Journal of Machine Learning Research 8 (2007) 1217-1260 Submitted 4/06; Revised 4/07; Published 6/07

Maximum Entropy Density Estimation with Generalized Regularization and an Application to Species Distribution Modeling

Miroslav Dudík MDUDIK@CS.PRINCETON.EDU
Princeton University
Department of Computer Science
35 Olden Street
Princeton, NJ 08540

Steven J. Phillips PHILLIPS@RESEARCH.ATT.COM
AT&T Labs – Research
180 Park Avenue
Florham Park, NJ 07932

Robert E. Schapire SCHAPIRE@CS.PRINCETON.EDU
Princeton University
Department of Computer Science
35 Olden Street
Princeton, NJ 08540

Editor: John Lafferty

Abstract

We present a unified and complete account of maximum entropy density estimation subject to constraints represented by convex potential functions or, alternatively, by convex regularization. We provide fully general performance guarantees and an algorithm with a complete convergence proof. As special cases, we easily derive performance guarantees for many known regularization types, including ℓ_1, ℓ_2, ℓ_2^2, and $\ell_1+\ell_2^2$ style regularization. We propose an algorithm solving a large and general subclass of generalized maximum entropy problems, including all discussed in the paper, and prove its convergence. Our approach generalizes and unifies techniques based on information geometry and Bregman divergences as well as those based more directly on compactness. Our work is motivated by a novel application of maximum entropy to species distribution modeling, an important problem in conservation biology and ecology. In a set of experiments on real-world data, we demonstrate the utility of maximum entropy in this setting. We explore effects of different feature types, sample sizes, and regularization levels on the performance of maxent, and discuss interpretability of the resulting models.

Keywords: maximum entropy, density estimation, regularization, iterative scaling, species distribution modeling

1. Introduction

The maximum entropy (maxent) approach to density estimation was first proposed by Jaynes (1957), and has since been used in many areas of computer science and statistical learning, especially natural language processing (Berger et al., 1996; Della Pietra et al., 1997). 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

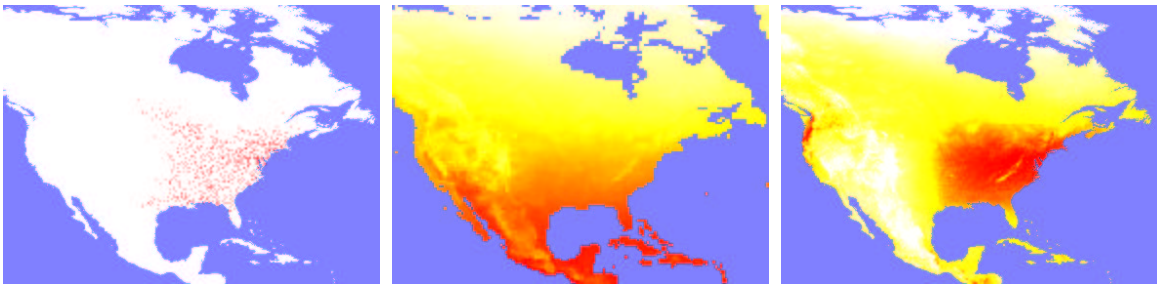

Figure 1: Left to right: Yellow-throated Vireo training localities from the first random partition, an example environmental variable (annual average temperature, higher values in red), maxent prediction using linear, quadratic and product features. Prediction strength is shown as white (weakest) to red (strongest); reds could be interpreted as suitable conditions for the species.

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 required to match its empirical average. By convex duality, this turns out to be the unique Gibbs distribution maximizing the likelihood of the samples, or, equivalently, minimizing the empirical log loss. (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. Input data for species distribution modeling consists of occurrence locations of a particular species in a region and of environmental variables for that region. Environmental variables may include topographical 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 type. Occurrence locations are commonly derived from specimen collections in natural history museums and herbaria. In the context of maxent, occurrences correspond to samples, the map divided into a finite number of cells is the sample space, and environmental variables or functions derived from them are features (see Figure 1 for an example). The number of occurrences for individual species is frequently quite small by machine learning standards, for example, a hundred or less.

It should not be surprising that maxent can severely overfit training data when the constraints on the output distribution are based on empirical averages, 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 features will almost never be equal to their true expectations, so the target distribution itself does not satisfy the constraints imposed on the output distribution. From the dual perspective, the family of Gibbs distributions is too expressive and the algorithm overfits. Common approaches to counter overfitting are parameter regularization (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001; Zhang, 2005), introduction of a prior (Williams, 1995; Goodman, 2004), feature selection (Berger et al., 1996; Della Pietra et al., 1997), discounting (Lau, 1994; Rosenfeld, 1996; Chen and Rosenfeld, 2000)

and constraint relaxation (Khudanpur, 1995; Kazama and Tsujii, 2003; Jedynak and Khudanpur, 2005). Thus, there are many ways of modifying maxent to control overfitting calling for a general treatment.

In this work, we study a generalized form of maxent. Although mentioned by other authors as *fuzzy maxent* (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001), we give the first complete theoretical treatment of this very general framework, including fully general and unified performance guarantees, algorithms, and convergence proofs. Independently, Altun and Smola (2006) derive a different theoretical treatment (see discussion below).

As special cases, our results allow us to easily derive performance guarantees for many known regularized formulations, including ℓ_1, ℓ_2, ℓ_2^2, and $\ell_1+\ell_2^2$ regularizations. More specifically, we derive guarantees on the performance of maxent solutions compared to the "best" Gibbs distribution $q^\star$ defined by a weight vector $\boldsymbol{\lambda}^\star$. Our guarantees are derived by bounding deviations of empirical feature averages from their expectations, a setting in which we can take advantage of a wide array of uniform convergence results. For example, for a finite set of features bounded in $[0,1]$, we can use Hoeffding's inequality and the union bound to show that the true log loss of the ℓ_1-regularized maxent solution will be with high probability worse by no more than an additive $O(\|\boldsymbol{\lambda}^\star\|_1\sqrt{(\ln n)/m})$ compared with the log loss of the Gibbs distribution $q^\star$, where n is the number of features and m is the number of samples. For an infinite set of binary features with VC-dimension d, the difference between the ℓ_1-regularized maxent solution and $q^\star$ is at most $O(\|\boldsymbol{\lambda}^\star\|_1\sqrt{d\ln(m^2/d)/m})$. Note that these bounds drop quickly with an increasing number of samples and depend only moderately on the number or complexity of the features, even admitting an extremely large number of features from a class of bounded VC-dimension. For maxent with ℓ_2 and ℓ_2^2-style regularization, it is possible to obtain bounds which are independent of the number of features, provided that the feature vector can be bounded in the ℓ_2 norm.

In the second part, we propose algorithms solving a large and general subclass of generalized maxent problems. We show convergence of our algorithms using a technique that unifies previous approaches and extends them to a more general setting. Specifically, our unified approach generalizes techniques based on information geometry and Bregman divergences (Della Pietra et al., 1997, 2001; Collins et al., 2002) as well as those based more directly on compactness. The main novel ingredient is a modified definition of an auxiliary function, a customary measure of progress, which we view as a surrogate for the difference between the primal and dual objective rather than a bound on the change in the dual objective.

Standard maxent algorithms such as iterative scaling (Darroch and Ratcliff, 1972; Della Pietra et al., 1997), gradient descent, Newton and quasi-Newton methods (Cesa-Bianchi et al., 1994; Malouf, 2002; Salakhutdinov et al., 2003), and their regularized versions (Lau, 1994; Williams, 1995; Chen and Rosenfeld, 2000; Kazama and Tsujii, 2003; Goodman, 2004; Krishnapuram et al., 2005) 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 (2002), and Lebanon and Lafferty (2001). This leads to a boosting-like approach permitting the selection of the best feature from a very large class. For instance, for ℓ_1-regularized maxent, 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. Other boosting-like approaches to density estimation have been proposed by Welling, Zemel, and Hinton (2003), and Rosset and Segal (2003).

For cases when the number of features is relatively small, yet we want to use benefits of regularization to prevent overfitting on small sample sets, it might be more efficient to solve generalized maxent by parallel updates. In Section 7, we give a parallel-update version of our algorithm with a proof of convergence.

In the last section, we return to species distribution modeling, and use it as a setting to test our ideas. In particular, we apply ℓ_1-regularized maxent to estimate distributions of bird species in North America. We present learning curves for several different feature classes derived for four species with a varying number of occurrence records. We also explore effects of regularization on the test log loss and interpretability of the resulting models. A more comprehensive set of experiments is evaluated by Phillips, Dudík, and Schapire (2004). The biological application is explored in more detail by Phillips, Anderson, and Schapire (2006).

1.1 Previous Work

There have been many studies of maxent and logistic regression, which is a conditional version of maxent, with ℓ_1-style regularization (Khudanpur, 1995; Williams, 1995; Kazama and Tsujii, 2003; Ng, 2004; Goodman, 2004; Krishnapuram et al., 2005), ℓ_2^2-style regularization (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001; Zhang, 2005) as well as some other types of regularization such as $\ell_1+\ell_2^2$-style (Kazama and Tsujii, 2003), ℓ_2-style regularization (Newman, 1977) and a smoothed version of ℓ_1-style regularization (Dekel et al., 2003). In a recent work, Altun and Smola (2006) derive duality and performance guarantees for settings in which the entropy is replaced by an arbitrary Bregman or Csiszár divergence and regularization takes the form of a norm raised to a power greater than one. With the exception of Altun and Smola's work and Zhang's work, the previous studies do not give performance guarantees applicable to our case, although Krishnapuram et al. (2005) and Ng (2004) prove guarantees for ℓ_1-regularized logistic regression. Ng also shows that ℓ_1-regularized logistic regression may be superior to the ℓ_2^2-regularized version in a scenario when the number of features is large and only a small number of them is relevant. Our results indicate a similar behavior for unconditional maxent.

