Hierarchical Maximum Entropy Density Estimation

Miroslav Dudík MDUDIK@CS.PRINCETON.EDU
David M. Blei BLEI@CS.PRINCETON.EDU
Robert E. Schapire SCHAPIRE@CS.PRINCETON.EDU
Princeton University, Department of Computer Science, 35 Olden Street, Princeton, NJ 08540

Abstract

We study the problem of simultaneously estimating several densities where the datasets are organized into overlapping groups, such as a hierarchy. For this problem, we propose a maximum entropy formulation, which systematically incorporates the groups and allows us to share the strength of prediction across similar datasets. We derive general performance guarantees, and show how some previous approaches, such as hierarchical shrinkage and hierarchical priors, can be derived as special cases. We demonstrate the proposed technique on synthetic data and in a real-world application to modeling the geographic distributions of species hierarchically grouped in a taxonomy. Specifically, we model the geographic distributions of species in the Australian wet tropics and Northeast New South Wales. In these regions, small numbers of samples per species significantly hinder effective prediction. Substantial benefits are obtained by combining information across taxonomic groups.

1. Introduction

Many real-world applications require solving multiple related learning problems. In this paper, we study the problem of simultaneously estimating several densities, whose datasets are organized into overlapping groups such as a hierarchy.

In problems of multiple estimation, we can typically either pool our data or treat each estimation problem individually. In pooling data, we obtain a confident estimate from a large sample but ignore the important differences between datasets. On the other hand, individual estimates address the separate nature of each dataset but may lead to poor generalization because of small sample sizes.

Here, we develop *hierarchical maximum entropy density estimation* (HME), a procedure that lies in the powerful middle-ground between these choices. The datasets are grouped, and the individual estimates are adjusted to reflect that grouping. With this approach, estimates from small sample sizes are influenced by the estimates for which we have more confidence; estimates from large sample sizes are less influenced by others. In statistics, this is known as *hierarchical/multi-level modeling* (Gelman & Hill, 2007) or *shrinkage*, introduced in the celebrated work of Stein (1956) and James and Stein (1961). In machine learning, hierarchical models have been used, for example, by McCallum et al. (1998) and Teh et al. (2004). These methods are also related to *multitask* or *transfer learning* (Caruana, 1993; Baxter, 2000; Raina et al., 2006)

As a running example, we consider the problem of estimating the distributions of a set of biological species in a region. We are given a set of locations, features describing them, and samples of where different species were observed. Our goal is to estimate the distribution of locations favored by each species based on the features of the kinds of places in which they are found. For example, we will consider species sampled from the *Australian wet tropics* (AWT), such as the golden bowerbird, the tooth-billed catbird, or the black treefern. All locations in the AWT are described by environmental variables such as annual mean temperature, annual precipitation, and annual mean radiation. This dataset is described in more detail in Section 6.

In recent solutions, each species distribution is modeled individually (Elith et al., 2006), even though some methods use combined data to aid variable selection (Ferrier et al., 2002; Leathwick et al., 2005). However, when modeling distributions of rare or endangered species, the number of occurrence records of a species is typically fewer than ten, and the resulting estimates are poor. With our approach, the information from several species is combined to produce better estimates for each individual species. Moreover, we can take advantage of the natural taxonomy of species. A bird's distribution is likely to be more similar to other bird distributions than it is to plant distributions. The results in Section 6 show significant improvements in predictive performance on real species data.

As a starting point, HME uses the maximum entropy approach where the equality constraints on the moments are relaxed to inequalities (Kazama & Tsujii, 2003; Dudík et al., 2004). This approach is formally equivalent to ℓ_1-regularized maximum likelihood and maximum *a posteriori* with a Laplace prior, but its alternative interpretation as a maximum entropy problem provides guidance in setting

Appearing in *Proceedings of the 24^{th} International Conference on Machine Learning*, Corvallis, OR, 2007.

hyperparameters and admits analysis of the generalization performance.

In HME, we assume that we are given a fixed class hierarchy. We fit the joint distribution of all classes, placing constraints on individual class distributions as well as on groups of classes defined by the hierarchy. We show that our approach is closely related to maximum *a posteriori* estimation with a hierarchical prior, or maximum likelihood estimation with hierarchical regularization (shrinkage). We apply the theory of maximum entropy with relaxed constraints and demonstrate how to choose hyperparameters in this setting. We prove strong generalization guarantees.

In Section 2, we introduce the objective function for HME. In Section 3, we derive an equivalent regularized maximum likelihood problem. Generalization guarantees are proved in Section 4. In Section 5, we discuss the relationship with hierarchical priors. Finally, in Section 6, we report the utility of HME on a small synthetic dataset and two large-scale real-world datasets.

2. Hierarchical Maximum Entropy

Our goal is to model multiple densities[1] over an identical sample space.[2] Density estimation problems are referred to as *classes* which are organized into *groups*; note that we are not performing classification. The set of classes will be denoted $\mathcal{Y}$, the shared sample space $\mathcal{X}$. Groups, jointly denoted $\mathcal{K}$, are formed as subsets $k \subseteq \mathcal{Y}$. Space $\mathcal{X}$ is described by real-valued functions $f : \mathcal{X} \to \mathbb{R}$, called *features*, jointly denoted as a set $\mathcal{F}$.

In our example, $\mathcal{X}$ is the set of pixels on the map of Australian wet tropics and features are equal to the environmental variables and their squares (linear and quadratic features). Classes $\mathcal{Y}$ correspond to 10 plant and 10 bird species. We introduce three groups: *plants* with 10 elements, *birds* with 10 elements, and *all species* with 20 elements. Note that we make no requirements on the composition of groups. In particular, groups can arbitrarily overlap. For example, we may have groups *rainforest plants* and *trees*, which will intersect in the set of rainforest trees.

Input consists of pairs $(X_1, Y_1), \ldots, (X_m, Y_m) \in \mathcal{X} \times \mathcal{Y}$, representing a pooled sample across all classes. In the AWT example, Y_1 may be the *golden bowerbird*, and X_1 geographic coordinates where it was observed. We assume that samples (X_i, Y_i) come from an unknown joint distribution π and use the maximum entropy principle to approximate π. Our interest, however, lies in approximating conditional distributions of locations given species, $\pi_y(x) = \pi(x \mid y)$. Therefore, the Y_i's do not need to be random. This is in contrast to logistic regression, where the goal is to approximate $\pi(y \mid x)$ for classification.

