Appearing in Proceedings of the 17th Annual Conference on Computational Learning Theory, 2004.

Performance Guarantees for Regularized Maximum Entropy Density Estimation

Miroslav Dudík[1], Steven J. Phillips[2], and Robert E. Schapire[1]

[1] Princeton University, Department of Computer Science,
35 Olden Street, Princeton, NJ 08544 USA,
{mdudik,schapire}@cs.princeton.edu,
[2] AT&T Labs − Research, 180 Park Avenue, Florham Park, NJ 07932 USA,
phillips@research.att.com

Abstract. We consider the problem of estimating an unknown probability distribution from samples using the principle of maximum entropy (maxent). To alleviate overfitting with a very large number of features, we propose applying the maxent principle with relaxed constraints on the expectations of the features. By convex duality, this turns out to be equivalent to finding the Gibbs distribution minimizing a regularized version of the empirical log loss. We prove non-asymptotic bounds showing that, with respect to the true underlying distribution, this relaxed version of maxent produces density estimates that are almost as good as the best possible. These bounds are in terms of the deviation of the feature empirical averages relative to their true expectations, a number that can be bounded using standard uniform-convergence techniques. In particular, this leads to bounds that drop quickly with the number of samples, and that depend very moderately on the number or complexity of the features. We also derive and prove convergence for both sequential-update and parallel-update algorithms. Finally, we briefly describe experiments on data relevant to the modeling of species geographical distributions.

1 Introduction

The maximum entropy (maxent) approach to probability density estimation was first proposed by Jaynes [9] in 1957, and has since been used in many areas of computer science and statistical learning, especially natural language processing [1, 6]. In maxent, one is given a set of samples from a target distribution over some space, and a set of known constraints on the distribution. The distribution is then estimated by a distribution of maximum entropy satisfying the given constraints. The constraints are often represented using a set of *features* (real-valued functions) on the space, with the expectation of every feature being required to match its empirical average. By convex duality, this turns out to be the unique Gibbs distribution maximizing the likelihood of the samples, where a Gibbs distribution is one that is exponential in a linear combination of the features. (Maxent and its dual are described more rigorously in Section 2.)

The work in this paper was motivated by a new application of maxent to the problem of modeling the distribution of a plant or animal species, a critical problem in conservation biology. This application is explored in detail in a companion paper [13]. Input

data for species distribution modeling consists of occurrence locations of a particular species in a certain region and of environmental variables for that region. Environmental variables may include topological layers, such as elevation and aspect, meteorological layers, such as annual precipitation and average temperature, as well as categorical layers, such as vegetation and soil types. Occurrence locations are commonly derived from specimen collections in natural history museums and herbaria. In the context of maxent, the sample space is a map divided into a finite number of cells, the modeled distribution is the probability that a random specimen of the species occurs in a given cell, samples are occurrence records, and features are environmental variables or functions thereof.

It should not be surprising that maxent can severely overfit training data when the constraints on the output distribution are based on feature expectations, as described above, especially if there is a very large number of features. For instance, in our application, we sometimes consider threshold features for each environmental variable. These are binary features equal to one if an environmental variable is larger than a fixed threshold and zero otherwise. Thus, there is a continuum of features for each variable, and together they force the output distribution to be non-zero only at values achieved by the samples. The problem is that in general, the empirical averages of the features will almost never be equal to their true expectation, so that the target distribution itself does not satisfy the constraints imposed on the output distribution. On the other hand, we do expect that empirical averages will be *close* to their expectations. In addition, we often have bounds or estimates on deviations of empirical feature averages from their expectations (empirical error bounds). In this paper, we propose a relaxation of feature-based maxent constraints in which we seek the distribution of maximum entropy subject to the constraint that feature expectations be *within empirical error bounds* of their empirical averages (rather than exactly equal to them).

As was the case for the standard feature-based maxent, the convex dual of this relaxed problem has a natural interpretation. In particular, this problem turns out to be equivalent to minimizing the empirical log loss of the sample points plus an ℓ_1-style regularization term. As we demonstrate, this form of regularization has numerous advantages, enabling the proof of meaningful bounds on the deviation between the density estimate and the true underlying distribution, as well as the derivation of simple algorithms for provably minimizing this regularized loss. Beginning with the former, we prove that the regularized (empirical) loss function itself gives an upper bound on the log loss with respect to the target distribution. This provides another sensible motivation for minimizing this function. More specifically, we prove a guarantee on the log loss over the target distribution in terms of empirical error bounds on features. Thus, to get exact bounds, it suffices to bound the empirical errors. For finite sets of features, we can use Chernoff bounds with a simple union bound; for infinite sets, we can choose from an array of uniform-convergence techniques. For instance, for a set of binary features with VC-dimension d, if given m samples, the log loss of the relaxed maxent solution on the target distribution will be worse by no more than $O(\|\boldsymbol{\lambda}^*\|_1 \sqrt{d \ln(m^2/d)/m})$ compared to the log loss of *any* Gibbs distribution defined by weight vector $\boldsymbol{\lambda}^*$ with ℓ_1-norm $\|\boldsymbol{\lambda}^*\|_1$. For a finite set of bounded, but not necessarily binary features, this difference is at most $O(\|\boldsymbol{\lambda}^*\|_1 \sqrt{(\ln n)/m})$ where n is the number of features. Thus, for a moderate number of samples, our method generates a density estimate that is almost

as good as the best possible, and the difference can be bounded non-asymptotically. Moreover, these bounds are very moderate in terms of the number or complexity of the features, even admitting an extremely large number of features from a class of bounded VC-dimension.

Previous work on maxent regularization justified modified loss functions as either constraint relaxations [2, 10], or priors over Gibbs distributions [2, 8]. Our regularized loss also admits these two interpretations. As a relaxed maxent, it has been studied by Kazama and Tsujii [10] and as a Laplace prior by Goodman [8]. These two works give experimental evidence showing benefits of ℓ_1-style regularization (Laplace prior) over ℓ_2^2-style regularization (Gaussian prior), but they do not provide any theoretical guarantees. In the context of neural nets, Laplace priors have been studied by Williams [20]. A smoothened version of ℓ_1-style regularization has been used by Dekel, Shalev-Shwartz and Singer [5].