In the context of linear models, ℓ_2^2, ℓ_1, and $\ell_1+\ell_2^2$ regularization have been used under the names *ridge regression* (Hoerl and Kennard, 1970), *lasso regression* (Tibshirani, 1996), and *elastic nets* (Zou and Hastie, 2005). Lasso regression, in particular, has provoked a lot of interest in recent statistical theory and practice. The frequently mentioned benefit of the lasso is its bias toward sparse solutions. The same bias is present also in ℓ_1-regularized maxent, but we do not analyze this bias in detail. Our interest is in deriving performance guarantees. Similar guarantees were derived by Donoho and Johnstone (1994) for linear models with the lasso penalty. The relationship between the lasso approximation and the sparsest approximation is explored, for example, by Donoho and Elad (2003).

Quite a number of approaches have been suggested for species distribution modeling, including neural nets, nearest neighbors, genetic algorithms, generalized linear models, generalized additive models, bioclimatic envelopes, boosted regression trees, and more; see Elith (2002) and Elith et al. (2006) for a comprehensive comparison. The latter work evaluates ℓ_1-regularized maxent as one of a group of twelve methods in the task of modeling species distributions. Maxent is among the best methods alongside boosted decision trees (Schapire, 2002; Leathwick et al., 2006), generalized dissimilarity models (Ferrier et al., 2002) and multivariate adaptive regression splines with the community level selection of basis functions (Moisen and Frescino, 2002; Leathwick et al., 2005).

Among these, however, maxent is the only method designed for presence-only data. It comes with a statistical interpretation that allows principled extensions, for example, to cases where the sampling process is biased (Dudík et al., 2005).

2. Preliminaries

Our goal is to estimate an unknown density π over a *sample space* $\mathcal{X}$ which, for the purposes of this paper, we assume to be finite.[1] As empirical information, we are typically given a set of *samples* $x_1, \ldots, x_m$ drawn independently at random according to π. The corresponding empirical distribution is denoted by $\tilde{\pi}$:

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

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

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

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, that is, the one that is closest to uniform. Here, as usual, the entropy of a distribution p on $\mathcal{X}$ is defined to be $\mathrm{H}(p) = -\sum_{x \in \mathcal{X}} p(x) \ln p(x)$.

However, the *default estimate* of π, that is, the distribution we would choose if we had no sample data, may be in some cases non-uniform. In a more general setup, we therefore seek a distribution that minimizes entropy relative to the default estimate q_0. The relative entropy, or Kullback-Leibler divergence, is an information theoretic measure defined as

$$\mathrm{D}(p \parallel q) = p[\ln(p/q)] \ .$$

Minimizing entropy relative to q_0 corresponds to choosing a distribution that is closest to q_0. When q_0 is uniform then minimizing entropy relative to q_0 is equivalent to maximizing entropy.

Instead of minimizing entropy relative to q_0, we can consider all *Gibbs distributions* of the form

$$q_{\boldsymbol{\lambda}}(x) = \frac{q_0(x) e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x)}}{Z_{\boldsymbol{\lambda}}}$$

where $Z_{\boldsymbol{\lambda}} = \sum_{x \in \mathcal{X}} q_0(x) e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x)}$ is a normalizing constant, and $\boldsymbol{\lambda} \in \mathbb{R}^n$. It can be proved (Della Pietra et al., 1997) that the maxent distribution is the same as the maximum likelihood distribution from the closure of the set of Gibbs distributions, that is, the distribution q that achieves the supremum of $\prod_{i=1}^m q_{\boldsymbol{\lambda}}(x_i)$ over all values of $\boldsymbol{\lambda}$, or equivalently, the infimum of the empirical log loss (negative normalized log likelihood)

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

1. In this paper, we are concerned with densities relative to the counting measure on $\mathcal{X}$. These correspond to probability mass functions.

The convex programs corresponding to the two optimization problems are

$$\min_{p\in\Delta} \mathrm{D}(p \parallel q_0) \text{ subject to } p[\mathbf{f}] = \tilde{\pi}[\mathbf{f}] \ , \tag{1}$$

$$\inf_{\boldsymbol{\lambda}\in\mathbb{R}^n} \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) \tag{2}$$

where Δ is the simplex of probability distributions over $\mathcal{X}$.

In general, we use

$$\mathrm{L}_r(\boldsymbol{\lambda}) = -r[\ln q_{\boldsymbol{\lambda}}]$$

to denote the log loss of $q_{\boldsymbol{\lambda}}$ relative to the distribution r. It differs from relative entropy $\mathrm{D}(r \parallel q_{\boldsymbol{\lambda}})$ only by the constant $\mathrm{H}(r)$. We will use the two interchangeably as objective functions.

3. Convex Analysis Background

Throughout this paper we make use of convex analysis. The necessary background is provided in this section. For a more detailed exposition see for example Rockafellar (1970), or Boyd and Vandenberghe (2004).

Consider a function $\psi : \mathbb{R}^n \to (-\infty, \infty]$. The *effective domain* of ψ is the set $\mathrm{dom}\,\psi = \{\mathbf{u} \in \mathbb{R}^n : \psi(\mathbf{u}) < \infty\}$. A point $\mathbf{u}$ where $\psi(\mathbf{u}) < \infty$ is called *feasible*. The *epigraph* of ψ is the set of points above its graph $\{(\mathbf{u}, t) \in \mathbb{R}^n \times \mathbb{R} : t \geq \psi(\mathbf{u})\}$. We say that ψ is *convex* if its epigraph is a convex set. A convex function is called *proper* if it is not uniformly equal to ∞. It is called *closed* if its epigraph is closed. For a proper convex function, closedness is equivalent to lower semi-continuity (ψ is lower semi-continuous if $\liminf_{\mathbf{u}'\to\mathbf{u}} \psi(\mathbf{u}') \geq \psi(\mathbf{u})$ for all $\mathbf{u}$).

If ψ is a closed proper convex function then its *conjugate* $\psi^* : \mathbb{R}^n \to (-\infty, \infty]$ is defined by

$$\psi^*(\boldsymbol{\lambda}) = \sup_{\mathbf{u}\in\mathbb{R}^n} [\boldsymbol{\lambda}\cdot\mathbf{u} - \psi(\mathbf{u})] \ .$$

The conjugate provides an alternative description of ψ in terms of tangents of ψ's epigraph. The definition of the conjugate immediately yields *Fenchel's inequality*

$$\forall \boldsymbol{\lambda}, \mathbf{u} : \boldsymbol{\lambda}\cdot\mathbf{u} \leq \psi^*(\boldsymbol{\lambda}) + \psi(\mathbf{u}) \ .$$

In fact, $\psi^*(\boldsymbol{\lambda})$ is defined to give the tightest bound of the form above. It turns out that ψ^* is also a closed proper convex function and $\psi^{**} = \psi$ (for a proof see Rockafellar, 1970, Corollary 12.2.1).

In this work we use several examples of closed proper convex functions. The first of them is relative entropy, viewed as a function of its first argument and extended to $\mathbb{R}^{\mathcal{X}}$ as follows:

$$\psi(p) = \begin{cases} \mathrm{D}(p \parallel q_0) & \text{if } p \in \Delta \\ \infty & \text{otherwise} \end{cases}$$

where $q_0 \in \Delta$ is assumed fixed. The conjugate of relative entropy is the log partition function

$$\psi^*(r) = \ln\left(\textstyle\sum_{x\in\mathcal{X}} q_0(x) e^{r(x)}\right)$$

where $r \in \mathbb{R}^{\mathcal{X}}$ and its components are denoted by $r(x)$.

The second example is the unnormalized relative entropy

$$\widetilde{\mathrm{D}}(p \parallel q_0) = \sum_{x\in\mathcal{X}} \left[p(x) \ln\left(\frac{p(x)}{q_0(x)}\right) - p(x) + q_0(x) \right] \ .$$

Fixing $q_0 \in [0,\infty)^{\mathcal{X}}$, it can be extended to a closed proper convex function of its first argument:

$$\psi(p) = \begin{cases} \widetilde{\mathrm{D}}(p \parallel q_0) & \text{if } p(x) \geq 0 \text{ for all } x \in \mathcal{X} \\ \infty & \text{otherwise.} \end{cases}$$

The conjugate of unnormalized relative entropy is a scaled exponential shifted to the origin:

$$\psi^*(r) = \sum_{x\in\mathcal{X}} q_0(x)(e^{r(x)} - 1) \ .$$

Both relative entropy and unnormalized relative entropy are examples of Bregman divergences (Bregman, 1967) which generalize some common distance measures including the squared Euclidean distance. We use two properties satisfied by any Bregman divergence $\mathrm{B}(\cdot \parallel \cdot)$:

(B1) $\mathrm{B}(\mathbf{a} \parallel \mathbf{b}) \geq 0$,
(B2) if $\mathrm{B}(\mathbf{a}_t \parallel \mathbf{b}_t) \to 0$ and $\mathbf{b}_t \to \mathbf{b}^\star$ then $\mathbf{a}_t \to \mathbf{b}^\star$.

It is not too difficult to check these properties explicitly both for relative entropy and unnormalized relative entropy.