The maximum entropy principle (maxent) (Jaynes, 1957) tells us to approximate π by the distribution of maximum entropy that satisfies a set of constraints expressed in terms of features. Ignoring group information, constraints are specified for each class separately, and typically require that feature expectations match their empirical averages. In the AWT example, this means that the model of the golden bowerbird should match the average altitude and the average squared altitude in which the golden bowerbird was observed. This is equivalent to matching the sample mean and sample variance.

When the number of samples is too small or the number of features too large, maxent overfits, because the true distribution does not match empirical averages exactly. We alleviate overfitting by relaxing the constraints so that feature expectations are required to be only close to sample averages.

In HME, we use the group information to leverage information across species. In addition to requiring that feature expectations of each individual class are close to their empirical averages, we also require that feature expectations for each group are close to the group empirical averages. Thus, in AWT, we require that the expectation of the altitude across all birds is not too far from the average altitude across all samples from the group *birds*. Since the total number of samples in the group *birds* is larger than, for example, the number of samples of the golden bowerbird, we can be more confident about our estimates of the means. This amounts to sharing information across all bird species.

Let $\tilde{\pi}$ denote the empirical distribution. We express both class and group constraints in terms of conditional expectations on the joint distribution:

$$
\begin{aligned}
\mathcal{P}: \max_{p \in \Delta} \mathrm{H}(p) \\
\text{s.t. } & p(y) = \tilde{\pi}(y) \text{ for all } y \in \mathcal{Y} \\
& \left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid Y = y] - \mathbf{E}_p[f(X) \mid Y = y]\right| \le \beta_{y,f} \\
& \qquad \text{for all } y \in \mathcal{Y}, f \in \mathcal{F} \\
& \left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid Y \in k] - \mathbf{E}_p[f(X) \mid Y \in k]\right| \le \beta_{k,f} \\
& \qquad \text{for all } k \in \mathcal{K}, f \in \mathcal{F}.
\end{aligned}
$$

Here, Δ is the simplex of probability distributions over $\mathcal{X} \times \mathcal{Y}$, $\mathrm{H}(p) = -\sum_{x,y} p(x,y) \ln p(x,y)$ is the entropy, $p(y)$ denotes the probability of $Y = y$, $\mathbf{E}_p$ denotes the expectation under distribution p, and $\beta_{k,f} \ge 0$ are errors that we allow in matching individual expectations.

Note that if $\mathcal{K} = \emptyset$, HME reduces to a series of maxent problems for each class: the joint entropy is maximized when all class entropies are maximized because class probabilities are fixed. When $\mathcal{K}$ is non-empty, the set of constraints in HME is more restrictive than a series of single-class maxent problems, so the resulting solutions differ.

Similar to single-class maxent, we will see that HME is equivalent to a regularized maximum likelihood problem. Specifically, the entropy is maximized by a distribution which takes the form $p(x,y) = \tilde{\pi}(y) q_{\boldsymbol{\lambda}_y}(x)$, where $q_{\boldsymbol{\lambda}_y}$ stands for a *Gibbs distribution* specified by a vector $\boldsymbol{\lambda}_y \in \mathbb{R}^{\mathcal{F}}$. For an arbitrary $\boldsymbol{\lambda} \in \mathbb{R}^{\mathcal{F}}$, the Gibbs distribution

[1] In this paper, we are concerned with densities relative to the counting measure on a discrete set. These correspond to probability mass functions.

[2] The restriction that the densities are over the same sample space simplifies the exposition, but it could be omitted.

$q_{\boldsymbol{\lambda}}$ is defined as

$$q_{\boldsymbol{\lambda}}(x) = \exp\left\{\sum_{f\in\mathcal{F}} \lambda_f f(x)\right\} / Z_{\boldsymbol{\lambda}},$$

where $Z_{\boldsymbol{\lambda}} = \sum_{x\in\mathcal{X}} \exp\{\sum_{f\in\mathcal{F}} \lambda_f f(x)\}$ is the normalization constant. We will see that the problem $\mathcal{P}$ is equivalent to the following regularized maximum likelihood problem:

$$\mathcal{Q}: \max_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{F}}\\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{K}\times\mathcal{F}}}} \left\{ \frac{1}{m}\sum_{i=1}^{m}\left(\ln q_{\boldsymbol{\lambda}_{Y_i}}(X_i)\right) - \sum_{y\in\mathcal{Y},f\in\mathcal{F}}\left(\tilde{\pi}(y)\beta_{y,f}\left|\lambda_{y,f} - \sum_{k:y\in k}\eta_{k,f}\right|\right) - \sum_{k\in\mathcal{K},f\in\mathcal{F}}\left(\tilde{\pi}(k)\beta_{k,f}\left|\eta_{k,f}\right|\right)\right\} .$$

We use $\tilde{\pi}(k)$ for the probability that $Y \in k$ under the distribution $\tilde{\pi}$. Vectors $\boldsymbol{\lambda}_y$ describe class distributions $q_{\boldsymbol{\lambda}_y}$, vectors $\boldsymbol{\eta}_k$ account for effects of membership in different groups. The dual objective is to optimize log likelihood of the data (the first term) under an ℓ_1-style penalty for deviating from group effects (the second term), which are themselves regularized by an ℓ_1-style penalty (the third term).

In Section 3, we show that solutions of $\mathcal{Q}$ correspond to solutions of $\mathcal{P}$; thus, we only need to optimize $\mathcal{Q}$ to solve the HME problem. In Section 4, we address the question how well the HME solutions perform on test data from the unknown distributions π_y. We show that the HME solutions converge to the best approximations of the π_y's by Gibbs distributions. The bounds will also indicate how the use of group information improves generalization.

3. HME Duality

In this section we show that $\mathcal{Q}$ is indeed a dual of $\mathcal{P}$, i.e., we establish the correspondence between maximum entropy and regularized maximum likelihood when estimating several distributions simultaneously.

Our setup will be slightly more general than the one discussed in Section 2. In particular, we make it possible to specify the importance of individual classes. The vector listing importance of individual classes will be denoted as $c \in \mathbb{R}^{\mathcal{Y}}$, with components referred to as $c(y)$. We assume that importance is scaled so that $\sum_{y\in\mathcal{Y}} c(y) = 1$; thus, c is a distribution over $\mathcal{Y}$.

To incorporate the importance c in the optimization, we modify our target distribution. We still assume that samples (X_i, Y_i) come from an unknown distribution π, but our goal is now to perform well relative to the distribution μ, which weights individual classes according to their importance:

$$\mu(x, y) = c(y)\pi(x \mid y) \ .$$

Equivalently, $\mu(y) = c(y)$ and $\mu(x \mid y) = \pi(x \mid y)$.