Standard maxent algorithms such as iterative scaling [4, 6], gradient descent, Newton and quasi-Newton methods [11, 16] and their regularized versions [2, 8, 10, 20] perform a sequence of feature weight updates until convergence. In each step, they update all feature weights. This is impractical when the number of features is very large. Instead, we propose a sequential update algorithm that updates only one feature weight in each iteration, along the lines of algorithms studied by Collins, Schapire and Singer [3]. This leads to a boosting-like approach permitting the selection of the best feature from a very large class. For instance, the best threshold feature associated with a single variable can be found in a single linear pass through the (pre-sorted) data, even though conceptually we are selecting from an infinite class of features. In Section 4, we describe our sequential-update algorithm and give a proof of convergence. Other boosting-like approaches to density estimation have been proposed by Welling, Zemel and Hinton [19] and Rosset and Segal [15].

For cases when the number of features is relatively small, yet we want to prevent overfitting on small sample sets, it might be more efficient to minimize the regularized log loss by parallel updates. In Section 5, we give the parallel-update version of our algorithm with a proof of convergence.

In the last section, we return to our application to species distribution modeling. We present learning curves for relaxed maxent for four species of birds with a varying number of occurrence records. We also explore the effects of regularization on the log loss over the test data. A more comprehensive set of experiments is evaluated in the companion paper [13].

2 Maximum Entropy with Relaxed Constraints

Our goal is to estimate an unknown probability distribution π over a *sample space* X which, for the purposes of this paper, we assume to be finite. We are given a set of *samples* $x_1, \ldots, x_m$ drawn independently at random according to π. The corresponding empirical distribution is denoted by $\tilde{\pi}$:

$$\tilde{\pi}(x) = \tfrac{1}{m}\big|\{1 \le i \le m : x_i = x\}\big|.$$

We also are given a set of *features* $f_1, \ldots, f_n$ where $f_j : X \to \mathbb{R}$. The vector of all n features is denoted by $\boldsymbol{f}$. For a distribution π and function f, we write $\pi[f]$ to denote the

expected value of f under distribution π (and sometimes use this notation even when π is not necessarily a probability distribution):

$$\pi[f] = \sum_{x \in X} \pi(x) f(x).$$

In general, $\tilde{\pi}$ may be quite distant, under any reasonable measure, from π. On the other hand, for a given function f, we do expect $\tilde{\pi}[f]$, the empirical average of f, to be rather close to its true expectation $\pi[f]$. It is quite natural, therefore, to seek an approximation p under which f_j's expectation is equal to $\tilde{\pi}[f_j]$ for every f_j. There will typically be many distributions satisfying these constraints. The *maximum entropy principle* suggests that, from among all distributions satisfying these constraints, we choose the one of maximum entropy, i.e., the one that is closest to uniform. Here, as usual, the entropy of a distribution p on X is defined to be $\mathrm{H}(p) = -\sum_{x \in X} p(x) \ln p(x)$.

Alternatively, we can consider all *Gibbs distributions* of the form

$$q_{\boldsymbol{\lambda}}(x) = \frac{e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x)}}{Z_{\boldsymbol{\lambda}}}$$

where $Z_{\boldsymbol{\lambda}} = \sum_{x \in X} e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x)}$ is a normalizing constant, and $\boldsymbol{\lambda} \in \mathbb{R}^n$. Then it can be proved [6] that the maxent distribution described above is the same as the maximum likelihood Gibbs distribution, i.e., the distribution $q_{\boldsymbol{\lambda}}$ that maximizes $\prod_{i=1}^{m} q_{\boldsymbol{\lambda}}(x_i)$, or equivalently, minimizes the empirical log loss (negative normalized log likelihood)

$$\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) = -\frac{1}{m} \sum_{i=1}^{m} \ln q_{\boldsymbol{\lambda}}(x_i) = -\tilde{\pi}[\ln q_{\boldsymbol{\lambda}}] \tag{1}$$

A related measure is the relative entropy (or Kullback-Leibler divergence), defined as

$$\mathrm{RE}(\tilde{\pi} \parallel q_{\boldsymbol{\lambda}}) = \tilde{\pi}[\ln(\tilde{\pi}/q_{\boldsymbol{\lambda}})].$$

The log loss and the relative entropy differ only by the constant $\mathrm{H}(\tilde{\pi})$. We will use the two interchangeably as objective functions.

Thus, the convex programs corresponding to the two optimization problems are

$$\mathcal{P}: \quad \max_{p \in \Delta} \mathrm{H}(p) \text{ subject to} \quad p[f_j] = \tilde{\pi}[f_j] \qquad \mathcal{Q}: \quad \min_{\boldsymbol{\lambda} \in \mathbb{R}^n} \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})$$

where Δ is the simplex of probability distributions over X.

This basic approach computes the maximum entropy distribution p for which $p[f_j] = \tilde{\pi}[f_j]$. However, we do not expect $\tilde{\pi}[f_j]$ to be *equal* to $\pi[f_j]$ but only close to it. Therefore, in keeping with the motivation above, we can soften these constraints to have the form

$$|p[f_j] - \tilde{\pi}[f_j]| \leq \beta_j \tag{2}$$

where β_j is an estimated upper bound of how close $\tilde{\pi}[f_j]$, being an empirical average, must be to its true expectation $\pi[f_j]$. Thus, the problem can be stated as follows:

$$\max_{p \in \Delta} \mathrm{H}(p) \text{ subject to}$$
$$\forall j: \ \left| p[f_j] - \tilde{\pi}[f_j] \right| \leq \beta_j$$

This corresponds to the convex program:

$$\begin{aligned}\mathcal{P}': \quad & \max_{p\in(\mathbb{R}^+)^X} \mathrm{H}(p) \text{ subject to} \\ & \textstyle\sum_{x\in X} p(x) = 1 && (\lambda_0) \\ & \forall j:\ \tilde{\pi}[f_j] - p[f_j] \le \beta_j && (\lambda_j^+) \\ & \forall j:\ p[f_j] - \tilde{\pi}[f_j] \le \beta_j && (\lambda_j^-)\end{aligned}$$

To compute the convex dual, we form the Lagrangian (dual variables are indicated next to constraints) to obtain the dual program

$$\min_{\substack{\lambda_0\in\mathbb{R}\\ \lambda_j^-,\lambda_j^+\in\mathbb{R}^+}} \max_{p\in(\mathbb{R}^+)^X} \Big[\mathrm{H}(p) - \lambda_0\left(\left[\textstyle\sum_{x\in X} p(x)\right] - 1\right) + \sum_j (\lambda_j^+ - \lambda_j^-)\left(p[f_j] - \tilde{\pi}[f_j]\right) + \sum_j \beta_j(\lambda_j^+ + \lambda_j^-)\Big].$$