Another example of a closed proper convex function is an *indicator function* of a closed convex set $C \subseteq \mathbb{R}^n$, denoted by I_C, which equals 0 when its argument lies in C and infinity otherwise. We will also use $\mathrm{I}(\mathbf{u} \in C)$ to denote $\mathrm{I}_C(\mathbf{u})$. The conjugate of an indicator function is a *support function*. For $C = \{\mathbf{u}_0\}$, we obtain $\mathrm{I}^*_{\{\mathbf{u}_0\}}(\boldsymbol{\lambda}) = \boldsymbol{\lambda} \cdot \mathbf{u}_0$. For a box $R = \{\mathbf{u} : |u_j| \leq \beta_j \text{ for all } j\}$, we obtain an ℓ_1-style conjugate $\mathrm{I}^*_R(\boldsymbol{\lambda}) = \sum_j \beta_j |\lambda_j|$. For a Euclidean ball $B = \{\mathbf{u} : \|\mathbf{u}\|_2 \leq \beta\}$, we obtain an ℓ_2-style conjugate, $\mathrm{I}^*_B(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_2$.

The final example is a square of the Euclidean norm $\psi(\mathbf{u}) = \|\mathbf{u}\|_2^2/(2\alpha)$, whose conjugate is also a square of the Euclidean norm $\psi^*(\boldsymbol{\lambda}) = \alpha\|\boldsymbol{\lambda}\|_2^2/2$.

The following identities can be proved from the definition of the conjugate function:

$$\text{if } \varphi(\mathbf{u}) = a\psi(b\mathbf{u} + \mathbf{c}) \qquad \text{then } \varphi^*(\boldsymbol{\lambda}) = a\psi^*(\boldsymbol{\lambda}/(ab)) - \boldsymbol{\lambda}\cdot\mathbf{c}/b \ , \tag{3}$$

$$\text{if } \varphi(\mathbf{u}) = \textstyle\sum_j \varphi_j(u_j) \qquad \text{then } \varphi^*(\boldsymbol{\lambda}) = \textstyle\sum_j \varphi_j^*(\lambda_j) \tag{4}$$

where $a > 0, b \neq 0$ and $\mathbf{c} \in \mathbb{R}^n$ are constants, and u_j, λ_j refer to the components of $\mathbf{u}, \boldsymbol{\lambda}$.

We conclude with a version of *Fenchel's Duality Theorem* which relates a convex minimization problem to a concave maximization problem using conjugates. The following result is essentially Corollary 31.2.1 of Rockafellar (1970) under a stronger set of assumptions.

Theorem 1 (Fenchel's Duality). *Let $\psi : \mathbb{R}^n \to (-\infty, \infty]$ and $\varphi : \mathbb{R}^m \to (-\infty, \infty]$ be closed proper convex functions and $\mathbf{A}$ a real-valued $m \times n$ matrix. Assume that $\operatorname{dom}\psi^* = \mathbb{R}^n$ or $\operatorname{dom}\varphi = \mathbb{R}^m$. Then*

$$\inf_{\mathbf{u}} \left[\psi(\mathbf{u}) + \varphi(\mathbf{A}\mathbf{u})\right] = \sup_{\boldsymbol{\lambda}} \left[-\psi^*(\mathbf{A}^\top\boldsymbol{\lambda}) - \varphi^*(-\boldsymbol{\lambda})\right] \ .$$

We refer to the minimization over $\mathbf{u}$ as the primal problem and the maximization over $\boldsymbol{\lambda}$ as the dual problem. When no ambiguity arises, we also refer to the minimization over $\boldsymbol{\lambda}$ of the negative dual objective as the dual problem. We call $\mathbf{u}$ a primal feasible point if the primal objective is finite at $\mathbf{u}$ and analogously define a dual feasible point.

4. Generalized Maximum Entropy

In this paper we study a generalized maxent problem

$$\mathcal{P}: \quad \min_{p\in\Delta}\big[\mathrm{D}(p \parallel q_0) + \mathrm{U}(p[\mathbf{f}])\big]$$

where $\mathrm{U}:\mathbb{R}^n \to (-\infty,\infty]$ is an arbitrary closed proper convex function. It is viewed as a *potential* for the maxent problem. We further assume that q_0 is positive on $\mathcal{X}$, that is, $\mathrm{D}(p \parallel q_0)$ is finite for all $p \in \Delta$ (otherwise we could restrict $\mathcal{X}$ to the support of q_0), and there exists a distribution whose vector of feature expectations is a feasible point of U (this is typically satisfied by the empirical distribution). These two conditions imply that the problem $\mathcal{P}$ is feasible.

The definition of generalized maxent captures many cases of interest including basic maxent, ℓ_1-regularized maxent and ℓ_2^2-regularized maxent. Basic maxent is obtained by using a point indicator potential $\mathrm{U}^{(0)}(\mathbf{u}) = \mathrm{I}(\mathbf{u} = \tilde{\pi}[\mathbf{f}])$. The ℓ_1-regularized version of maxent, as shown by Kazama and Tsujii (2003), corresponds to the relaxation of equality constraints to box constraints

$$|\tilde{\pi}[f_j] - p[f_j]| \le \beta_j \ .$$

This choice can be motivated by an observation that we do not expect $\tilde{\pi}[f_j]$ to be *equal* to $\pi[f_j]$ but only close to it. Box constraints are represented by the potential $\mathrm{U}^{(1)}(\mathbf{u}) = \mathrm{I}(|\tilde{\pi}[f_j] - u_j| \le \beta_j \text{ for all } j)$. Finally, as pointed out by Chen and Rosenfeld (2000) and Lebanon and Lafferty (2001), ℓ_2^2-regularized maxent is obtained using the potential $\mathrm{U}^{(2)}(\mathbf{u}) = \|\tilde{\pi}[\mathbf{f}] - \mathbf{u}\|_2^2/(2\alpha)$ which incurs an ℓ_2^2-style penalty for deviating from empirical averages.

The primal objective of generalized maxent will be referred to as P:

$$P(p) = \mathrm{D}(p \parallel q_0) + \mathrm{U}(p[\mathbf{f}]) \ .$$

Note that P attains its minimum over Δ, because Δ is compact and P is lower semi-continuous. The minimizer of P is unique by strict convexity of $\mathrm{D}(p \parallel q_0)$.

To derive the dual of $\mathcal{P}$, define the matrix $\mathbf{F}_{jx} = f_j(x)$ and use Fenchel's duality:

$$\begin{aligned}
\min_{p\in\Delta}[\mathrm{D}(p \parallel q_0) + \mathrm{U}(p[\mathbf{f}])] &= \min_{p\in\Delta}[\mathrm{D}(p \parallel q_0) + \mathrm{U}(\mathbf{F}p)] \\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^n}\left[-\ln\left(\textstyle\sum_{x\in\mathcal{X}} q_0(x)\exp\{(\mathbf{F}^\top\boldsymbol{\lambda})_x\}\right) - \mathrm{U}^*(-\boldsymbol{\lambda})\right] \qquad (5)\\
&= \sup_{\boldsymbol{\lambda}\in\mathbb{R}^n}[-\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}^*(-\boldsymbol{\lambda})] \ . \qquad (6)
\end{aligned}$$

In Equation (5), we apply Theorem 1. We use $(\mathbf{F}^\top\boldsymbol{\lambda})_x$ to denote the entry of $\mathbf{F}^\top\boldsymbol{\lambda}$ indexed by x. In Equation (6), we note that $(\mathbf{F}^\top\boldsymbol{\lambda})_x = \boldsymbol{\lambda}\cdot\mathbf{f}(x)$ and thus the expression inside the logarithm is the normalization constant of $q_{\boldsymbol{\lambda}}$. The dual objective will be referred to as Q:

$$Q(\boldsymbol{\lambda}) = -\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}^*(-\boldsymbol{\lambda}) \ .$$

There are two formal differences between generalized maxent and basic maxent. The first difference is that the constraints of the basic primal (1) are stated relative to the empirical expectations whereas the potential of the generalized primal $\mathcal{P}$ makes no reference to $\tilde{\pi}[\mathbf{f}]$. This difference is only superficial. It is possible to "hard-wire" the distribution $\tilde{\pi}$ in the potential U, as we saw on

	potential (absolute and relative)	conjugate potential
generalized maxent:		
$\mathrm{U}(\mathbf{u})$	$\mathrm{U}(\mathbf{u})$	$\mathrm{U}^*(\boldsymbol{\lambda})$
$\mathrm{U}_r(\mathbf{u})$	$\mathrm{U}(r[\mathbf{f}]-\mathbf{u})$	$\mathrm{U}^*(-\boldsymbol{\lambda})+\boldsymbol{\lambda}\cdot r[\mathbf{f}]$
$\mathrm{U}_{\tilde{\pi}}(\mathbf{u})$	$\mathrm{U}(\tilde{\pi}[\mathbf{f}]-\mathbf{u})$	$\mathrm{U}^*(-\boldsymbol{\lambda})+\boldsymbol{\lambda}\cdot\tilde{\pi}[\mathbf{f}]$
basic constraints:		
$\mathrm{U}^{(0)}(\mathbf{u})$	$\mathrm{I}(\mathbf{u}=\tilde{\pi}[\mathbf{f}])$	$\boldsymbol{\lambda}\cdot\tilde{\pi}[\mathbf{f}]$
$\mathrm{U}^{(0)}_r(\mathbf{u})$	$\mathrm{I}(\mathbf{u}=r[\mathbf{f}]-\tilde{\pi}[\mathbf{f}])$	$\boldsymbol{\lambda}\cdot(r[\mathbf{f}]-\tilde{\pi}[\mathbf{f}])$
$\mathrm{U}^{(0)}_{\tilde{\pi}}(\mathbf{u})$	$\mathrm{I}(\mathbf{u}=\mathbf{0})$	0
box constraints:		
$\mathrm{U}^{(1)}(\mathbf{u})$	$\mathrm{I}(\lvert\tilde{\pi}[f_j]-u_j\rvert\le\beta_j \text{ for all } j)$	$\boldsymbol{\lambda}\cdot\tilde{\pi}[\mathbf{f}]+\sum_j\beta_j\lvert\lambda_j\rvert$
$\mathrm{U}^{(1)}_r(\mathbf{u})$	$\mathrm{I}(\lvert u_j-(r[f_j]-\tilde{\pi}[f_j])\rvert\le\beta_j \text{ for all } j)$	$\boldsymbol{\lambda}\cdot(r[\mathbf{f}]-\tilde{\pi}[\mathbf{f}])+\sum_j\beta_j\lvert\lambda_j\rvert$
$\mathrm{U}^{(1)}_{\tilde{\pi}}(\mathbf{u})$	$\mathrm{I}(\lvert u_j\rvert\le\beta_j \text{ for all } j)$	$\sum_j\beta_j\lvert\lambda_j\rvert$
ℓ_2^2 penalty:		
$\mathrm{U}^{(2)}(\mathbf{u})$	$\lVert\tilde{\pi}[\mathbf{f}]-\mathbf{u}\rVert_2^2/(2\alpha)$	$\boldsymbol{\lambda}\cdot\tilde{\pi}[\mathbf{f}]+\alpha\lVert\boldsymbol{\lambda}\rVert_2^2/2$
$\mathrm{U}^{(2)}_r(\mathbf{u})$	$\lVert\mathbf{u}-(r[\mathbf{f}]-\tilde{\pi}[\mathbf{f}])\rVert_2^2/(2\alpha)$	$\boldsymbol{\lambda}\cdot(r[\mathbf{f}]-\tilde{\pi}[\mathbf{f}])+\alpha\lVert\boldsymbol{\lambda}\rVert_2^2/2$
$\mathrm{U}^{(2)}_{\tilde{\pi}}(\mathbf{u})$	$\lVert\mathbf{u}\rVert_2^2/(2\alpha)$	$\alpha\lVert\boldsymbol{\lambda}\rVert_2^2/2$