To simplify the exposition, we assume that constraint widths $\beta_{y,f}, \beta_{k,f}$ are feature independent, i.e., they depend only on the class y or group k. The duality results and performance guarantees generalize to feature dependent widths.

We will now modify the constraints of HME from Section 2 to reflect reweighting of the classes.

$$\mathcal{P}': \max_{p\in\Delta} \mathrm{H}(p)$$

$$\text{s.t. } p(y) = c(y) \text{ for all } y \in \mathcal{Y} \tag{1}$$

$$\left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid y] - \mathbf{E}_p[f(X) \mid y]\right| \le \beta_y \quad \text{for all } y \in \mathcal{Y}, f \in \mathcal{F} \tag{2}$$

$$\left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid k] - \sum_{y\in k}\tilde{\pi}(y \mid k)\mathbf{E}_p[f(X) \mid y]\right| \le \beta_k \quad \text{for all } k \in \mathcal{K}, f \in \mathcal{F}. \tag{3}$$

In conditional expectations, we abbreviated events $Y \in y$ and $Y \in k$ as y and k. Eq. (1) reflects our assumption $\mu(y) = c(y)$. Eq. (2) reflects our assumption $\mu(x \mid y) = \pi(x \mid y)$, and captures the approximation

$$\mathbf{E}_{\tilde{\pi}}[f(X) \mid y] \approx \mathbf{E}_{\pi}[f(X) \mid y] = \mathbf{E}_{\mu}[f(X) \mid y] \ .$$

In Eq. (3), we express the approximation

$$\begin{aligned}\mathbf{E}_{\pi}[f(X) \mid k] &\approx \textstyle\sum_{y\in k}\tilde{\pi}(y \mid k)\mathbf{E}_{\pi}[f(X) \mid y] \\ &= \textstyle\sum_{y\in k}\tilde{\pi}(y \mid k)\mathbf{E}_{\mu}[f(X) \mid y] \ .\end{aligned}$$

Note that if we set $c(y) = \tilde{\pi}(y)$, i.e., we set the importance of each class according to its empirical probability, then we obtain the primal $\mathcal{P}$.

Next we show that the maximum entropy solution to $\mathcal{P}'$ can be obtained from a regularized max log likelihood problem:

Theorem 1. *Let $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{\mathcal{Y}\times\mathcal{F}}$ optimize*

$$\mathcal{Q}': \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{F}}\\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{K}\times\mathcal{F}}}} \left\{ \frac{1}{m}\sum_{i=1}^{m}\left(\frac{c(Y_i)}{\tilde{\pi}(Y_i)}\ln q_{\boldsymbol{\lambda}_{Y_i}}(X_i)\right) - \sum_{y\in\mathcal{Y}} c(y)\beta_y\left\|\boldsymbol{\lambda}_y - \sum_{k:y\in k}\frac{\tilde{\pi}(y \mid k)}{c(y \mid k)}\boldsymbol{\eta}_k\right\|_1 - \sum_{k\in\mathcal{K}} c(k)\beta_k\|\boldsymbol{\eta}_k\|_1\right\} .$$

Then $p(x, y) = c(y)q_{\hat{\boldsymbol{\lambda}}_y}(x)$ solves the primal $\mathcal{P}'$, with terms $q_{\hat{\boldsymbol{\lambda}}_y}$ possibly replaced by their limit distributions.

Sketch of proof. The claim of the theorem follows from generalized maxent duality (Dudík & Schapire, 2006). To apply maxent duality, however, we need to express $\mathcal{P}'$ in terms of convex constraints of *unconditional* feature expectations. To remove conditional expectations, we observe that $p(y) = c(y)$ at the solution. Thus, conditioning on $Y = y$ can be replaced by the division by $c(y)$. To obtain a set of constraints equivalent to Eqs. (1–3), we introduce a new set of features, defined on $\mathcal{X}\times\mathcal{Y}$, and indexed by y and (y, f) respectively:

$$\begin{aligned}h_y(x', y') &= \delta(y, y') \\ g_{y,f}(x', y') &= \delta(y, y')f(x, y)/c(y) \ ,\end{aligned}$$

where $\delta(y, y')$ equals 1 if $y = y'$ and equals 0 otherwise. Maximizing entropy under the constraints (1–3) is thus equivalent to maximizing entropy under the constraints

$$\begin{aligned}
&\mathbf{E}_p[h_y(X,Y)] = c(y) \text{ for all } y \\
&\left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid y] - \mathbf{E}_p[g_{y,f}(X,Y)]\right| \leq \beta_y \text{ for all } y, f \\
&\left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid k] - \sum_{y \in k} \tilde{\pi}(y \mid k)\mathbf{E}_p[g_{y,f}(X,Y)]\right| \leq \beta_k \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } k, f.
\end{aligned}$$

The objective of $\mathcal{Q}'$ is now derived by taking the convex conjugate of the indicator function of the constraint set (see, e.g., Boyd & Vandenberghe, 2004). An alternative proof, which only proves the non-limit case, is obtained by applying the method of Lagrange multipliers. ■

4. Performance Guarantees

In Section 3, we have seen that class distributions under HME take the form of Gibbs distributions. However, these distributions need not accurately represent true class densities $\pi_y = \pi(\cdot \mid y)$. In fact, if the feature set is poorly chosen, it is possible that no Gibbs distributions are good approximations of π_y. Therefore, it only makes sense to compare performance of the HME solutions against the best performance among all Gibbs distributions.

In this section, we derive guarantees on HME performance relative to arbitrary Gibbs distributions. We will see that the performance depends directly on the regularization and this will help us choose the appropriate widths β_y and β_k.

As a measure of performance, we use *relative entropy*. For distributions p_1, p_2 over a set $\mathcal{X}$, relative entropy measures their information-theoretic distance, and is defined as

$$\mathrm{RE}(p_1 \parallel p_2) = \textstyle\sum_{x \in \mathcal{X}} p_1(x) \ln[p_1(x)/p_2(x)] \ .$$

It differs from the negative of test log likelihood

$$-\mathbf{E}_{p_1}[\ln p_2(X)]$$

only by the constant $\mathrm{H}(p_1)$. Thus, minimizing $\mathrm{RE}(p_1 \parallel p_2)$ corresponds to maximizing the log likelihood $\mathbf{E}_{p_1}[\ln p_2(X)]$.