Note that we have retained use of the notation $p[f]$ and $\mathrm{H}(p)$, with the natural definitions, even though p is no longer necessarily a probability distribution. Without loss of generality we may assume that in the solution, at most one in each pair λ_j^+, λ_j^- is nonzero. Otherwise, we could decrease them both by a positive value, decreasing the value of the third sum while not affecting the remainder of the expression. Thus, if we set $\lambda_j = \lambda_j^+ - \lambda_j^-$ then we obtain a simpler program

$$\min_{\lambda_0,\lambda_j\in\mathbb{R}} \max_{p\in(\mathbb{R}^+)^X} \Big[\mathrm{H}(p) - \lambda_0\left(\left[\textstyle\sum_{x\in X} p(x)\right] - 1\right) + \sum_j \lambda_j\left(p[f_j] - \tilde{\pi}[f_j]\right) + \sum_j \beta_j|\lambda_j|\Big].$$

The inner expression is differentiable and concave in $p(x)$. Setting partial derivatives with respect to $p(x)$ equal to zero yields that p must be a Gibbs distribution with parameters corresponding to dual variables λ_j and $\ln Z_{\boldsymbol{\lambda}} = \lambda_0 + 1$. Hence the program becomes

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^n} \Big[\mathrm{H}(q_{\boldsymbol{\lambda}}) + \boldsymbol{\lambda}\cdot\left(q_{\boldsymbol{\lambda}}[\boldsymbol{f}] - \tilde{\pi}[\boldsymbol{f}]\right) + \sum_j \beta_j|\lambda_j|\Big]. \tag{3}$$

Note that

$$\mathrm{H}(q_{\boldsymbol{\lambda}}) = -q_{\boldsymbol{\lambda}}[\ln q_{\boldsymbol{\lambda}}] = -q_{\boldsymbol{\lambda}}[\boldsymbol{\lambda}\cdot\boldsymbol{f} - \ln Z_{\boldsymbol{\lambda}}] = -\boldsymbol{\lambda}\cdot q_{\boldsymbol{\lambda}}[\boldsymbol{f}] + \ln Z_{\boldsymbol{\lambda}}.$$

Hence, the inner expression of Eq. (3) becomes

$$-\boldsymbol{\lambda}\cdot\tilde{\pi}[\boldsymbol{f}] + \ln Z_{\boldsymbol{\lambda}} + \sum_j \beta_j|\lambda_j| = \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j|\lambda_j|. \tag{4}$$

(See Eq. (5) below.) Denoting this function by $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda})$, we obtain the final version of the dual program

$$\mathcal{Q}': \quad \min_{\boldsymbol{\lambda}} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}).$$

Thus, we have shown that maxent with relaxed constraints is equivalent to minimizing $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. This modified objective function consists of an empirical loss term $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})$ plus an additional term $\sum_j \beta_j|\lambda_j|$ that can be interpreted as a form of regularization limiting how large the weights λ_j can become.

3 Bounding the Loss on the Target Distribution

In this section, we derive bounds on the performance of relaxed maxent relative to the true distribution π. That is, we are able to bound $\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})$ in terms of $\mathrm{L}_\pi(\boldsymbol{\lambda}^*)$ when $\hat{\boldsymbol{\lambda}}$ minimizes the regularized loss and $q_{\boldsymbol{\lambda}^*}$ is an arbitrary Gibbs distribution, in particular, the Gibbs distribution minimizing the true loss. Note that $\mathrm{RE}(\pi \parallel q_{\boldsymbol{\lambda}})$ differs from $\mathrm{L}_\pi(\boldsymbol{\lambda})$ only by the constant term $\mathrm{H}(\pi)$, so analogous bounds also hold for $\mathrm{RE}(\pi \parallel q_{\hat{\boldsymbol{\lambda}}})$.

We begin with the following simple lemma on which all of the bounds in this section are based. The lemma states that the difference between the true and empirical loss of any Gibbs distribution can be bounded in terms of the magnitude of the weights λ_j and the deviation of feature averages from their means.

Lemma 1. *Let $q_{\boldsymbol{\lambda}}$ be a Gibbs distribution. Then*

$$\left|\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) - \mathrm{L}_\pi(\boldsymbol{\lambda})\right| \leq \sum_{j=1}^{n} \left|\lambda_j\right| \left|\tilde{\pi}[f_j] - \pi[f_j]\right|$$

Proof. Note that

$$\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) = -\tilde{\pi}[\ln q_{\boldsymbol{\lambda}}] = -\tilde{\pi}[\boldsymbol{\lambda} \cdot \boldsymbol{f} - \ln Z_{\boldsymbol{\lambda}}] = -\boldsymbol{\lambda} \cdot \tilde{\pi}[\boldsymbol{f}] + \ln Z_{\boldsymbol{\lambda}}. \tag{5}$$

Using an analogous identity for $\mathrm{L}_\pi(\boldsymbol{\lambda})$, we obtain

$$\begin{aligned}\left|\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) - \mathrm{L}_\pi(\boldsymbol{\lambda})\right| &= \left|-\boldsymbol{\lambda} \cdot \tilde{\pi}[\boldsymbol{f}] + \ln Z_{\boldsymbol{\lambda}} + \boldsymbol{\lambda} \cdot \pi[\boldsymbol{f}] - \ln Z_{\boldsymbol{\lambda}}\right| \\ &= \left|\boldsymbol{\lambda} \cdot (\tilde{\pi}[\boldsymbol{f}] - \pi[\boldsymbol{f}])\right| \leq \sum_{j=1}^{n} \left|\lambda_j\right| \left|\tilde{\pi}[f_j] - \pi[f_j]\right| \qquad \square\end{aligned}$$

This lemma yields an alternative motivation for minimizing $\mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}$. For if we have bounds $\left|\tilde{\pi}[f_j] - \pi[f_j]\right| \leq \beta_j$, then the lemma implies that $\mathrm{L}_\pi(\boldsymbol{\lambda}) \leq \mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\boldsymbol{\lambda})$. Thus, in minimizing $\mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\boldsymbol{\lambda})$, we also minimize an upper bound on $\mathrm{L}_\pi(\boldsymbol{\lambda})$, the true log loss of $\boldsymbol{\lambda}$.

Next, we prove that the distribution produced using maxent cannot be much worse than the best Gibbs distribution (with bounded weight vector), assuming the empirical errors of the features are not too large.