Table 1: Absolute and relative potentials, and their conjugates for various versions of maxent.

the example of $\mathrm{U}^{(0)}$. In the latter case, it would be more correct, but perhaps overly pedantic and somewhat clumsy, to make the dependence of the potential on $\tilde{\pi}$ explicit and use the notation $\mathrm{U}^{(0),\tilde{\pi}}$.

The second difference, which seems more significant, is the difference between the duals. The objective of the basic dual (2) equals the log loss relative to the empirical distribution $\tilde{\pi}$, but the log loss does not appear in the generalized dual. However, we will see that the generalized dual can be expressed in terms of the log loss. In fact, it can be expressed in terms of the log loss relative to an arbitrary distribution, including the empirical distribution $\tilde{\pi}$ as well as the unknown distribution π.

We next describe *shifting*, the transformation of an "absolute" potential to a "relative" potential. Shifting is a technical tool which will simplify some of the proofs in Sections 5 and 6, and will also be used to rewrite the generalized dual in terms of the log loss.

4.1 Shifting

For an arbitrary distribution r and a potential U, let U_r denote the function

$$\mathrm{U}_r(\mathbf{u}) = \mathrm{U}(r[\mathbf{f}]-\mathbf{u}) \ .$$

This function will be referred to as the *potential relative to r* or simply the *relative potential*. The original potential U will be in contrast referred to as the *absolute potential*. In Table 1, we list potentials discussed so far, alongside their versions relative to an arbitrary distribution r, and relative to $\tilde{\pi}$ in particular.

From the definition of a relative potential, we see that the absolute potential can be expressed as $\mathrm{U}(\mathbf{u}) = \mathrm{U}_r(r[\mathbf{f}]-\mathbf{u})$. Thus, it is possible to implicitly define a potential U by defining a relative potential U_r for a particular distribution r. The potentials $\mathrm{U}^{(0)}$, $\mathrm{U}^{(1)}$, $\mathrm{U}^{(2)}$ of basic maxent, maxent with box constraints, and maxent with ℓ_2^2 penalty could thus have been specified by defining $\mathrm{U}^{(0)}_{\tilde{\pi}}(\mathbf{u}) = \mathrm{I}(\mathbf{u}=\mathbf{0})$, $\mathrm{U}^{(1)}_{\tilde{\pi}}(\mathbf{u}) = \mathrm{I}(\lvert u_j\rvert\le\beta_j \text{ for all } j)$ and $\mathrm{U}^{(2)}_{\tilde{\pi}}(\mathbf{u}) = \lVert\mathbf{u}\rVert_2^2/(2\alpha)$.

The conjugate of a relative potential, the *conjugate relative potential*, is obtained, according to Equation (3), by adding a linear function to the conjugate of U:

$$\mathrm{U}_r^*(\boldsymbol{\lambda}) = \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot r[\mathbf{f}] \ . \tag{7}$$

Table 1 lists $\mathrm{U}^{(0)*}$, $\mathrm{U}^{(1)*}$, $\mathrm{U}^{(2)*}$, and the conjugates of the corresponding relative potentials.

4.2 The Generalized Dual as the Minimization of a Regularized Log Loss

We will now show how the dual objective $Q(\boldsymbol{\lambda})$ can be expressed in terms of the log loss relative to an arbitrary distribution r. This will highlight how the dual of the generalized maxent extends the dual of the basic maxent. Using Equation (7), we rewrite $Q(\boldsymbol{\lambda})$ as follows:

$$\begin{aligned} Q(\boldsymbol{\lambda}) &= -\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}^*(-\boldsymbol{\lambda}) = -\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}_r^*(\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot r[\mathbf{f}] \\ &= -r[\ln q_0] + r[\ln q_0 + \boldsymbol{\lambda} \cdot \mathbf{f} - \ln Z_{\boldsymbol{\lambda}}] - \mathrm{U}_r^*(\boldsymbol{\lambda}) \\ &= \mathrm{L}_r(\mathbf{0}) - \mathrm{L}_r(\boldsymbol{\lambda}) - \mathrm{U}_r^*(\boldsymbol{\lambda}) \ . \end{aligned} \tag{8}$$

Since the first term in Equation (8) is a constant independent of $\boldsymbol{\lambda}$, the maximization of $Q(\boldsymbol{\lambda})$ is equivalent to the minimization of $\mathrm{L}_r(\boldsymbol{\lambda}) + \mathrm{U}_r^*(\boldsymbol{\lambda})$. Setting $r = \tilde{\pi}$ we obtain a dual analogous to the basic dual (2):

$$Q_{\tilde{\pi}} : \quad \inf_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^*(\boldsymbol{\lambda}) \right] \ .$$

From Equation (8), it follows that the $\boldsymbol{\lambda}$ minimizing $\mathrm{L}_r(\boldsymbol{\lambda}) + \mathrm{U}_r^*(\boldsymbol{\lambda})$ does not depend on a particular choice of r. As a result, the minimizer of $Q_{\tilde{\pi}}$ is also the minimizer of $\mathrm{L}_\pi(\boldsymbol{\lambda}) + \mathrm{U}_\pi^*(\boldsymbol{\lambda})$. This observation will be used in Section 5 to prove performance guarantees.

The objective of $Q_{\tilde{\pi}}$ has two terms. The first of them is the empirical log loss. The second one is the regularization term penalizing "complex" solutions. The regularization term need not be non-negative and it does not necessarily increase with any norm of $\boldsymbol{\lambda}$. On the other hand, it is a proper closed convex function and if $\tilde{\pi}$ is feasible then by Fenchel's inequality the regularization is bounded from below by $-\mathrm{U}_{\tilde{\pi}}(\mathbf{0})$. From a Bayesian perspective, $\mathrm{U}_{\tilde{\pi}}^*$ corresponds to negative log of the prior, and minimizing $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^*(\boldsymbol{\lambda})$ is equivalent to maximizing the posterior.

In the case of basic maxent, we obtain $\mathrm{U}_{\tilde{\pi}}^{(0)*}(\boldsymbol{\lambda}) = 0$ and recover the basic dual. For the box potential, we obtain $\mathrm{U}_{\tilde{\pi}}^{(1)*}(\boldsymbol{\lambda}) = \sum_j \beta_j |\lambda_j|$, which corresponds to an ℓ_1-style regularization and a Laplace prior. For the ℓ_2^2 potential, we obtain $\mathrm{U}_{\tilde{\pi}}^{(2)*}(\boldsymbol{\lambda}) = \alpha \|\boldsymbol{\lambda}\|_2^2/2$, which corresponds to an ℓ_2^2-style regularization and a Gaussian prior.

In all the cases discussed in this paper, it is natural to consider the dual objective relative to $\tilde{\pi}$ as we have seen in the previous examples. In other cases, the empirical distribution $\tilde{\pi}$ need not be available, and there may be no natural distribution relative to which a potential could be specified, yet it is possible to define a meaningful absolute potential (Dudík et al., 2005; Dudík and Schapire, 2006). To capture the more general case, we formulate the generalized maxent using the absolute potential.

4.3 Maxent Duality

We know from Equation (6) that the generalized maxent primal and dual have equal *values*. In this section, we show the equivalence of the primal and dual *optimizers*. Specifically, we show that the maxent primal $\mathcal{P}$ is solved by the Gibbs distribution whose parameter vector $\boldsymbol{\lambda}$ solves the dual

(possibly in a limit). This parallels the result of Della Pietra, Della Pietra, and Lafferty (1997) for the basic maxent and gives additional motivation for the view of the dual objective as the regularized log loss.