In our guarantees, we compare $\mathrm{RE}(\pi_y \parallel q_{\hat{\boldsymbol{\lambda}}_y})$, where $q_{\hat{\boldsymbol{\lambda}}_y}$ are HME solutions, with $\mathrm{RE}(\pi_y \parallel q_{\boldsymbol{\lambda}_y^{\text{OPT}}})$, where $q_{\boldsymbol{\lambda}_y^{\text{OPT}}}$ are arbitrary Gibbs distributions; in particular, these can be Gibbs distributions which best approximate π_y. The performance across classes is weighted according to $c(y)$.

Before proving specific performance guarantees, we derive a general lemma. It relates the set of constraints in HME primal with the regularization in the dual. In particular, it states that if true feature means satisfy HME constraints then the gap in performance between the maxent solution and an arbitrary Gibbs distribution is at most twice the value of regularization of that Gibbs distribution. Thus, the guarantee reflects the notion that the regularization quantifies the complexity of the Gibbs distribution.

Lemma 2. *Let $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{\mathcal{Y} \times \mathcal{F}}$ solve the regularized log likelihood problem $\mathcal{Q}'$. Assume that feature expectations with respect to true class densities π_y satisfy*

$$\begin{aligned}
&\left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid y] - \mathbf{E}_{\pi_y}[f(X)]\right| \leq \beta_y \text{ for all } y \in \mathcal{Y}, f \in \mathcal{F} \\
&\left|\mathbf{E}_{\tilde{\pi}}[f(X) \mid k] - \textstyle\sum_{y \in k} \tilde{\pi}(y \mid k)\mathbf{E}_{\pi_y}[f(X)]\right| \leq \beta_k \\
&\qquad\qquad\qquad\qquad \text{for all } k \in \mathcal{K}, f \in \mathcal{F}.
\end{aligned}$$

Then for all $\boldsymbol{\lambda}^{\text{OPT}} \in \mathbb{R}^{\mathcal{Y} \times \mathcal{F}}, \boldsymbol{\eta}^{\text{OPT}} \in \mathbb{R}^{\mathcal{K} \times \mathcal{F}}$

$$\begin{aligned}
\sum_{y \in \mathcal{Y}} c(y)\mathrm{RE}\big(\pi_y \parallel q_{\hat{\boldsymbol{\lambda}}_y}\big) \leq \sum_{y \in \mathcal{Y}} c(y)\mathrm{RE}\big(\pi_y \parallel q_{\boldsymbol{\lambda}_y^{\text{OPT}}}\big) \\
+ 2\sum_{y \in \mathcal{Y}} c(y)\beta_y \Big\|\boldsymbol{\lambda}_y^{\text{OPT}} - \sum_{k: y \in k} \frac{\tilde{\pi}(y \mid k)}{c(y \mid k)} \boldsymbol{\eta}_k^{\text{OPT}}\Big\|_1 \\
+ 2\sum_{k \in \mathcal{K}} c(k)\beta_k \|\boldsymbol{\eta}_k^{\text{OPT}}\|_1 \ .
\end{aligned}$$

Lemma 2 can be derived from the corresponding lemma for generalized maxent by the same transformation as in the proof of Theorem 1. It is also possible to prove Lemma 2 without an explicit use of convex conjugacy, similar to the single-class case (Dudík et al., 2004). The complete proof of Lemma 2 will appear in an extended version of the paper (it is omitted here for the sake of brevity).

Lemma 2 guides the choice of β_y, β_k. In particular, β_y and β_k should be chosen as small as possible, so that true class densities satisfy HME constraints with high probability. As in the single-class case, this amounts to bounding deviations of empirical averages from their means. There are many statistical techniques available for this (see, e.g., Devroye et al., 1996).

We now derive a specific bound for the case when $\mathcal{F}$ is a finite set of bounded features. We let m_y and m_k denote the number of examples with $Y_i = y$ and $Y_i \in k$. Without loss of generality, we assume that features are scaled, so their values lie within the interval $[0, 1]$. This case covers linear and quadratic features in the AWT example.

Theorem 3. *Assume that $f : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ for all $f \in \mathcal{F}$, and $\mathcal{F}$ is finite. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ maximize regularized likelihood $\mathcal{Q}'$ with $\beta_y = \beta_0/\sqrt{m_y}, \beta_k = \beta_0/\sqrt{m_k}$ where $\beta_0 = \sqrt{\ln\left(2|\mathcal{Y}||\mathcal{F}| + 2|\mathcal{K}||\mathcal{F}|\right)/2}$. Then with probability at least $1 - \delta$, for all $\boldsymbol{\lambda}^{\text{OPT}} \in \mathbb{R}^{\mathcal{Y} \times \mathcal{F}}, \boldsymbol{\eta}^{\text{OPT}} \in \mathbb{R}^{\mathcal{K} \times \mathcal{F}}$,*

$$\begin{aligned}
\sum_{y \in \mathcal{Y}} c(y)\mathrm{RE}\big(\pi_y \parallel q_{\hat{\boldsymbol{\lambda}}_y}\big) \leq \sum_{y \in \mathcal{Y}} c(y)\mathrm{RE}\big(\pi_y \parallel q_{\boldsymbol{\lambda}_y^{\text{OPT}}}\big) \\
+ 2\beta_0 \sum_{y \in \mathcal{Y}} \frac{c(y)}{\sqrt{m_y}} \Big\|\boldsymbol{\lambda}_y^{\text{OPT}} - \sum_{k: y \in k} \frac{\tilde{\pi}(y \mid k)}{c(y \mid k)} \boldsymbol{\eta}_k^{\text{OPT}}\Big\|_1 \\
+ 2\beta_0 \sum_{k \in \mathcal{K}} \frac{c(k)}{\sqrt{m_k}} \big\|\boldsymbol{\eta}_k^{\text{OPT}}\big\|_1 \ .
\end{aligned}$$