Theorem 1. *Assume that for each j, $\left|\pi[f_j] - \tilde{\pi}[f_j]\right| \leq \beta_j$. Let $\hat{\boldsymbol{\lambda}}$ minimize the regularized log loss $\mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\boldsymbol{\lambda})$. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^*}$*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + 2\sum_{j=1}^{n} \beta_j \left|\lambda_j^*\right|.$$

Proof.

$$\begin{aligned}\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\leq \mathrm{L}_{\tilde{\pi}}(\hat{\boldsymbol{\lambda}}) + \textstyle\sum_j \beta_j \left|\hat{\lambda}_j\right| = \mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\hat{\boldsymbol{\lambda}}) && (6)\\ &\leq \mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\boldsymbol{\lambda}^*) = \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}^*) + \textstyle\sum_j \beta_j \left|\lambda_j^*\right| && (7)\\ &\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + 2\textstyle\sum_j \beta_j \left|\lambda_j^*\right|. && (8)\end{aligned}$$

Eqs. (6) and (8) follow from Lemma 1, Eq. (7) follows from the optimality of $\hat{\boldsymbol{\lambda}}$. $\square$

Thus, if we can bound $|\pi[f_j] - \tilde{\pi}[f_j]|$, then we can use Theorem 1 to obtain a bound on the true loss $\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})$. Fortunately, this is just a matter of bounding the difference between an empirical average and its expectation, a problem for which there exists a huge array of techniques. For instance, when the features are bounded, we can prove the following:

Corollary 1. *Assume that features $f_1, \ldots, f_n$ are bounded in $[0, 1]$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\boldsymbol{\lambda})$ with $\beta_j = \beta = \sqrt{\ln(2n/\delta)/(2m)}$ for all j. Then with probability at least $1 - \delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^*}$,*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + 2\left\|\boldsymbol{\lambda}^*\right\|_1 \sqrt{\frac{\ln(2n/\delta)}{2m}}.$$

Proof. By Hoeffding's inequality, for a fixed j, the probability that $|\pi[f_j] - \tilde{\pi}[f_j]|$ exceeds β is at most $e^{-2\beta^2 m} = \delta/n$. By the union bound, the probability of this happening for any j is at most δ. The corollary now follows immediately from Theorem 1. □

Similarly, when the f_j's are selected from a possibly larger class of binary features with VC-dimension d, we can prove the following corollary. This will be the case, for instance, when using threshold features on k variables, a class with VC-dimension $O(\ln k)$.

Corollary 2. *Assume that features are binary with VC-dimension d. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}(\boldsymbol{\lambda})$ with $\beta_j = \beta = \sqrt{[d\ln(em^2/d) + \ln(1/\delta) + \ln(4e^8)]/(2m)}$ for all j. Then with probability at least $1 - \delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^*}$,*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + 2\left\|\boldsymbol{\lambda}^*\right\|_1 \sqrt{\frac{d\ln(em^2/d) + \ln(1/\delta) + \ln(4e^8)}{2m}}.$$

Proof. In this case, a uniform-convergence result of Devroye [7], combined with Sauer's Lemma, can be used to argue that $|\pi[f_j] - \tilde{\pi}[f_j]| \leq \beta$ for all f_j simultaneously, with probability at least $1 - \delta$. □

As noted in the introduction, these corollaries show that the difference in performance between the density estimate computed by minimizing $\mathrm{L}^{\boldsymbol{\beta}}_{\tilde{\pi}}$ and the best Gibbs distribution (of bounded norm), becomes small rapidly as the number of samples m increases. Moreover, the dependence of this difference on the number or complexity of the features is quite moderate.

4 A Sequential-update Algorithm and Convergence Proof

There are a number of algorithms for finding the maxent distribution, especially iterative scaling and its variants [4, 6]. In this section, we describe and prove the convergence of a sequential-update algorithm that modifies one weight λ_j at a time, as explored by Collins, Schapire and Singer [3] in a similar setting. This style of coordinate-wise descent is convenient when working with a very large (or infinite) number of features.

Input: Finite domain X
features $f_1, \ldots, f_n$ where $f_j : X \to [0,1]$
examples $x_1, \ldots, x_m \in X$
nonnegative regularization parameters $\beta_1, \ldots, \beta_n$
Output: $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ minimizing $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda})$
Let $\boldsymbol{\lambda}_1 = \mathbf{0}$
For $t = 1, 2, \ldots$:
- let $(j, \delta) = \arg\min_{(j,\delta)} F_j(\boldsymbol{\lambda}_t, \delta)$
  where $F_j(\boldsymbol{\lambda}, \delta)$ is the expression appearing in Eq. (12)
- $\lambda_{t+1,j'} = \begin{cases} \lambda_{t,j} + \delta & \text{if } j' = j \\ \lambda_{t,j'} & \text{else} \end{cases}$

Fig. 1. A sequential-update algorithm for optimizing the regularized log loss.

As explained in Section 2, the goal of the algorithm is to find $\boldsymbol{\lambda}$ minimizing the objective function $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ given in Eq. (4). Our algorithm works by iteratively adjusting the single weight λ_j that will maximize (an approximation of) the change in $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}$. To be more precise, suppose we add δ to λ_j. Let $\boldsymbol{\lambda}'$ be the resulting vector of weights, identical to $\boldsymbol{\lambda}$ except that $\lambda'_j = \lambda_j + \delta$. Then the change in $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}$ is

$$\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}') - \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \boldsymbol{\lambda} \cdot \tilde{\pi}[\boldsymbol{f}] - \boldsymbol{\lambda}' \cdot \tilde{\pi}[\boldsymbol{f}] + \ln Z_{\boldsymbol{\lambda}'} - \ln Z_{\boldsymbol{\lambda}} + \beta_j(|\lambda'_j| - |\lambda_j|) \quad (9)$$

$$= -\delta\tilde{\pi}[f_j] + \ln(q_{\boldsymbol{\lambda}}\left[e^{\delta f_j}\right]) + \beta_j(|\lambda_j + \delta| - |\lambda_j|) \quad (10)$$

$$\leq -\delta\tilde{\pi}[f_j] + \ln(q_{\boldsymbol{\lambda}}\left[1 + (e^{\delta} - 1)f_j\right]) + \beta_j(|\lambda_j + \delta| - |\lambda_j|) \quad (11)$$

$$= -\delta\tilde{\pi}[f_j] + \ln(1 + (e^{\delta} - 1)q_{\boldsymbol{\lambda}}[f_j]) + \beta_j(|\lambda_j + \delta| - |\lambda_j|). \quad (12)$$

Eq. (9) follows from Eq. (5). Eq. (10) uses

$$Z_{\boldsymbol{\lambda}'} = \sum_{x \in X} e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x) + \delta f_j(x)} = Z_{\boldsymbol{\lambda}} \sum_{x \in X} q_{\boldsymbol{\lambda}}(x) e^{\delta f_j(x)}. \quad (13)$$

Eq. (11) is because $e^{\delta x} \leq 1 + (e^{\delta} - 1)x$ for $x \in [0,1]$.

