alpha \sim p_0}[\langle p_\alpha(\mathbf{x}) - p_{\mathrm{Bayes}}(\mathbf{x}), \nabla \phi(p_{\mathrm{Bayes}}(\mathbf{x})) \rangle] \\ &\overset{(a)}{=} \phi(p_{\mathrm{Bayes}}(\mathbf{x})) - \phi(p'(\mathbf{x})) - \langle p_{\mathrm{Bayes}}(\mathbf{x}) - p'(\mathbf{x}), \nabla \phi(p'(\mathbf{x})) \rangle \\ &= d_\phi(p_{\mathrm{Bayes}}(\mathbf{x}), p'(\mathbf{x})) \ \geq \ 0, \end{aligned}$$

where (a) follows from the linearity of expectation and since

$$E_{\alpha \sim p_0}[p_\alpha(\mathbf{x})] = \int_\alpha p_\alpha(\mathbf{x}) p_0(\alpha) d\alpha = p_{\mathrm{Bayes}} \ .$$

Hence, for all $p' \in \Delta^{k-1}$, we have

$$E_{\alpha \sim p_0}[d_\phi(p_\alpha(\mathbf{x}), p_{\mathrm{Bayes}}(\mathbf{x}))] \leq E_{\alpha \sim p_0}[d_\phi(p_\alpha(\mathbf{x}), p'(\mathbf{x}))] \ ,$$

so that $p_{\mathrm{Bayes}}(\mathbf{x})$ is a minimizer of the expected disagreement. From the strict convexity of ϕ, it follows that $d_\phi(p_{\mathrm{Bayes}}(\mathbf{x}), p'(\mathbf{x}))$ is 0 if and only if $p'(\mathbf{x}) = p_{\mathrm{Bayes}}(\mathbf{x})$, implying that $p_{\mathrm{Bayes}}(\mathbf{x})$ is the unique minimizer. ∎

Theorem 9 gives a new perspective on the optimality of Bayesian models. In particular, it shows that weighted vote used by a Bayesian predictor is the one that has minimum expected disagreement with all the individual predictors. Interestingly, the proof argument is a straight-forward extension of a similar argument due to [5], where it was used to study partitional clustering using Bregman divergences.

6 Conclusion

Over the years, logistic models have been successfully used on a wide variety of practical problems. Logistic regression has been one of the most widely used probabilistic classification techniques due to its robustness and high quality predictions. Further, there are several other advantages to logistic regression: (i) It can be extended straightforwardly to the multi-class case [22], which can be problematic for some of the other classification methods; (ii) It can be extended to incorporate kernels since all relevant information are in the form of dot-products between the parameters and the data; and (iii) It makes a probabilistic prediction which may be desirable to get an idea in the confidence or entropy of the prediction. Conditional random fields are logistic models for sequences/graphs which have gained wide popularity in the recent years. In this paper, we have given an exact quantification of the bias [17] of logistic models. In particular, we have shown that the log-odds ratio of the posterior probabilities in affine if and only if each of the class conditional densities belong to the same exponential families. Thus, making a specific exponential family generative model assumption has significantly higher bias compared to the corresponding logistic models. Our results further explain some of the existing experimental as well as theoretical results comparing logistic regression to generative models such as naive-Bayes [34].

Bayesian logistic regression has gained popularity in the recent past due to both theoretical developments [25, 26] as well as empirical success in large scale

real-life problems [29, 26]. We have analyzed the performance of Bayesian logistic regression predictors in an online setting, where data points are coming one at a time and may be potentially be generated by an adversary. Without making any assumptions about the data, we have shown the cumulative loss incurred by the incrementally updated Bayesian logistic regression classifier is comparable to the loss incurred by the best single batch logistic classifier that can be trained by using the entire dataset. The incremental nature of the algorithm is desirable in several settings, such as in truly online settings, where more data becomes available over time, or when the data set size is huge and batch processing is costly both in terms of computation as well as memory requirements. It is reassuring to know that the incrementally updated classifier will track the best classifier, irrespective of the dimensionality of the data or the number of data points.