Proof. Instead of drawing pairs (X_i, Y_i) independently from π, we first draw Y_i's independently from π and then draw each X_i from π_{Y_i}. It suffices to show that for any choice of Y_i's, the statement of the theorem is true with probability at least $1 - \delta$ over the draw of X_i's. We first consider the constraints conditioned on $Y \in k$. If the Y_i's are fixed then for an arbitrary f and k, the empirical

mean $\mathbf{E}_{\tilde{\pi}}[f(X) \mid k]$ is an average of m_k independent (but not identically distributed!) random variables bounded in $[0, 1]$. Expectation of this empirical mean, conditioned on the Y_i's, is

$$\sum_{y \in k} \tilde{\pi}(y \mid k) \mathbf{E}_{\pi}[f(X) \mid y] \ .$$

Thus, by Hoeffding's inequality, the probability that the deviation

$$\left| \mathbf{E}_{\tilde{\pi}}[f(X) \mid k] - \sum_{y \in k} \tilde{\pi}(y \mid k) \mathbf{E}_{\pi}[f(X) \mid y] \right|$$

exceeds β_k is at most $\delta / (|\mathcal{Y}||\mathcal{F}| + |\mathcal{K}||\mathcal{F}|)$. Similarly, the probability that any particular constraint conditioned on $Y = y$ is not satisfied is at most $\delta / (|\mathcal{Y}||\mathcal{F}| + |\mathcal{K}||\mathcal{F}|)$. Hence, by the union bound, the probability that this will happen for any $k \in \mathcal{K}, f \in \mathcal{F}$ or $y \in \mathcal{Y}, f \in \mathcal{F}$ is at most δ. ■

Theorem 3 provides an insight how the group information improves learning. For instance, consider a simple scenario of estimating distributions of birds, all of which have an equal number of occurrences and equal importance, i.e., $m_y = m/|\mathcal{Y}|$ and $c(y) = 1/|\mathcal{Y}|$ for all y. Further, assume that distributions of these birds are similarly influenced by about half the features, and distinctly influenced by the other half. For example, the birds are influenced in the same way by precipitation and vegetation, but different birds respond differently to temperature. Denote the first subset of features as *shared* and the second subset of features as *distinct*. We compare how our generalization guarantees change if we introduce the group *birds*.

First, fix parameters $\boldsymbol{\lambda}_y^{\text{OPT}}$ of the optimal Gibbs distributions. Since species depend similarly on *shared*, we assume that the slices of parameters $\boldsymbol{\lambda}_{y,\textit{shared}}^{\text{OPT}}$ corresponding to *shared* are roughly equal; denote the shared parameter values as $\boldsymbol{\lambda}_{\textit{shared}}^{\text{OPT}}$. For an empty hierarchy, the gap between the maxent solutions and the best Gibbs distributions is

$$2\beta_0 \sum_{y \in \mathcal{Y}} \frac{c(y)}{\sqrt{m_y}} \|\boldsymbol{\lambda}_y^{\text{OPT}}\|_1 = \frac{2\beta_0}{\sqrt{m|\mathcal{Y}|}} \sum_{y \in \mathcal{Y}} \|\boldsymbol{\lambda}_y^{\text{OPT}}\|_1$$
$$= \frac{2\beta_0}{\sqrt{m|\mathcal{Y}|}} \sum_{y \in \mathcal{Y}} \|\boldsymbol{\lambda}_{y,\textit{distinct}}^{\text{OPT}}\|_1 + 2\beta_0 \sqrt{\frac{|\mathcal{Y}|}{m}} \|\boldsymbol{\lambda}_{\textit{shared}}^{\text{OPT}}\|_1 \ . \quad (4)$$

Now, add the group *birds*, and set $\boldsymbol{\eta}_{\textit{birds,shared}}^{\text{OPT}} = \boldsymbol{\lambda}_{\textit{shared}}^{\text{OPT}}$ and $\boldsymbol{\eta}_{\textit{birds,distinct}}^{\text{OPT}} = \mathbf{0}$. The first term of Eq. (4) remains unchanged except for β_0, which slightly increases to reflect $|\mathcal{K}| = 1$. The second term, however, becomes zero, and an additional term appears accounting for the group *birds*:

$$2\beta_0 \frac{c(\textit{birds})}{\sqrt{m_{\textit{birds}}}} \|\boldsymbol{\eta}_{\textit{birds}}^{\text{OPT}}\| = \frac{2\beta_0}{\sqrt{m}} \|\boldsymbol{\lambda}_{\textit{shared}}^{\text{OPT}}\|_1 \ .$$

Thus, when the group *birds* is introduced, the second term of Eq. (4) is effectively divided by the square root of the number of species. Already for a moderate number of species, for example, 10 or 20, this may constitute a significant decrease. Assuming that the relevance of *distinct* is similar to the relevance of *shared*, i.e., $\|\boldsymbol{\lambda}_{y,\textit{distinct}}\|_1 \approx \|\boldsymbol{\lambda}_{y,\textit{shared}}\|_1$, the gap in performance between the maxent distributions and the best Gibbs distributions is reduced almost twofold.

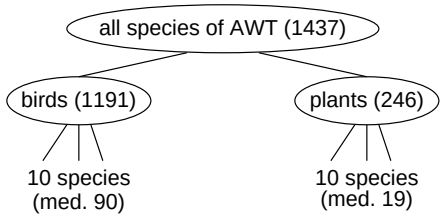

Figure 1. The hierarchy of species in the Australian wet tropics dataset. Numbers in parentheses indicate the number of training records. At the lowest level, we list only the number of species and report the median number of training records.

Note that the guarantee of Theorem 3 grows very moderately with the number of features. In particular, the bound is meaningful as long as the number of features grows subexponentially with the number of training examples.

5. HME as MAP with a hierarchical prior

So far, we have considered two interpretations of the HME problem. The first interpretation is the maximization of entropy subject to constraints on conditional expectations. The second interpretation is the maximization of regularized log likelihood. Here, we introduce a third interpretation. We show that when $\mathcal{K}$ describes a tree hierarchy, HME can be viewed as maximum *a posteriori* under a hierarchical Laplace prior. The HME interpretation is more general since it allows arbitrary groups. In addition, it guides the process of choosing hyperparameters and provides insights into generalization properties.

In this section, we limit our attention to tree hierarchies, such as the AWT hierarchy in Fig. 1. In this case it is natural to set up a hierarchical model, in which we associate a vector of Gibbs distribution parameters $\boldsymbol{\lambda}_n$ with each node n. Let $\mathcal{N}$ denote the set of all nodes in the hierarchy, including leaves y corresponding to our individual classes. A hierarchical Laplace prior, conditioned on $Y_1, \ldots, Y_m$, can then be specified as

$$\boldsymbol{\lambda}_{\textit{root}} \sim e^{-\alpha_{\textit{root}} \|\boldsymbol{\lambda}_{\textit{root}}\|_1} \quad (5)$$
$$\boldsymbol{\lambda}_n \sim e^{-\alpha_n \|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}_{\textit{parent}(n)}\|_1} \text{ for all } n \neq \textit{root} \quad (6)$$
$$X_i \mid \boldsymbol{\lambda}_{Y_i} \sim q_{\boldsymbol{\lambda}_{Y_i}} \text{ for all } i. \quad (7)$$

This corresponds to the directed graphical model with the structure identical to the hierarchy, with a different $\boldsymbol{\lambda}_n$ variable assigned to each node. The root is distributed according to Eq. (5), the remaining nodes depend on their parents according to Eq. (6), and observations, described by Eq. (7), are attached at the bottom.