Let $F_j(\boldsymbol{\lambda}, \delta)$ denote the expression in Eq. (12). This function can be minimized over all choices of $\delta \in \mathbb{R}$ via a simple case analysis on the sign of $\lambda_j + \delta$. In particular, using calculus, we see that we only need consider the possibility that $\delta = -\lambda_j$ or that δ is equal to

$$\ln\left(\frac{(\tilde{\pi}[f_j] - \beta_j)(1 - q_{\boldsymbol{\lambda}}[f_j])}{(1 - \tilde{\pi}[f_j] + \beta_j)q_{\boldsymbol{\lambda}}[f_j]}\right) \quad \text{or} \quad \ln\left(\frac{(\tilde{\pi}[f_j] + \beta_j)(1 - q_{\boldsymbol{\lambda}}[f_j])}{(1 - \tilde{\pi}[f_j] - \beta_j)q_{\boldsymbol{\lambda}}[f_j]}\right)$$

where the first and second of these can be valid only if $\lambda_j + \delta \geq 0$ and $\lambda_j + \delta \leq 0$, respectively.

This case analysis is repeated for all features f_j. The pair (j, δ) minimizing $F_j(\boldsymbol{\lambda}, \delta)$ is then selected and δ is added to λ_j. The complete algorithm is shown in Figure 1.

The following theorem shows that this algorithm is guaranteed to produce a sequence of $\boldsymbol{\lambda}_t$'s minimizing the objective function $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}$ in the case of interest where all

the β_j's are positive. A modified proof can be used in the unregularized case in which all the β_j's are zero.

Theorem 2. *Assume all the β_j's are strictly positive. Then the algorithm of Figure 1 produces a sequence $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ for which*

$$\lim_{t\to\infty} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t) = \min_{\boldsymbol{\lambda}} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}).$$

Proof. Let us define the vectors $\boldsymbol{\lambda}^+$ and $\boldsymbol{\lambda}^-$ in terms of $\boldsymbol{\lambda}$ as follows: for each j, if $\lambda_j \geq 0$ then $\lambda_j^+ = \lambda_j$ and $\lambda_j^- = 0$, and if $\lambda_j \leq 0$ then $\lambda_j^+ = 0$ and $\lambda_j^- = -\lambda_j$. Vectors $\hat{\boldsymbol{\lambda}}^+$, $\hat{\boldsymbol{\lambda}}^-$, $\boldsymbol{\lambda}_t^+$, $\boldsymbol{\lambda}_t^-$, etc. are defined analogously.

We begin by rewriting the function F_j. For any λ, δ, we have that

$$|\lambda + \delta| - |\lambda| = \min\{\delta^+ + \delta^- \mid \delta^+ \geq -\lambda^+, \delta^- \geq -\lambda^-, \delta^+ - \delta^- = \delta\}. \qquad (14)$$

This can be seen by a simple case analysis on the signs of λ and $\lambda + \delta$. Plugging into the definition of F_j gives

$$F_j(\boldsymbol{\lambda}, \delta) = \min\{G_j(\boldsymbol{\lambda}, \delta^+, \delta^-) \mid \delta^+ \geq -\lambda^+, \delta^- \geq -\lambda^-, \delta^+ - \delta^- = \delta\}$$

where

$$G_j(\boldsymbol{\lambda}, \delta^+, \delta^-) = (\delta^- - \delta^+)\tilde{\pi}[f_j] + \ln\left(1 + (e^{(\delta^+ - \delta^-)} - 1)q_{\boldsymbol{\lambda}}[f_j]\right) + \beta_j(\delta^+ + \delta^-).$$

Combined with Eq. (12) and our choice of j and δ, this gives that

$$\begin{aligned}\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_{t+1}) - \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t) &\leq \min_j \min_\delta F_j(\boldsymbol{\lambda}_t, \delta) \\ &= \min_j \min\{G_j(\boldsymbol{\lambda}_t, \delta^+, \delta^-) \mid \delta^+ \geq -\lambda_{t,j}^+, \delta^- \geq -\lambda_{t,j}^-\} \qquad (15)\end{aligned}$$

Let $\min G(\boldsymbol{\lambda}_t)$ denote this last expression.

Since $G_j(\boldsymbol{\lambda}, 0, 0) = 0$, it follows that $\min G(\boldsymbol{\lambda}_t)$ is not positive and hence $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t)$ is nonincreasing in t. Since log loss is nonnegative, this means that

$$\sum_j \beta_j |\lambda_{t,j}| \leq \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_1) < \infty.$$

Therefore, using our assumption that the β_j's are strictly positive, we see that the $\boldsymbol{\lambda}_t$'s must belong to a compact space.

Since $\hat{\boldsymbol{\lambda}}_t$'s come from a compact space, in Eq. (15) it suffices to consider updates δ^+ and δ^- that come from a compact space themselves. Functions G_j are uniformly continuous over these compact spaces, hence the function $\min G$ is continuous.

The fact that $\hat{\boldsymbol{\lambda}}_t$'s come from a compact space also implies that they must have a subsequence converging to some vector $\hat{\boldsymbol{\lambda}}$. Clearly, $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}$ is nonnegative, and we already noted that $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t)$ is nonincreasing. Therefore, $\lim_{t\to\infty} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t)$ exists and is equal, by continuity, to $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\hat{\boldsymbol{\lambda}})$. Moreover, the differences $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_{t+1}) - \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t)$ must be converging

to zero, so $\min G(\lambda_t)$, which is nonpositive, also must be converging to zero by Eq. (15). By continuity, this means that $\min G(\hat{\boldsymbol{\lambda}}) = 0$. In particular, for each j, we have

$$\min\{G_j(\hat{\boldsymbol{\lambda}}, \delta^+, \delta^-) \mid \delta^+ \geq -\hat{\lambda}_j^+, \delta^- \geq -\hat{\lambda}_j^-\} = 0. \tag{16}$$