Theorem 2 (Maxent Duality). *Let q_0, U, P, Q be as above. Then*

$$\min_{p \in \Delta} P(p) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^n} Q(\boldsymbol{\lambda}) \ . \tag{9}$$

Moreover, for a sequence $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ such that

$$\lim_{t \to \infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^n} Q(\boldsymbol{\lambda})$$

the sequence of $q_t = q_{\boldsymbol{\lambda}_t}$ has a limit and

$$P\left(\lim_{t \to \infty} q_t\right) = \min_{p \in \Delta} P(p) \ . \tag{10}$$

Proof. Equation (9) is a consequence of Fenchel's duality as was shown earlier. It remains to prove Equation (10). We will use an alternative expression for the dual objective. Let r be an arbitrary distribution. Adding and subtracting $\mathrm{H}(r)$ from Equation (8) yields

$$Q(\boldsymbol{\lambda}) = -\mathrm{D}(r \parallel q_{\boldsymbol{\lambda}}) + \mathrm{D}(r \parallel q_0) - \mathrm{U}_r^*(\boldsymbol{\lambda}) \ . \tag{11}$$

Let $\hat{p}$ be the minimizer of P and $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ maximize Q in the limit. Then

$$\begin{aligned} \mathrm{D}(\hat{p} \parallel q_0) + \mathrm{U}_{\hat{p}}(\mathbf{0}) = P(\hat{p}) &= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^n} Q(\boldsymbol{\lambda}) = \lim_{t \to \infty} Q(\boldsymbol{\lambda}_t) \\ &= \lim_{t \to \infty} \left[-\mathrm{D}(\hat{p} \parallel q_t) + \mathrm{D}(\hat{p} \parallel q_0) - \mathrm{U}_{\hat{p}}^*(\boldsymbol{\lambda}_t) \right] \ . \end{aligned}$$

Denoting terms with the limit 0 by $o(1)$ and rearranging yields

$$\mathrm{U}_{\hat{p}}(\mathbf{0}) + \mathrm{U}_{\hat{p}}^*(\boldsymbol{\lambda}_t) = -\mathrm{D}(\hat{p} \parallel q_t) + o(1) \ .$$

The left-hand side is non-negative by Fenchel's inequality, so $\mathrm{D}(\hat{p} \parallel q_t) \to 0$ by the non-negativity of relative entropy. Therefore, by property (B2), every convergent subsequence of $q_1, q_2, \ldots$ has the limit $\hat{p}$. Since the q_t's come from the compact set Δ, we obtain $q_t \to \hat{p}$. ∎

Thus, in order to solve the primal, it suffices to find a sequence of $\boldsymbol{\lambda}$'s maximizing the dual. This will be the goal of algorithms in Sections 6 and 7.

5. Bounding the Loss on the Target Distribution

In this section, we derive bounds on the performance of generalized maxent relative to the true distribution π. That is, we are able to bound $\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})$ in terms of $\mathrm{L}_\pi(\boldsymbol{\lambda}^\star)$ when $q_{\hat{\boldsymbol{\lambda}}}$ maximizes the dual objective Q and $q_{\boldsymbol{\lambda}^\star}$ is either an arbitrary Gibbs distribution, or in some cases, a Gibbs distribution with a bounded norm of $\boldsymbol{\lambda}^\star$. In particular, bounds hold for the Gibbs distribution minimizing the true loss (in some cases, among Gibbs distributions with a bounded norm of $\boldsymbol{\lambda}^\star$). Note that $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}})$ differs from $\mathrm{L}_\pi(\boldsymbol{\lambda})$ only by the constant term $\mathrm{H}(\pi)$, so identical bounds also hold for $\mathrm{D}(\pi \parallel q_{\hat{\boldsymbol{\lambda}}})$ in terms of $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}^\star})$.

Our results are stated for the case when the supremum of Q is attained at $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^n$, but they easily extend to the case when the supremum is only attained in a limit. The crux of our method is the lemma below. Even though its proof is remarkably simple, it is sufficiently general to cover all the cases of interest.

Lemma 3. *Let $\hat{\boldsymbol{\lambda}}$ maximize Q. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\mathrm{U}(\pi[\mathbf{f}]) + \mathrm{U}^*(\boldsymbol{\lambda}^\star) + \mathrm{U}^*(-\boldsymbol{\lambda}^\star) \ , \tag{12}$$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\mathrm{U}_{\tilde{\pi}}(\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}]) + \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}^\star) + \mathrm{U}^*_{\tilde{\pi}}(-\boldsymbol{\lambda}^\star) \ , \tag{13}$$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + (\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}) \cdot (\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]) + \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}^\star) - \mathrm{U}^*_{\tilde{\pi}}(\hat{\boldsymbol{\lambda}}) \ . \tag{14}$$

Proof. Optimality of $\hat{\boldsymbol{\lambda}}$ with respect to $\mathrm{L}_\pi(\boldsymbol{\lambda}) + \mathrm{U}^*_\pi(\boldsymbol{\lambda}) = -Q(\boldsymbol{\lambda}) + \mathit{const.}$ yields

$$\begin{aligned}\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{U}^*_\pi(\boldsymbol{\lambda}^\star) - \mathrm{U}^*_\pi(\hat{\boldsymbol{\lambda}}) \\ &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + (\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}) \cdot \pi[\mathbf{f}] + \mathrm{U}^*(-\boldsymbol{\lambda}^\star) - \mathrm{U}^*(-\hat{\boldsymbol{\lambda}}) \ . \end{aligned} \tag{15}$$

In Equation (15), we express U^*_π in terms of U^* using Equation (7). Now Equation (12) is obtained by applying Fenchel's inequality to the second term of Equation (15):

$$(\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}) \cdot \pi[\mathbf{f}] \le \mathrm{U}^*(\boldsymbol{\lambda}^\star) + \mathrm{U}(\pi[\mathbf{f}]) + \mathrm{U}^*(-\hat{\boldsymbol{\lambda}}) + \mathrm{U}(\pi[\mathbf{f}]) \ .$$

Equations (13) and (14) follow from Equations (12) and (15) by shifting potentials and their conjugates to $\tilde{\pi}$. ∎

Remark. Notice that π and $\tilde{\pi}$ in the statement and the proof of the lemma can be replaced by arbitrary distributions p_1 and p_2.

A special case which we discuss in more detail is when U is an indicator of a closed convex set C, such as $\mathrm{U}^{(0)}$ and $\mathrm{U}^{(1)}$ of the previous section. In that case, the right hand side of Lemma 3.12 will be infinite unless $\pi[\mathbf{f}]$ lies in C. In order to apply Lemma 3.12, we ensure that $\pi[\mathbf{f}] \in C$ with high probability. Therefore, we choose C as a confidence region for $\pi[\mathbf{f}]$. If $\pi[\mathbf{f}] \in C$ then for any Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_C(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_C(-\boldsymbol{\lambda}^\star) \ . \tag{16}$$

For a fixed $\boldsymbol{\lambda}^\star$ and a non-empty C, $\mathrm{I}^*_C(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_C(-\boldsymbol{\lambda}^\star)$ is always non-negative and proportional to the size of C's projection onto a line in the direction $\boldsymbol{\lambda}^\star$. Thus, smaller confidence regions yield better performance guarantees.

A common method of obtaining confidence regions is to bound the difference between empirical averages and true expectations. There exists a huge array of techniques to achieve this. Before moving to specific examples, we state a general result which follows directly from Lemma 3.13 analogously to Equation (16).

Theorem 4. *Assume that $\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}] \in C_0$ where C_0 is a closed convex set symmetric around the origin. Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{I}^*_{C_0}(\boldsymbol{\lambda})$. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\mathrm{I}^*_{C_0}(\boldsymbol{\lambda}^\star) \ .$$

Proof. Setting $\mathrm{U}_{\tilde{\pi}}(\mathbf{u}) = \mathrm{I}_{C_0}(\mathbf{u})$ and assuming $\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}] \in C_0$, we obtain by Lemma 3.13

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_{C_0}(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_{C_0}(-\boldsymbol{\lambda}^\star) \ .$$

The result now follows by the symmetry of C_0, which implies the symmetry of I_{C_0}, which in turn implies the symmetry of $\mathrm{I}^*_{C_0}$. ∎

5.1 Maxent with ℓ_1 Regularization

We now apply the foregoing general results to some specific cases of interest. To begin, we consider the box indicator $\mathrm{U}^{(1)}$ of Section 4. In this case it suffices to bound $|\tilde{\pi}[f_j]-\pi[f_j]|$ and use Theorem 4 to obtain a bound on the true loss $\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})$. For instance, when the features are bounded, we can prove the following:

Corollary 5. *Assume that features $f_1,\ldots,f_n$ are bounded in $[0,1]$. Let $\delta>0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\beta\|\boldsymbol{\lambda}\|_1$ with $\beta=\sqrt{\ln(2n/\delta)/(2m)}$. Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,*

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

Proof. By Hoeffding's inequality, for a fixed j, the probability that $|\tilde{\pi}[f_j]-\pi[f_j]|$ exceeds β is at most $2e^{-2\beta^2 m}=\delta/n$. By the union bound, the probability of this happening for any j is at most δ. The claim now follows immediately from Theorem 4. ■

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 6. *Assume that features are binary with VC-dimension d. Let $\delta>0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\beta\|\boldsymbol{\lambda}\|_1$ with*

$$\beta=\sqrt{\frac{d\ln(em^2/d)+\ln(1/\delta)+\ln(4e^8)}{2m}} \ .$$

Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star)+\frac{\|\boldsymbol{\lambda}^\star\|_1}{\sqrt{m}}\sqrt{2[d\ln(em^2/d)+\ln(1/\delta)+\ln(4e^8)]} \ .$$

Proof. Here, a uniform-convergence result of Devroye (1982), combined with Sauer's Lemma, can be used to argue that $|\tilde{\pi}[f_j]-\pi[f_j]|\le\beta$ for all f_j simultaneously with probability at least $1-\delta$. ■

The final result for ℓ_1-regularized maxent is motivated by the Central Limit Theorem approximation $|\tilde{\pi}[f_j]-\pi[f_j]|=O(\sigma[f_j]/\sqrt{m})$, where $\sigma[f_j]$ is the standard deviation of f_j under π. We bound $\sigma[f_j]$ from above using McDiarmid's inequality for the empirical estimate of variance

$$\tilde{\sigma}^2[f_j]=\frac{m\big(\tilde{\pi}[f_j^2]-\tilde{\pi}[f_j]^2\big)}{m-1} \ ,$$

and then obtain non-asymptotic bounds on $|\tilde{\pi}[f_j]-\pi[f_j]|$ by Bernstein's inequality (for a complete proof see Appendix A).