For example, in AWT, the process of drawing samples $X_1, \ldots, X_m$ given $Y_1, \ldots, Y_m$ can be described as first drawing the parameter $\boldsymbol{\lambda}_{\textit{all species}}$ according to its prior, then choosing $\boldsymbol{\lambda}_{\textit{birds}}$ and $\boldsymbol{\lambda}_{\textit{plants}}$ conditioned on $\boldsymbol{\lambda}_{\textit{all species}}$, then drawing $\boldsymbol{\lambda}_y$ conditioned on the respective groups, such as $\boldsymbol{\lambda}_{\textit{golden bowerbird}}$ conditioned on $\boldsymbol{\lambda}_{\textit{birds}}$, and finally choosing observations X_i in which $Y_i = \textit{golden bowerbird}$, conditioned on $\boldsymbol{\lambda}_{\textit{golden bowerbird}}$.

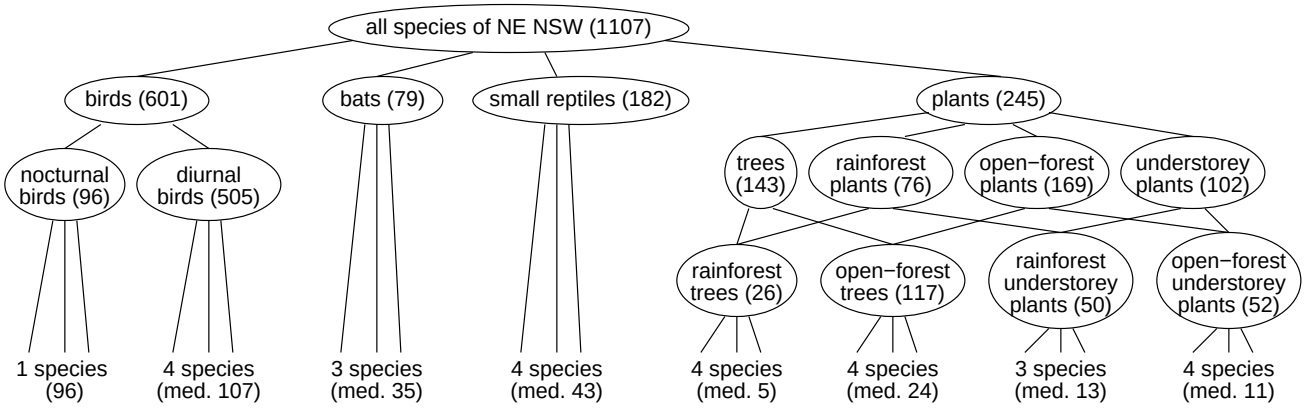

Figure 2. The hierarchy of species in the North-east New South Wales dataset. Numbers in parentheses indicate the number of training records. At the lowest level, we list only the number of species and report the median number of training records. Note that the children of *plants* correspond to overlapping groups. This hierarchy, therefore, cannot be represented as a tree.

To derive the equivalence of $\mathcal{Q}'$ with a hierarchical Laplace prior, we set class importance equal to empirical probabilities and multiply the objective of $\mathcal{Q}'$ by m:

$$\sum_{i=1}^{m} \left(\ln q_{\boldsymbol{\lambda}_{Y_i}}(X_i)\right) - \sum_{y\in\mathcal{Y}} \left(m_y \beta_y \left\| \boldsymbol{\lambda}_y - \sum_{k:y\in k} \boldsymbol{\eta}_k \right\|_1 \right) - \sum_{k\in\mathcal{K}} \left(m_k \beta_k \left\| \boldsymbol{\eta}_k \right\|_1 \right) \ . \tag{8}$$

The first term is the log likelihood. The second and third terms, corresponding to regularization, can be viewed as the log of a prior, and the entire expression as the log of a posterior. Thus, maximizing the regularized log likelihood corresponds to maximizing the posterior.

To show that the regularization in Eq. (8) corresponds to the hierarchical prior described above, we identify each inner node n with the set $k(n) \subseteq \mathcal{Y}$ containing all classes y which are descendants of n. We set $\mathcal{K} = \{k(n) : n \text{ is an inner node}\}$ and establish the correspondence by setting $\boldsymbol{\lambda}_n$, for each inner node n, equal to the sum of contributions $\boldsymbol{\eta}_{k(n')}$ over n' on the path from the root to the node n. The second and third terms in Eq. (8) then become

$$- \sum_{y\in\mathcal{Y}} \left(m_y \beta_y \|\boldsymbol{\lambda}_y - \boldsymbol{\lambda}_{\textit{parent}(y)}\|_1 \right) - m_{\textit{root}} \beta_{\textit{root}} \|\boldsymbol{\lambda}_{\textit{root}}\|_1 - \sum_{n\in\mathcal{N}\setminus\mathcal{Y}\setminus\{\textit{root}\}} \left(m_n \beta_n \|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}_{\textit{parent}(n)}\|_1 \right)$$

where m_n and β_n are shorthand for $m_{k(n)}$ and $\beta_{k(n)}$. The equivalence with the hierarchical Laplace prior is now obtained by setting $\alpha_n = m_n \beta_n$.

6. Experiments

We evaluate HME on synthetic and real-world data. In both cases, we use the sequential update algorithm of Dudík and Schapire (2006). Class importance in all of our experiments equals empirical probabilities. For regularization widths, we use a tighter setting than discussed in Section 4. Instead of using an identical β_y and β_k across all feature expectations conditioned on $Y = y$ and $Y \in k$, we use feature specific settings. Specifically, since $\beta_{k,f}$ and $\beta_{y,f}$ correspond to deviations of empirical averages from true means, we use central limit approximations and set

$$\beta_{y,f} = \beta_0 \sqrt{\mathbf{V}_{\tilde{\pi}}[f(X) \mid y]/m_y}$$
$$\beta_{k,f} = \beta_0 \sqrt{\mathbf{V}_{\tilde{\pi}}[f(X) \mid k]/m_k} \ ,$$

where β_0 is a single tuning parameter and $\mathbf{V}_{\tilde{\pi}}$ is the empirical variance.