We will complete the proof by showing that this equation implies that $\hat{\boldsymbol{\lambda}}^+$ and $\hat{\boldsymbol{\lambda}}^-$ together with $q_{\hat{\boldsymbol{\lambda}}}$ satisfy the KKT (Kuhn-Tucker) conditions [14] for the convex program $\mathcal{P}'$, and thus form a solution to this optimization problem as well as to its dual $\mathcal{Q}'$, the minimization of $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}$. For $p = q_{\hat{\boldsymbol{\lambda}}}$, these conditions work out to be the following for all j:

$$\hat{\lambda}_j^+ \geq 0, \;\; \tilde{\pi}[f_j] - q_{\hat{\boldsymbol{\lambda}}}[f_j] \leq \beta_j, \;\; \hat{\lambda}_j^+(\tilde{\pi}[f_j] - q_{\hat{\boldsymbol{\lambda}}}[f_j] - \beta_j) = 0 \tag{17}$$

$$\hat{\lambda}_j^- \geq 0, \;\; q_{\hat{\boldsymbol{\lambda}}}[f_j] - \tilde{\pi}[f_j] \leq \beta_j, \;\; \hat{\lambda}_j^-(q_{\hat{\boldsymbol{\lambda}}}[f_j] - \tilde{\pi}[f_j] - \beta_j) = 0. \tag{18}$$

Recall that $G_j(\hat{\boldsymbol{\lambda}}, 0, 0) = 0$. Thus, by Eq. (16), if $\hat{\lambda}_j^+ > 0$ then $G_j(\hat{\boldsymbol{\lambda}}, \delta^+, 0)$ is nonnegative in a neighborhood of $\delta^+ = 0$, and so has a local minimum at this point. That is,

$$0 = \frac{\partial G_j(\hat{\boldsymbol{\lambda}}, \delta^+, 0)}{\partial \delta^+}\Big|_{\delta^+=0} = -\tilde{\pi}[f_j] + q_{\hat{\boldsymbol{\lambda}}}[f_j] + \beta_j.$$

If $\hat{\lambda}_j^+ = 0$, then Eq. (16) gives that $G_j(\hat{\boldsymbol{\lambda}}, 0, 0) \geq 0$ for $\delta^+ \geq 0$. Thus, $G_j(\boldsymbol{\lambda}, \delta^+, 0)$ cannot be decreasing at $\delta^+ = 0$. Therefore, the partial derivative evaluated above must be nonnegative. Together, these arguments exactly prove Eq. (17). Eq. (18) is proved analgously.

Thus, we have proved that

$$\lim_{t\to\infty} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t) = \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\hat{\boldsymbol{\lambda}}) = \min_{\boldsymbol{\lambda}} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}). \qquad \square$$

5 A Parallel-update Algorithm

Much of this paper has tried to be relevant to the case in which we are faced with a very large number of features. However, when the number of features is relatively small, it may be reasonable to minimize the regularized loss $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ using an algorithm that updates all features simultaneously on every iteration. There are quite a few algorithms that do this for the unregularized case, such as iterative scaling [4, 6], gradient descent, Newton and quasi-Newton methods [11, 16].

Williams [20] outlines how to modify any gradient based search to include ℓ_1-style regularization. Kazama and Tsujii [10] use a gradient based method that imposes additional linear constraints to avoid discontinuities in the first derivative. Regularized variants of iterative scaling were proposed by Goodman [8], but without a complete proof of convergence. In this section, we describe a variant of iterative scaling with a proof of convergence. Note that the gradient based or Newton methods might be faster in practice.

Throughout this section, we make the assumption (without loss of generality) that, for all $x \in X$, $f_j(x) \geq 0$ and $\sum_j f_j(x) \leq 1$. Like the algorithm of Section 4, our

parallel-update algorithm is based on an approximation of the change in the objective function $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}$, in this case the following, where $\boldsymbol{\lambda}' = \boldsymbol{\lambda} + \boldsymbol{\delta}$:

$$\begin{aligned}
\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}') - \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}) &= \boldsymbol{\lambda} \cdot \tilde{\pi}[\boldsymbol{f}] - \boldsymbol{\lambda}' \cdot \tilde{\pi}[\boldsymbol{f}] + \ln Z_{\boldsymbol{\lambda}'} - \ln Z_{\boldsymbol{\lambda}} + \sum_j \beta_j (|\lambda'_j| - |\lambda_j|) \\
&= -\boldsymbol{\delta} \cdot \tilde{\pi}[\boldsymbol{f}] + \ln q_{\boldsymbol{\lambda}}[\exp(\boldsymbol{\delta} \cdot \boldsymbol{f})] + \sum_j \beta_j (|\lambda_j + \delta_j| - |\lambda_j|) \quad (19) \\
&\leq \sum_j \Big[-\delta_j \tilde{\pi}[f_j] + q_{\boldsymbol{\lambda}}[f_j](e^{\delta_j} - 1) + \beta_j (|\lambda_j + \delta_j| - |\lambda_j|) \Big]. \quad (20)
\end{aligned}$$

Eq. (19) uses Eq. (13). For Eq. (20), note first that, if $x_j \in \mathbb{R}$ and $p_j \geq 0$ with $\sum_j p_j \leq 1$ then

$$\exp\left(\textstyle\sum_j p_j x_j\right) - 1 \leq \displaystyle\sum_j p_j (e^{x_j} - 1).$$

(See Collins, Schapire and Singer [3] for a proof.) Thus,

$$\begin{aligned}
\ln q_{\boldsymbol{\lambda}} \Big[\exp\left(\textstyle\sum_j \delta_j f_j\right) \Big] &\leq \ln q_{\boldsymbol{\lambda}} \Big[1 + \textstyle\sum_j f_j (e^{\delta_j} - 1) \Big] \\
&= \ln \left(1 + \textstyle\sum_j q_{\boldsymbol{\lambda}}[f_j](e^{\delta_j} - 1) \right) \\
&\leq \displaystyle\sum_j q_{\boldsymbol{\lambda}}[f_j](e^{\delta_j} - 1)
\end{aligned}$$

since $\ln(1 + x) \leq x$ for all $x > -1$.

Our algorithm, on each iteration, minimizes Eq. (20) over all choices of the δ_j's. With a case analysis on the sign of $\lambda_j + \delta_j$, and some calculus, we see that the minimizing δ_j must occur when $\delta_j = -\lambda_j$, or when δ_j is either

$$\ln\left(\frac{\tilde{\pi}[f_j] - \beta_j}{q_{\boldsymbol{\lambda}}[f_j]}\right) \quad \text{or} \quad \ln\left(\frac{\tilde{\pi}[f_j] + \beta_j}{q_{\boldsymbol{\lambda}}[f_j]}\right)$$

where the first and second of these can be valid only if $\lambda_j + \delta_j \geq 0$ and $\lambda_j + \delta_j \leq 0$, respectively. The full algorithm is shown in Figure 2. As before, we can prove the convergence of this algorithm when the β_j's are strictly positive.