We believe that this type of result may in practice be more useful than Corollaries 5 and 6, because it allows differentiation between features depending on empirical error estimates computed from the sample data. Motivated by Corollary 7 below, in Section 8 we describe experiments that use $\beta_j=\beta_0\tilde{\sigma}[f_j]/\sqrt{m}$, where β_0 is a single tuning constant. This approach is equivalent to using features scaled to the unit sample variance, that is, features $f'_j(x)=f_j(x)/\tilde{\sigma}[f_j]$, and a regularization parameter independent of features, $\beta'_j=\beta_0/\sqrt{m}$, as is a common practice in statistics. Corollary 7 justifies this practice and also suggests replacing the sample variance by a slightly larger value $\tilde{\sigma}^2[f_j]+O(1/\sqrt{m})$.

Corollary 7. *Assume that features $f_1, \ldots, f_n$ are bounded in $[0,1]$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$ with*

$$\beta_j = \sqrt{\frac{2\ln(4n/\delta)}{m}} \cdot \sqrt{\tilde{\sigma}^2[f_j] + \sqrt{\frac{\ln(2n/\delta)}{2m}} + \frac{\ln(4n/\delta)}{18m}} + \frac{\ln(4n/\delta)}{3m} \quad .$$

Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\sum_j \beta_j |\lambda_j^\star| \quad .$$

Corollaries of this section show that the difference in performance between the distribution computed by minimizing ℓ_1-regularized log loss and the best Gibbs distribution becomes small rapidly as the number of samples m increases. Note that this difference depends only moderately on the number or complexity of features.

Another feature of ℓ_1 regularization is that it induces sparsity (Tibshirani, 1996). Note that a maxent solution $\hat{\boldsymbol{\lambda}}$ is "truly" sparse, that is, some of its components are "truly" zero, only if they remain zero under perturbations in the regularization parameters β_j and the expectations $\tilde{\pi}[f_j]$; in other words, the fact that the components of $\hat{\boldsymbol{\lambda}}$ are zero is not just a lucky coincidence. To see how ℓ_1 regularization induces this property, notice that its partial derivatives are discontinuous at $\lambda_j = 0$. As a consequence, if the regularized log loss is uniquely minimized at a point where the j_0-th component $\hat{\lambda}_{j_0}$ equals zero, then the optimal $\hat{\lambda}_{j_0}$ will remain zero even if the parameters β_j and the expectations $\tilde{\pi}[f_j]$ are slightly perturbed.

5.2 Maxent with Smoothed ℓ_1 Regularization

While the guarantees for ℓ_1-style regularization have many favorable properties, the fact that the ℓ_1 norm is not strictly convex and its first derivative is discontinuous at zero may sometimes be problematic. The lack of strict convexity may lead to infinitely many $\boldsymbol{\lambda}$'s optimizing the dual objective,[2] and the discontinuous derivatives may cause problems in certain convex optimization algorithms. To prevent these problems, smooth approximations of ℓ_1 regularization may be necessary.

In this section, we analyze a smooth approximation similar to one used by Dekel, Shalev-Shwartz, and Singer (2003):

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}(\boldsymbol{\lambda}) = \sum_j \alpha_j \beta_j \ln\cosh(\lambda_j/\alpha_j) = \sum_j \alpha_j \beta_j \ln\left(\frac{e^{\lambda_j/\alpha_j} + e^{-\lambda_j/\alpha_j}}{2}\right) \quad .$$

Constants $\alpha_j > 0$ control the tightness of fit to the ℓ_1 norm while constants $\beta_j \ge 0$ control scaling. Note that $\cosh x \le e^{|x|}$ hence

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}(\boldsymbol{\lambda}) \le \sum_j \alpha_j \beta_j \ln e^{|\lambda_j|/\alpha_j} = \sum_j \alpha_j \beta_j |\lambda_j|/\alpha_j = \sum_j \beta_j |\lambda_j| \quad . \tag{17}$$

The potential corresponding to $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}$ is

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}(\mathbf{u}) = \sum_j \alpha_j \beta_j \mathrm{D}\left(\frac{1+u_j/\beta_j}{2} \,\middle\|\, \frac{1}{2}\right)$$

2. This may only happen if features are not linearly independent.

where $\mathrm{D}(a \parallel b)$ is a shorthand for $\mathrm{D}((a,1-a) \parallel (b,1-b))$ (for a derivation of $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$ see Appendix B). This potential can be viewed as a smooth upper bound on the box potential $\mathrm{U}_{\tilde{\pi}}^{(1)}$ in the sense that the gradient of $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$ is continuous on the interior of the effective domain of $\mathrm{U}_{\tilde{\pi}}^{(1)}$ and its norm approaches ∞ on the border. Note that if $|u_j| \leq \beta_j$ for all j then $\mathrm{D}\big(\frac{1+u_j/\beta_j}{2} \,\big\|\, \frac{1}{2}\big) \leq \mathrm{D}\big(0 \,\big\|\, \frac{1}{2}\big) = \ln 2$ and hence

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}(\mathbf{u}) \leq (\ln 2) \textstyle\sum_j \alpha_j \beta_j \quad . \tag{18}$$

Applying bounds (17) and (18) in Lemma 3.13 we obtain an analog of Theorem 4.

Theorem 8. *Assume that for each j, $|\tilde{\pi}[f_j] - \pi[f_j]| \leq \beta_j$. Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}(\boldsymbol{\lambda})$. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\textstyle\sum_j \beta_j |\lambda_j^\star| + (2\ln 2)\sum_j \alpha_j \beta_j \quad .$$

To obtain guarantees analogous to those of ℓ_1-regularized maxent, it suffices to choose sufficiently small α_j. For example, in order to perform well relative to distributions $q_{\boldsymbol{\lambda}^\star}$ with $\sum_j \beta_j |\lambda_j^\star| \leq L$, it suffices to choose $\alpha_j = (\varepsilon L)/(n\beta_j \ln 2)$ and obtain

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2(1+\varepsilon)L \quad .$$

For example, we can derive an analog of Corollary 5. We relax the constraint that features are bounded in $[0,1]$ and, instead, provide a guarantee in terms of the ℓ_∞ diameter of the feature space.

Corollary 9. *Let $D_\infty = \sup_{x,x' \in \mathcal{X}} \|\mathbf{f}(x) - \mathbf{f}(x')\|_\infty$ be the ℓ_∞ diameter of $\mathbf{f}(\mathcal{X})$. Let $\delta, \varepsilon, L_1 > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\beta \sum_j \ln\cosh(\lambda_j/\alpha)$ with*

$$\alpha = \frac{\varepsilon L_1}{n \ln 2} \quad , \qquad \beta = D_\infty \sqrt{\frac{\ln(2n/\delta)}{2m}} \quad .$$

Then with probability at least $1-\delta$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \inf_{\|\boldsymbol{\lambda}^\star\|_1 \leq L_1} \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{(1+\varepsilon) L_1 D_\infty}{\sqrt{m}} \cdot \sqrt{2\ln(2n/\delta)} \quad .$$

Thus, maxent with smoothed ℓ_1 regularization performs almost as well as ℓ_1-regularized maxent, provided that we specify an upper bound on the ℓ_1 norm of $\boldsymbol{\lambda}^\star$ in advance. As a result of removing discontinuities in the gradient, smoothed ℓ_1 regularization lacks the sparsity inducing properties of ℓ_1 regularization.

As $\alpha \to 0$, the guarantees for smoothed ℓ_1 regularization converge to those for ℓ_1 regularization, but at the price of reducing smoothness of the objective in some regions and increasing its flatness in others. For many methods of convex optimization, this leads to a worse runtime. For example, the number of iterations of gradient descent increases with an increasing condition number of the Hessian of the objective, which in our case grows as $\alpha \to 0$. Similarly, the number of iterations of Newton's method depends both on the condition number and the Lipschitz constant of the Hessian, both of which increase as $\alpha \to 0$. Thus, in choosing α, we trade an improvement in performance guarantees for an increase in runtime.

5.3 Maxent with ℓ_2 Regularization

In some cases, tighter performance guarantees are obtained by using confidence regions which take the shape of a Euclidean ball. More specifically, we consider the potential and conjugate

$$\mathrm{U}_{\tilde{\pi}}^{(\sqrt{2})}(\mathbf{u}) = \begin{cases} 0 & \text{if } \|\mathbf{u}\|_2 \le \beta \\ \infty & \text{otherwise} \end{cases} \quad , \qquad \mathrm{U}_{\tilde{\pi}}^{(\sqrt{2})*}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_2 \quad .$$

We first derive an ℓ_2 version of Hoeffding's inequality (Lemma 10 below, proved in Appendix C) and then use Theorem 4 to obtain performance guarantees.