Synthetic data. We first study a synthetic toy-example. Our map consists of 100 pixels described by two features: precipitation (*prec*) and temperature (*temp*). Values of *prec* are equally spaced in $[0, 1]$ and *temp* is defined as $\textit{temp} = (2 \cdot \textit{prec} - 1)^2$ (we make no claims about physical plausibility of this model). We study two synthetic species: *icebird* and *sunbird*. Both prefer low precipitation, but they differ in their temperature requirements: *icebird* prefers low temperatures while *sunbird* prefers high temperatures. We assume that true distributions of *icebird* and *sunbird* are Gibbs distributions with parameters $\boldsymbol{\lambda}_{\textit{icebird}} = (-5, -2)$, $\boldsymbol{\lambda}_{\textit{sunbird}} = (-3, 1)$.

We have 100 observations of *sunbird* and vary the number of observations of *icebird* between 3 and 10,000. For each number of occurrences, we estimate the distribution of *icebird* using both single-class maxent, and HME with a single group $\textit{birds} = \{\textit{icebird}, \textit{sunbird}\}$. The tuning parameter β_0 is set to 0.5.

In Figure 3, we present our results. For each HME run, we report values of the HME parameters of *icebird*, *sunbird*, and the group *birds*. For *temp*, the HME parameters of *icebird* agree with its single-class parameters. This matches the intuition behind the bound of Section 4: the temperature requirements of *icebird* and *sunbird* are different, so pooled estimates provide no advantage; the best setting of the *birds* parameter is zero and the best setting of the *icebird* parameter matches the single-class case. For *prec*, the situation is rather different. The parameter $\eta_{\textit{birds,prec}}$ shows that *birds* prefer low precipitation. This information is used with small sample sizes of *icebird*: $\lambda_{\textit{icebird,prec}}$ matches $\eta_{\textit{birds,prec}}$. As the number of samples increases, single-class estimates for *icebird* become more accurate

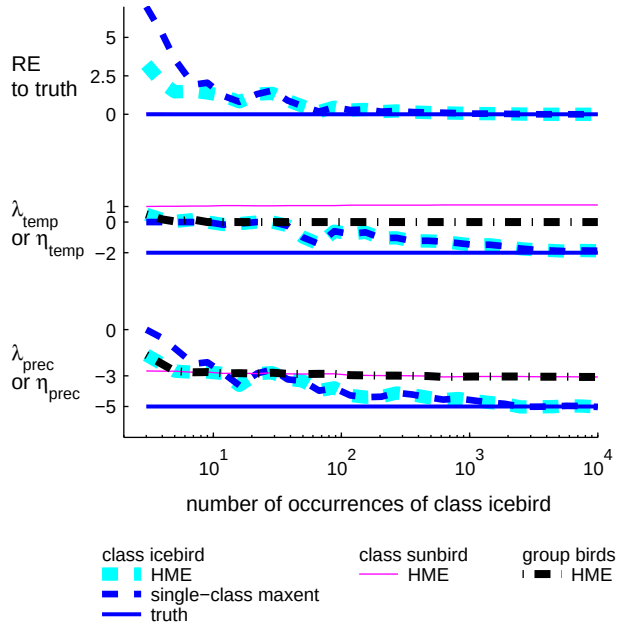

Figure 3. Synthetic experiments. The precipitation and temperature parameters of classes *icebird* and *sunbird*, and the group *birds*, are fitted by HME as the number of occurrences of *icebird* increases (the number of occurrences of *sunbird* is fixed at 100). Performance of the *icebird* models is reported in terms of relative entropy to the truth. Both the HME models and the single-species models of *icebird* converge to the truth, but HME performs better for small sample sizes, taking advantage of the group estimate of the precipitation parameter.

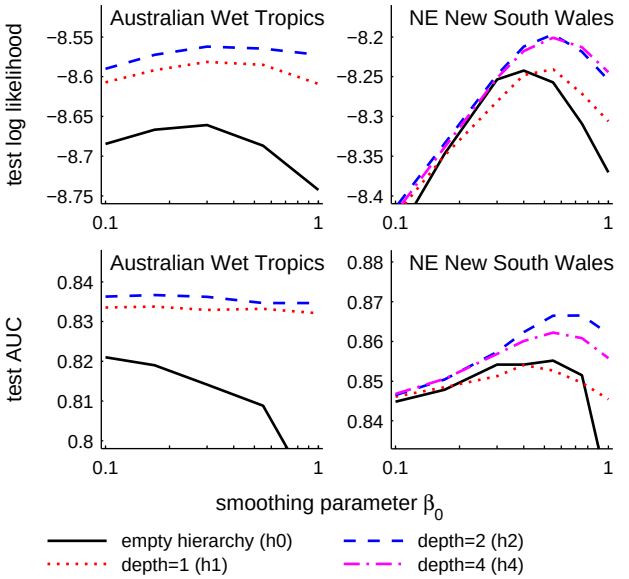

Figure 4. Performance of hierarchies with different depth over a range of smoothing parameters β_0. In AWT, the hierarchies *h1* and *h2* perform significantly better and are more robust to changes in β_0 than the empty hierarchy *h0*; no hierarchy beyond the depth two is available. In NSW, *h1*, *h2*, and *h4* perform significantly better than *h0*. The average performance of *h1* and *h0* appears similar, but *h1* improves the log likelihood of 18 out of 27 species, a significant departure from random improvements.

than group estimates, which is reflected in the HME parameters. In the top plot of Fig. 3, we see that the HME model performs better than the single-class model. As expected, the improvement is especially dramatic for small sample sizes. For moderate sample sizes, the HME estimates match single-class estimates exactly. This is qualitatively different from the James-Stein estimator, which always shrinks towards the pooled estimates.

Real-world data. Next, we demonstrate the performance of HME on a real-world dataset of species from the *Australian wet tropics* (AWT) and *Northeast New South Wales* (NSW). Species sample locations and environmental variables were all produced and used as part of the working group "Testing alternative methodologies for modeling species' ecological niches and predicting geographic distributions" at the National Center for Ecological Analysis and Synthesis (NCEAS). The working group compared modeling methods across a variety of species and regions. The training set contained presence-only data from unplanned surveys and incidental records. The test set contained presence-absence data from rigorously planned independent surveys. Single-class maxent was among the top methods in the NCEAS comparison (Elith et al., 2006).