Theorem 3. *Assume all the β_j's are strictly positive. Then the algorithm of Figure 2 produces a sequence $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ for which*

$$\lim_{t \to \infty} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t) = \min_{\boldsymbol{\lambda}} \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}).$$

Proof. The proof mostly follows the same lines as for Theorem 2. Here we sketch the main differences.

Let us redefine F_j and G_j as follows:

$$F_j(\boldsymbol{\lambda}, \delta) = -\delta \tilde{\pi}[f_j] + q_{\boldsymbol{\lambda}}[f_j](e^{\delta} - 1) + \beta_j (|\lambda_j + \delta| - |\lambda_j|)$$

Input: Finite domain X
features $f_1, \ldots, f_n$ where $f_j : X \to [0, 1]$
and $\sum_j f_j(x) \leq 1$ for all $x \in X$
examples $x_1, \ldots, x_m \in X$
nonnegative regularization parameters $\beta_1, \ldots, \beta_n$
Output: $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ minimizing $\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda})$
Let $\boldsymbol{\lambda}_1 = \mathbf{0}$
For $t = 1, 2, \ldots$:
- for each j, let $\delta_j = \arg\min_{\delta} \left(-\delta\tilde{\pi}[f_j] + q_{\boldsymbol{\lambda}}[f_j](e^{\delta} - 1) + \beta_j(|\lambda_j + \delta| - |\lambda_j|)\right)$
- update $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \boldsymbol{\delta}$

Fig. 2. A parallel-update algorithm for optimizing the regularized log loss.

and

$$G_j(\boldsymbol{\lambda}, \delta^+, \delta^-) = (\delta^- - \delta^+)\tilde{\pi}[f_j] + q_{\boldsymbol{\lambda}}[f_j](e^{\delta^+ - \delta^-} - 1) + \beta_j(\delta^+ + \delta^-).$$

Then by Eq. (14),

$$F_j(\boldsymbol{\lambda}, \delta) = \min\{G_j(\boldsymbol{\lambda}, \delta^+, \delta^-) \mid \delta^+ \geq -\lambda_j^+, \delta^- \geq -\lambda_j^-, \delta = \delta^+ - \delta^-\}.$$

So, by Eq. (20),

$$\begin{aligned}\mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_{t+1}) - \mathrm{L}_{\tilde{\pi}}^{\boldsymbol{\beta}}(\boldsymbol{\lambda}_t) &\leq \min_{\boldsymbol{\delta}} \textstyle\sum_j F_j(\boldsymbol{\lambda}_t, \delta_j) \\ &= \textstyle\sum_j \min_{\delta_j} F_j(\boldsymbol{\lambda}_t, \delta_j) \\ &= \textstyle\sum_j \min\{G_j(\boldsymbol{\lambda}_t, \delta_j^+, \delta_j^-) \mid \delta_j^+ \geq -\lambda_j^+, \delta_j^- \geq -\lambda_j^-\}.\end{aligned}$$

Note that $G_j(\boldsymbol{\lambda}, 0, 0) = 0$, so none of the terms in this sum can be positive. As in the proof of Theorem 2, the $\boldsymbol{\lambda}_t$'s have a convergent subsequence converging to some $\hat{\boldsymbol{\lambda}}$ for which

$$\textstyle\sum_j \min\{G_j(\hat{\boldsymbol{\lambda}}, \delta_j^+, \delta_j^-) \mid \delta_j^+ \geq -\lambda_j^+, \delta_j^- \geq -\lambda_j^-\} = 0.$$

This fact, in turn, implies that $\hat{\boldsymbol{\lambda}}^+$, $\hat{\boldsymbol{\lambda}}^-$ and $q_{\hat{\boldsymbol{\lambda}}}$ satisfy the KKT conditions for convex program $\mathcal{P}'$. This follows using the same arguments on the derivatives of G_j as in Theorem 2. □

6 Experiments

In order to evaluate the effect of regularization on real data, we used maxent to model the distribution of some bird species, based on occurrence records in the North American Breeding Bird Survey [17]. Experiments described in this section overlap with the (much more extensive) experiments given in the companion paper [13].

We selected four species with a varying number of occurrence records: Hutton's Vireo (198 occurrences), Blue-headed Vireo (973 occurrences), Yellow-throated Vireo

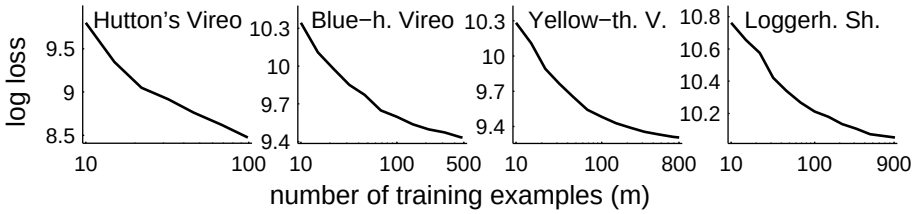

Fig. 3. *Learning curves.* Log loss averaged over 10 partitions as a function of the number of training examples. Numbers of training examples are plotted on a logarithmic scale.

(1611 occurrences) and Loggerhead Shrike (1850 occurrences). The occurrence data of each species was divided into ten random partitions: in each partition, 50% of the occurrence localities were randomly selected for the training set, while the remaining 50% were set aside for testing. The environmental variables (coverages) use a North American grid with 0.2 degree square cells. We used seven coverages: elevation, aspect, slope, annual precipitation, number of wet days, average daily temperature and temperature range. The first three derive from a digital elevation model for North America [18], and the remaing four were interpolated from weather station readings [12]. Each coverage is defined over a 386×286 grid, of which 58,065 points have data for all coverages.

In our experiments, we used threshold features derived from all environmental variables. We reduced the β_j to a single regularization parameter β as follows. We expect $|\pi[f_j] - \tilde{\pi}[f_j]| \approx \sigma[f_j]/\sqrt{m}$, where $\sigma[f_j]$ is the standard deviation of f_j under π. We therefore approximated $\sigma[f_j]$ by the sample deviation $\tilde{\sigma}[f_j]$ and used $\beta_j = \beta\tilde{\sigma}[f_j]/\sqrt{m}$. We believe that this method is more practical than the uniform convergence bounds from section 3, because it allows differentiation between features depending on empirical error estimates computed from the sample data. In order to analyze this method, we could, for instance, bound errors in standard deviation estimates using uniform convergence results.

We ran two types of experiments. First, we ran maxent on increasing subsets of the training data and evaluated log loss on the test data. We took an average over ten partitions and plotted the log loss as a function of the number of training examples. These plots are referred to as learning curves. Second, we also varied the regularization parameter β and plotted the log loss for fixed numbers of training examples as functions of β. These curves are referred to as sensitivity curves. In addition to these curves, we give examples of Gibbs distributions returned by maxent with and without regularization.