Our results open up two important new directions of research worth exploring. One is the application of logistic models in unsupervised and semi-supervised settings. In unsupervised settings, one often uses a mixture of exponential family distributions [5] to model data densities, which have been appropriately extended to the semi-supervised setting as well [7]. Our results indicate that instead of using mixture of exponential family models, one can use a mixture of logistic models, whose parameters can be learned using an application of the EM algorithm [15, 5, 33]. It will be interesting to study how such models perform in realistic tasks. The second research direction is to obtain practical algorithms for Bayesian logistic regression. In practice, one often uses the MAP estimate instead of the full Bayesian posterior, resulting in a quantifiable loss in accuracy [25]. Directly computing the posterior is difficult since the Gaussian (or any other distribution) is not the conjugate prior to the logistic model [21]. It will be interesting to study if a practical approximation of the Bayesian posterior is possible while still maintaining the guarantees we have established in this paper.

References

[1] L. V. Ahlfors. *Complex Analysis.* McGraw-Hill, 1966.

[2] N. I. Akhizer. *The Classical Moment Problem and some related questions in analysis.* Hafner Publishing Company, 1965.

[3] J. A. Anderson. *Handbook of Statistics*, volume 2, chapter Logistic Discrimination, pages 169–191. P. R. Krishnaiah and L. Kanal (Eds), North-Holland, 1982.

[4] A. Banerjee. On Bayesian bounds. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[5] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[6] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* John Wiley and Sons, 1978.

[7] S. Basu, M. Bilenko, A. Banerjee, and R. Mooney. *Semi-Supervised Learning*, chapter Probabilistic Semi-supervised Clustering with Constraints. MIT Press, 2006. In preparation.

[8] C. Berg, J. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.* Springer-Verlag, 1984.

[9] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[11] S. Bochner. Monotone funktionen, stieltjes integrale und harmonische analyse. *Math. Ann.*, 108:378–410, 1933.

[12] L. M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[13] Y. Censor and S. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications.* Oxford University Press, 1998.

[14] W. Cheney and W. Light. *A Course in Approximation Theory.* Brooks/Cole Publishng Company, 2000.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[16] A. Devinatz. The representation of functions as Laplace-Stieltjes integrals. *Duke Mathematical Journal*, 24:481–498, 1955.

[17] T. G. Dietterich and E. B. Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University, 1995.

[18] W. F. Donoghue. *Distributions and Fourier Transforms*, volume 32 of *Pure and Applied Mathematics.* Academic Press, New York, 1969.

[19] W. Ehm, M. G. Genton, and T. Gneiting. Stationary covariances associated with exponentially convex functions. *Bernoulli*, 9(4):607–615, 2003.

[20] Y. Freund, R. Schapire, Y. Singer, and M. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing*, pages 334–343, 1997.

[21] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, 2003.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Ele-*

ments of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001.

[23] D. Helmbold, E. Scahpire, Y. Singer, and M. Warmuth. Online portfolio setection using multiplicative weights. *Mathematical Finance*, 8(4):325–347, 1998.

[24] M. I. Jordan. Why the logistic function? A tutorial discussion on probabilities and neural networks . Computational Cognitive Science Report 9503, MIT, 1995.

[25] S. M. Kakade and A. Ng. Online bounds for Bayesian algorithms. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2004.

[26] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. Sparse multinomial logistic regression: fast algorithms, and generalization bounds. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

[27] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[28] J. Langford and J. Shawe-Taylor. PAC-Bayes and margins. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, 2002.

[29] D. Madigan, A. Genkin, D. D. Lewis, and D. Fradkin. Bayesian multinomial logistic regression for author identification. In *Maxent Conference*, 2005.

[30] D. McAllester. PAC-Bayesian model averaging. *Machine Learning Journal*, 5:5–21, 2003.

[31] D. McAllester. Simplified PAC-Bayesian margin bounds. In *Proceedings of the 16th Annual Conference on Learning Theory*, pages 203–215, 2003.

[32] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004.

[33] T. Minka. A comparison of numerical optimizers for logistic regression, 2003.

[34] A. Ng and M. Jordan. On discrminative vs generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the 15th Annual Conference on Neural Information Processing Systems*, 2001.

[35] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, 2002.

[36] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[37] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2):257–286, 1989.

[38] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1970.

[39] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[40] H. M. Wallach. Conditional random fields: An introduction. Technical Report MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania, 2004.

[41] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with application to novelty detection. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2004.