In our experiments, we use only the training portion of the NCEAS dataset (to avoid problems with sample-selection bias). Specifically, we use a randomly chosen half of species in both AWT and NSW (we withhold the other half for future experiments). We use linear and quadratic features derived from 13 environmental variables in AWT and 12 environmental variables in NSW. We evaluate the performance of HME using five-fold cross-validation. The complete hierarchies, with the average number of training occurrences across all folds, are given in Figs. 1 and 2.

We run HME with three types of hierarchy for AWT and four types of hierarchy for NSW. In both regions we consider empty hierarchies, hierarchies of depth one, with the single group *all species*, and hierarchies of depth two. In AWT, the hierarchy of depth two is the complete hierarchy, in NSW, the hierarchy of depth two includes the groups *all species*, *birds*, *bats*, *small reptiles*, and *plants*. In NSW, we also consider a hierarchy of depth four (the complete hierarchy). Note that this hierarchy contains overlapping groups, so it cannot be expressed as a tree; however, this is not a problem for our setup. The hierarchies are referred to as *h0*, *h1*, *h2*, and *h4*, according to their depth.

As a metric of evaluation we use log likelihood and the area under the ROC curve (AUC). AUC is typically defined as a probability of ranking a randomly chosen positive above a randomly chosen negative. Since there are no negatives in our dataset, we treat all points in the region as negatives. Thus AUC corresponds to the probability that maxent prediction will rank a randomly chosen test sample above a sample chosen uniformly from the entire region. Thus, AUC is always below 1.0 and prediction by the uniform distribution receives the AUC of 0.5.

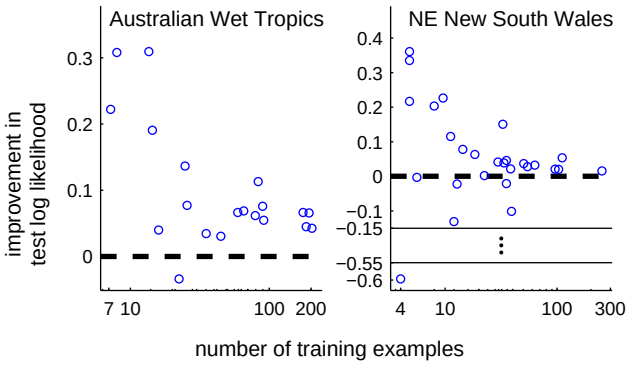

Figure 5. Improvement in performance using HME. We report the difference in test log likelihood between HME and single-class maxent for every species. The depth of the hierarchy is two. The improvement is the most dramatic for small sample sizes. The performance is significantly worse only in one case: a NE New South Wales species with only four training occurrence records.

In Figure 4, we report results for a range of smoothing parameters β_0. In each region, we show the average across all species. In AWT, performance of HME improves, both in terms of log likelihood and AUC, as the hierarchy gets more specific. The results are highly significant for log loss across all values of β_0: *h1* improves over *h0* on 19 species out of 20, and *h2* improves over *h1* on 15 species out of 20.

In NSW, we still observe that larger hierarchies tend to perform better than the smaller ones. Even though the differences appear rather small, it turns out that all three non-empty hierarchies are significantly better than *h0*. Specifically, *h1*, *h2*, and *h4* improve the log loss compared with *h0* over 18, 19, and 17 species out of 27, respectively, across all values β_0.

We use the word "significant" loosely to evoke the comparison with a random change. This null hypothesis is not entirely justified, because in non-empty hierarchies the models for different species influence one another. Note that we do not analyze the choice of the smoothing parameter β_0. We assume that in a concrete application, β_0 is set to a fixed value or determined by model selection.

The main benefit of HME should be observed on species with small numbers of samples. In Fig. 5, we show how the improvement due to the use of the group information varies across sample sizes. We use *h2*, with $\beta_0 = 0.3$ for AWT and $\beta_0 = 0.4$ for NSW. In AWT, the improvement is extremely consistent, and it appears to agree with the difference in relative entropy that we observed in synthetic experiments (Fig. 3). In NSW, we see the same trend on the vast majority of species. However, the performance is significantly worse in one case. It is the species with the smallest number of training occurrences — four.

7. Conclusion

We have applied the maximum entropy formalism to hierarchical models, where we simultaneously solve related estimation problems to improve each individual solution. We note that this method is not restricted to ℓ_1 regularization (or Laplace priors). It can be immediately generalized to arbitrary convex constraints (or log concave priors) along the same lines as the single-class maxent. This includes many widely used priors, such as those in the exponential family. The maximum entropy interpretation enhances our understanding of their generalization properties.

Acknowledgments. R. Schapire and M. Dudík received support through NSF grant CCR-0325463. D. Blei is supported by a grant from Google.

References

Baxter, J. (2000). A model of inductive bias learning. *J. Artif. Intell. Res.*, *12*, 149–198.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Caruana, R. (1993). Multitask learning: A knowledge-based source of inductive bias. *Machine Learning: Proceedings of the Tenth International Conference* (pp. 41–48).

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

Dudík, M., Phillips, S. J., & Schapire, R. E. (2004). Performance guarantees for regularized maximum entropy density estimation. *Proceedings of the Seventeenth Annual Conference on Learning Theory* (pp. 472–486). Springer-Verlag.

Dudík, M., & Schapire, R. E. (2006). Maximum entropy distribution estimation with generalized regularization. *Proc. Nineteenth Annual Conference on Learning Theory* (pp. 123–138).

Elith, J., Graham, C. H., & NCEAS Species Distribution Modelling Group (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

Ferrier, S., Drielsma, M., Manion, G., & Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. II. Community-level modelling. *Biodiv. Cons.*, *11*, 2309–2338.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Stat. Prob.* (pp. 311–319).

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Reviews*, *106*, 620–630.

Kazama, J., & Tsujii, J. (2003). Evaluation and extension of maximum entropy models with inequality constraints. *Conf. Empirical Methods in Natural Language Processing* (pp. 137–144).

Leathwick, J. R., Rowe, D., Richardson, J., Elith, J., & Hastie, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, *50*, 2034–2051.

McCallum, A., Rosenfeld, R., Mitchell, T. M., & Ng, A. Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. *Machine Learning: Proceedings of the Fifteenth International Conference* (pp. 359–367).

Raina, R., Ng, A. Y., & Koller, D. (2006). Constructing informative priors using transfer learning. *Proc. of the Twenty-Third International Conference on Machine Learning* (pp. 713–720).

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Stat. Prob.* (pp. 197–206).

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. *NIPS 17*.