Fig. 3 shows learning curves for the four studied species. In all our runs we set $\beta = 1.0$. This choice is justified by the sensitivity curve experiments described below. In the absence of regularization, maxent would exactly fit the training data with delta functions around sample values of the environmental variables. This would result in severe overfitting even when the number of examples is large. As the learning curves show, the regularized maxent does not exhibit this behavior, and finds better and better distributions as the number of training examples increases.

In order to see how regularization facilitates learning, we examine the resulting distributions. In Fig. 4, we show Gibbs distributions returned by a regularized and an insuf-

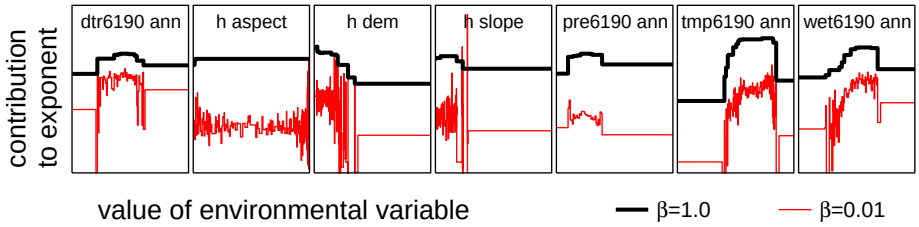

Fig. 4. *Feature profiles learned on the first partition of the Yellow-throated Vireo.* For every environmental variable, its additive contribution to the exponent of the Gibbs distribution is given as a function of its value. Profiles for the two values of β have been shifted for clarity — this corresponds to adding a constant in the exponent; it has, however, no effect on the resulting model since constants in the exponent cancel out with the normalization factor.

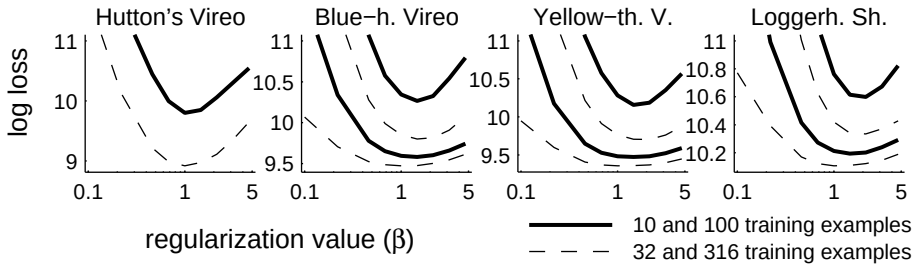

Fig. 5. *Sensitivity curves.* Log loss averaged over 10 partitions as a function of β for a varying number of training examples. For a fixed value of β, maxent finds better solutions (with smaller log loss) as the number of examples grows. We ran maxent with 10, 32, 100 and 316 training examples. Curves from top down correspond to these numbers; curves for higher numbers are missing where fewer training examples were available. Values of β are plotted on a log scale.

ficently regularized run of maxent on the first partition of the Yellow-throated Vireo. To represent Gibbs distributions, we use feature profiles. For each environmental variable, we plot the contribution to the exponent by all the derived threshold features as a function of the value of the environmental variable. This contribution is just the sum of step functions corresponding to threshold features weighted by the corresponding lambdas. As we can see, the value of $\beta = 0.01$ only prevents components of $\boldsymbol{\lambda}$ from becoming arbitrarily large, but it does little to prevent heavy overfitting with many peaks capturing single training examples. Raising β to 1.0 completely eliminates these peaks.

Fig. 5 shows the sensitivity of maxent to the regularization value β. Note that the minimum log loss is achieved consistently around $\beta = 1.0$ for all studied species. This suggests that for the purposes of maxent regularization, $\tilde{\sigma}[f_j]$ are good estimates of $\left|\tilde{\pi}[f_j] - \pi[f_j]\right|$ and that the maxent criterion models the underlying distribution well, at least for threshold features. Log loss minima for other feature types may be less consistent accross different species [13].

Acknowledgements: R. Schapire and M. Dudík received support through NSF grant CCR-0325463. M. Dudík was also partially supported by a Gordon Wu fellowship.

References

1. Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.
2. S. F. Chen and R. Rosenfeld. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50, January 2000.
3. Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1):253–285, 2002.
4. J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
5. Ofer Dekel, Shai Shalev-Shwartz, and Yoram Singer. Smooth ϵ-insensitive regression by loss symmetrization. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, pages 433–447. Springer, 2003.
6. Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):1–13, April 1997.
7. Luc Devroye. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.
8. Joshua Goodman. Exponential priors for maximum entropy models. Technical report, Microsoft Research, 2003. (Available from http://research.microsoft.com/˜joshuago/longexponentialprior.ps).
9. E. T. Jaynes. Information theory and statistical mechanics. *Physics Reviews*, 106:620–630, 1957.
10. Jun'ichi Kazama and Jun'ichi Tsujii. Evaluation and extension of maximum entropy models with inequality constraints. In *Conference on Empirical Methods in Natural Language Processing*, pages 137–144, 2003.
11. Robert Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 49–55, 2002.
12. Mark New, Mike Hulme, and Phil Jones. Representing twentieth-century space-time climate variability. Part 1: Development of a 1961-90 mean monthly terrestrial climatology. *Journal of Climate*, 12:829–856, 1999.
13. Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.
14. R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
15. Saharon Rosset and Eran Segal. Boosting density estimation. In *Advances in Neural Information Processing Systems 15*, pages 641–648. MIT Press, 2003.
16. Ruslan Salakhutdinov, Sam T. Roweis, and Zoubin Ghahramani. On the convergence of bound optimization algorithms. In *Uncertainty in Artificial Intelligence 19*, pages 509–516, 2003.
17. J. R. Sauer, J. E. Hines, and J. Fallon. The North American breeding bird survey, results and analysis 1966–2000, Version 2001.2. http://www.mbr-pwrc.usgs.gov/bbs/bbs.html, 2001. USGS Patuxent Wildlife Research Center, Laurel, MD.
18. USGS. HYDRO 1k, elevation derivative database. Available at http://edcdaac.usgs.gov/gtopo30/hydro/, 2001. United States Geological Survey, Sioux Falls, South Dakota.
19. Max Welling, Richard S. Zemel, and Geoffrey E. Hinton. Self supervised boosting. In *Advances in Neural Information Processing Systems 15*, pages 665–672. MIT Press, 2003.
20. Peter M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.