Lemma 10. *Let $D_2 = \sup_{x,x' \in \mathcal{X}} \|\mathbf{f}(x) - \mathbf{f}(x')\|_2$ be the ℓ_2 diameter of $\mathbf{f}(\mathcal{X})$ and let $\delta > 0$. Then with probability at least $1-\delta$*

$$\|\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}]\|_2 \le \frac{D_2}{\sqrt{2m}}\left[1 + \sqrt{\ln(1/\delta)}\right] \quad .$$

Theorem 11. *Let D_2 be the ℓ_2 diameter of $\mathbf{f}(\mathcal{X})$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\|\boldsymbol{\lambda}\|_2$ with $\beta = D_2\left[1 + \sqrt{\ln(1/\delta)}\right]/\sqrt{2m}$. Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\boldsymbol{\lambda}^\star\|_2 D_2}{\sqrt{m}}\left(\sqrt{2} + \sqrt{2\ln(1/\delta)}\right) \quad .$$

Unlike results of the previous sections, this bound does not explicitly depend on the number of features and only grows with the ℓ_2 diameter of the feature space. The ℓ_2 diameter is small for example when the feature space consists of sparse binary vectors.

An analogous bound can also be obtained for ℓ_1-regularized maxent in terms of the ℓ_∞ diameter of the feature space (relaxing the requirement of Corollary 5 that features be bounded in $[0,1]$):

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\boldsymbol{\lambda}^\star\|_1 D_\infty}{\sqrt{m}}\sqrt{2\ln(2n/\delta)} \quad .$$

This bound increases with the ℓ_∞ diameter of the feature space and also grows slowly with the number of features. It provides some insight for when we expect ℓ_1 regularization to perform better than ℓ_2 regularization. For example, consider a scenario when the total number of features is large, but the best approximation of π can be derived from a small number of relevant features. Increasing the number of irrelevant features, we may keep $\|\boldsymbol{\lambda}^\star\|_1$, $\|\boldsymbol{\lambda}^\star\|_2$ and D_∞ fixed while increasing D_2 as $\Omega(\sqrt{n})$. The guarantee for ℓ_2-regularized maxent then grows as $\Omega(\sqrt{n})$ while the guarantee for ℓ_1-regularized maxent grows only as $\Omega(\sqrt{\ln n})$. Note, however, that in practice the distribution returned by ℓ_2-regularized maxent may perform better than indicated by this guarantee. For a comparison of ℓ_1 and ℓ_2^2 regularization in the context of logistic regression see Ng (2004).

5.4 Maxent with ℓ_2^2 Regularization

So far we have considered potentials that take the form of an indicator function or its smooth approximation. In this section we present a result for the ℓ_2^2 potential $\mathrm{U}_{\tilde{\pi}}^{(2)}$ of Section 4 and the corresponding conjugate $\mathrm{U}_{\tilde{\pi}}^{(2)*}$:

$$\mathrm{U}_{\tilde{\pi}}^{(2)}(\mathbf{u}) = \frac{\|\mathbf{u}\|_2^2}{2\alpha} \quad , \qquad \mathrm{U}_{\tilde{\pi}}^{(2)*}(\boldsymbol{\lambda}) = \frac{\alpha\|\boldsymbol{\lambda}\|_2^2}{2} \quad .$$

The potential $\mathrm{U}_{\tilde{\pi}}^{(2)}$ grows continuously with an increasing distance from empirical averages while the conjugate $\mathrm{U}_{\tilde{\pi}}^{(2)*}$ corresponds to ℓ_2^2 regularization.

In the case of ℓ_2^2-regularized maxent it is possible to derive guarantees on the expected performance in addition to probabilistic guarantees. However, these guarantees require an *a priori* bound on $\|\boldsymbol{\lambda}^\star\|_2$ and thus are not entirely uniform. Our expectation guarantees are analogous to those derived by Zhang (2005) for the conditional case. However, we are able to obtain a better multiplicative constant.

Note that we could derive expectation guarantees by simply applying Lemma 3.13 and taking the expectation over a random sample:

$$\begin{aligned} \mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2^2}{\alpha} + \alpha\|\boldsymbol{\lambda}^\star\|_2^2 \\ \mathrm{E}\big[\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})\big] &\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\operatorname{tr}\boldsymbol{\Sigma}}{\alpha m} + \alpha\|\boldsymbol{\lambda}^\star\|_2^2 \ . \end{aligned} \tag{19}$$

Here, $\boldsymbol{\Sigma}$ is the covariance matrix of features with respect to π and tr denotes the trace of a matrix. We improve this guarantee by using Lemma 3.14 with $q_{\boldsymbol{\lambda}^\star}$ chosen to minimize $\mathrm{L}_\pi(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^{(2)*}(\boldsymbol{\lambda})$, and explicitly bounding $(\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}) \cdot (\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}])$ using a stability result similarly to Zhang (2005).

Lemma 12. *Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ where $\alpha > 0$. Then for every $q_{\boldsymbol{\lambda}^\star}$*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2^2}{\alpha} + \frac{\alpha\|\boldsymbol{\lambda}^\star\|_2^2}{2} \ .$$

Proof of Lemma 12 is given in Appendix D. Lemma 12 improves on (19) in the leading constant of $\|\boldsymbol{\lambda}^\star\|_2^2$ which is $\alpha/2$ instead of α. Taking the expectation over a random sample and bounding the trace of $\boldsymbol{\Sigma}$ in terms of the ℓ_2 diameter (see Lemma 22 of Appendix C), we obtain an expectation guarantee. We can also use Lemma 10 to bound $\|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2^2$ with high probability, and obtain a probabilistic guarantee. The two results are presented in Theorem 13 with the tradeoff between the guarantees controlled by the parameter s.

Theorem 13. *Let D_2 be the ℓ_2 diameter of $\mathbf{f}(\mathcal{X})$ and let $L_2, s > 0$. Let $\hat{\boldsymbol{\lambda}}$ minimize the ℓ_2^2-regularized log loss $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ with $\alpha = sD_2/(L_2\sqrt{m})$. Then*

$$\mathrm{E}\big[\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})\big] \leq \inf_{\|\boldsymbol{\lambda}^\star\|_2 \leq L_2} \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{L_2 D_2}{\sqrt{m}} \cdot \frac{s + s^{-1}}{2}$$

and if $\delta > 0$ then with probability at least $1 - \delta$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \inf_{\|\boldsymbol{\lambda}^\star\|_2 \leq L_2} \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{L_2 D_2}{\sqrt{m}} \cdot \frac{s + s^{-1}\big(1 + \sqrt{\ln(1/\delta)}\big)^2}{2} \ .$$

The bounds of Theorem 13 have properties similar to probabilistic guarantees of ℓ_2-regularized maxent. As mentioned earlier, they differ in the crucial fact that the norm $\|\boldsymbol{\lambda}^\star\|_2$ needs to be bounded *a priori* by a constant L_2. It is this constant rather than a possibly smaller norm $\|\boldsymbol{\lambda}^\star\|_2$ that enters the bound.

Note that bounds of this section generalize to arbitrary quadratic potentials $\mathrm{U}_{\tilde{\pi}}(\mathbf{u}) = \mathbf{u}^\top\mathbf{A}^{-1}\mathbf{u}/2$ and respective conjugates $\mathrm{U}_{\tilde{\pi}}^*(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^\top\mathbf{A}\boldsymbol{\lambda}/2$ where $\mathbf{A}$ is a symmetric positive definite matrix. Applying the transformation

$$\mathbf{f}'(x) = \alpha^{1/2}\mathbf{A}^{-1/2}\mathbf{f}(x) \ , \qquad \boldsymbol{\lambda}' = \alpha^{-1/2}\mathbf{A}^{1/2}\boldsymbol{\lambda}$$

where $\mathbf{A}^{1/2}$ is the unique symmetric positive definite matrix such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$, the guarantees for quadratic-regularized maxent in terms of $\mathbf{f}(\mathcal{X})$ and $\boldsymbol{\lambda}^\star$ reduce to the guarantees for ℓ_2^2-regularized maxent in terms of $\mathbf{f}'(\mathcal{X})$ and $\boldsymbol{\lambda}^{\star\prime}$.

5.5 Maxent with ℓ_2 Regularization versus ℓ_2^2 Regularization

In the previous two sections we have seen that performance guarantees for maxent with ℓ_2 and ℓ_2^2 regularization differ whenever we require that β and α be fixed before running the algorithm. We now show that if all possible values of β and α are considered then the sets of models generated by the two maxent versions are the same.

Let $\Lambda^{(\sqrt{2}),\beta}$ and $\Lambda^{(2),\alpha}$ denote the respective solution sets for maxent with ℓ_2 and ℓ_2^2 regularization:

$$\Lambda^{(\sqrt{2}),\beta} = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\|\boldsymbol{\lambda}\|_2\right] \tag{20}$$

$$\Lambda^{(2),\alpha} = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^n} \left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2\right] \quad . \tag{21}$$

If $\beta,\alpha > 0$ then $\Lambda^{(\sqrt{2}),\beta}$ and $\Lambda^{(2),\alpha}$ are non-empty because the objectives are lower semi-continuous and approach infinity as $\|\boldsymbol{\lambda}\|_2$ increases. For $\beta = 0$ and $\alpha = 0$, Equations (20) and (21) reduce to the basic maxent. Thus, $\Lambda^{(\sqrt{2}),0}$ and $\Lambda^{(2),0}$ contain the $\boldsymbol{\lambda}$'s for which $q_{\boldsymbol{\lambda}}[\mathbf{f}] = \tilde{\pi}[\mathbf{f}]$. This set will be empty if the basic maxent solutions are attained only in a limit.

Theorem 14. *Let* $\Lambda^{(\sqrt{2})} = \bigcup_{\beta\in[0,\infty]} \Lambda^{(\sqrt{2}),\beta}$ *and* $\Lambda^{(2)} = \bigcup_{\alpha\in[0,\infty]} \Lambda^{(2),\alpha}$. *Then* $\Lambda^{(\sqrt{2})} = \Lambda^{(2)}$.

Proof. First note that $\Lambda^{(\sqrt{2}),\infty} = \Lambda^{(2),\infty} = \{\mathbf{0}\}$. Next, we will show that $\Lambda^{(\sqrt{2})} \setminus \{\mathbf{0}\} = \Lambda^{(2)} \setminus \{\mathbf{0}\}$. Taking derivatives in Equations (20) and (21), we obtain that $\boldsymbol{\lambda} \in \Lambda^{(\sqrt{2}),\beta} \setminus \{\mathbf{0}\}$ if and only if

$$\boldsymbol{\lambda} \neq \mathbf{0} \quad \text{and} \quad \nabla \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_2 = 0 \quad .$$

Similarly, $\boldsymbol{\lambda} \in \Lambda^{(2),\alpha} \setminus \{\mathbf{0}\}$ if and only if

$$\boldsymbol{\lambda} \neq \mathbf{0} \quad \text{and} \quad \nabla \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\boldsymbol{\lambda} = 0 \quad .$$

Thus, any $\boldsymbol{\lambda} \in \Lambda^{(\sqrt{2}),\beta} \setminus \{\mathbf{0}\}$ is also in the set $\Lambda^{(2),\beta/\|\boldsymbol{\lambda}\|_2} \setminus \{\mathbf{0}\}$, and conversely any $\boldsymbol{\lambda} \in \Lambda^{(2),\alpha} \setminus \{\mathbf{0}\}$ is also in the set $\Lambda^{(\sqrt{2}),\alpha\|\boldsymbol{\lambda}\|_2} \setminus \{\mathbf{0}\}$. ■

The proof of Theorem 14 rests on the fact that the contours of regularization functions $\|\boldsymbol{\lambda}\|_2$ and $\|\boldsymbol{\lambda}\|_2^2$ coincide. We could easily extend the proof to include the equivalence of $\Lambda^{(\sqrt{2})}$, $\Lambda^{(2)}$ with the set of solutions to $\min\{\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) : \|\boldsymbol{\lambda}\|_2 \leq 1/\gamma\}$ where $\gamma \in [0,\infty]$. Similarly, one could show the equivalence of the solutions for regularizations $\beta\|\boldsymbol{\lambda}\|_1$, $\alpha\|\boldsymbol{\lambda}\|_1^2/2$ and $\mathrm{I}(\|\boldsymbol{\lambda}\|_1 \leq 1/\gamma)$.

The main implication of Theorem 14 is for maxent density estimation with model selection, for example, by minimization of held-out or cross-validated empirical error. In those cases, maxent versions with ℓ_2, ℓ_2^2 (and an ℓ_2-ball indicator) regularization yield the same solution. Thus, we prefer to use the computationally least intensive method. This will typically be ℓ_2^2-regularized maxent whose potential and regularization are smooth.

The solution sets $\Lambda^{(\sqrt{2}),\beta}$ and $\Lambda^{(2),\alpha}$ differ in their "sparsity" properties. We put the sparsity inside quotation marks because there are only two sparsity levels for ℓ_2 regularization: either all coordinates of $\boldsymbol{\lambda}$ remain zero under perturbations, or none of them. This is because the sole discontinuity of the gradient of the ℓ_2-regularized log loss is at $\boldsymbol{\lambda} = \mathbf{0}$. On the other hand, ℓ_2^2 regularization is smooth and therefore does not induce sparsity.

5.6 Maxent with $\ell_1+\ell_2^2$ Regularization

In this section, we consider regularization that has both ℓ_1-style and ℓ_2^2-style terms. To simplify the discussion, we do not distinguish between coordinates and use a weighted sum of the ℓ_1 norm and the square of the ℓ_2 norm:

$$\mathrm{U}_{\tilde{\pi}}^{(1+2)*}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_1 + \frac{\alpha\|\boldsymbol{\lambda}\|_2^2}{2} \quad , \qquad \mathrm{U}_{\tilde{\pi}}^{(1+2)}(\mathbf{u}) = \sum_j \frac{\left||u_j|-\beta\right|_+^2}{2\alpha} \quad .$$

Here α and β are positive constants, and $|x|_+ = \max\{0,x\}$ denotes the positive part of x. For a derivation of $\mathrm{U}_{\tilde{\pi}}^{(1+2)}$ see Appendix E.

For this type of regularization we are able to prove both probabilistic and expectation guarantees. Using similar techniques as in the previous sections we can derive, for example, the following theorem.

Theorem 15. *Let D_2, D_∞ be the ℓ_2 and ℓ_∞ diameters of $\mathbf{f}(\mathcal{X})$ respectively. Let $\delta, L_2 > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\|\boldsymbol{\lambda}\|_1 + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ with $\alpha = (D_2\min\{1/\sqrt{2}, \sqrt{m\delta}\})/(2L_2\sqrt{m})$ and $\beta = D_\infty\sqrt{\ln(2n/\delta)/(2m)}$. Then*

$$\mathrm{E}\left[\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})\right] \le \inf_{\|\boldsymbol{\lambda}^\star\|_2 \le L_2} \left[\mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{D_\infty\|\boldsymbol{\lambda}^\star\|_1}{\sqrt{m}}\sqrt{2\ln(2n/\delta)}\right] + \frac{D_2 L_2}{\sqrt{m}} \cdot \min\left\{\frac{1}{\sqrt{2}}, \sqrt{m\delta}\right\}$$

and with probability at least $1-\delta$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \inf_{\|\boldsymbol{\lambda}^\star\|_2 \le L_2} \left[\mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{D_\infty\|\boldsymbol{\lambda}^\star\|_1}{\sqrt{m}}\sqrt{2\ln(2n/\delta)}\right] + \frac{D_2 L_2}{\sqrt{m}} \cdot \frac{1}{2}\min\left\{\frac{1}{\sqrt{2}}, \sqrt{m\delta}\right\} \quad .$$

Proof. We only need to bound $\mathrm{U}_{\tilde{\pi}}^{(1+2)}(\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}])$ and its expectation and use Lemma 3.14. By Hoeffding's inequality and the union bound, the potential is zero with probability at least $1-\delta$, immediately yielding the second claim. Otherwise,

$$\mathrm{U}_{\tilde{\pi}}^{(1+2)}(\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}]) \le \frac{\|\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}]\|_2^2}{2\alpha} \le \frac{D_2^2}{2\alpha}$$

hence $\mathrm{E}\left[\mathrm{U}_{\tilde{\pi}}^{(1+2)}(\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}])\right] \le (\delta D_2^2)/(2\alpha)$. On the other hand, we can bound the trace of the feature covariance matrix by Lemma 22 of Appendix C and obtain

$$\mathrm{E}\left[\mathrm{U}_{\tilde{\pi}}^{(1+2)}(\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}])\right] \le \frac{\mathrm{E}\left[\|\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}]\|_2^2\right]}{2\alpha} = \frac{\mathrm{tr}\,\boldsymbol{\Sigma}}{2m\alpha} \le \frac{D_2^2}{4m\alpha} \quad .$$

Hence

$$\mathrm{E}\left[\mathrm{U}_{\tilde{\pi}}^{(1+2)}(\tilde{\pi}[\mathbf{f}] - \pi[\mathbf{f}])\right] \le \frac{D_2^2}{2m\alpha} \cdot \min\left\{\frac{1}{2}, m\delta\right\}$$

and the first claim follows. ∎

Setting $\delta = s/m$, we bound the difference in performance between the maxent distribution and any Gibbs distribution of a bounded weight vector by $O\left(\left(D_\infty\|\boldsymbol{\lambda}^\star\|_1\sqrt{\ln(2mn/s)} + D_2L_2\sqrt{s}\right)/\sqrt{m}\right)$. Now the constant s can be tuned to achieve the optimal tradeoff between $D_\infty\|\boldsymbol{\lambda}^\star\|_1$ and D_2L_2. Notice that the sparsity inducing properties of ℓ_1 regularization are preserved in $\ell_1+\ell_2^2$ regularization, because the partial derivatives of $\beta\|\boldsymbol{\lambda}\|_1 + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ are discontinuous at zero.

5.7 Extensions to Other Regularization Types

Regularizations explored in previous sections are derived from ℓ_1 and ℓ_2 norms. However, performance guarantees easily extend to arbitrary norms using the corresponding concentration bounds and conjugacy relationships in the spirit of (Altun and Smola, 2006).

For instance, for an arbitrary norm $\|\cdot\|_{\mathcal{B}^*}$, the regularization function $\beta\|\boldsymbol{\lambda}\|_{\mathcal{B}^*}$ corresponds to the potential $\mathrm{I}_B(\mathbf{u})$ where $B = \{\|\mathbf{u}\|_{\mathcal{B}} \leq \beta\}$ and $\|\cdot\|_{\mathcal{B}}$ is the dual norm of $\|\cdot\|_{\mathcal{B}^*}$. Similarly, for a norm $\|\cdot\|_{\mathcal{A}^*}$, the regularization function $\alpha\|\boldsymbol{\lambda}\|^2_{\mathcal{A}^*}/2$ corresponds to the potential $\|\mathbf{u}\|^2_{\mathcal{A}}/(2\alpha)$ where $\|\cdot\|_{\mathcal{A}}$ is the dual norm of $\|\cdot\|_{\mathcal{A}^*}$. For the combined regularization $\mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_{\mathcal{B}^*} + \alpha\|\boldsymbol{\lambda}\|^2_{\mathcal{A}^*}/2$, we can perform a similar analysis as for $\ell_1 + \ell_2^2$ regularization using the bound

$$\mathrm{U}_{\tilde{\pi}}(\mathbf{u}) \leq \min\{\mathrm{I}_B(\mathbf{u}), \|\mathbf{u}\|^2_{\mathcal{A}}/(2\alpha)\} \ .$$

Of course, the framework presented here is not limited to regularization functions derived from norms; however, the corresponding concentration bounds are typically less readily available.

6. Selective-Update Algorithm

In the previous section, we have discussed performance bounds of various types of regularization. Now we turn our attention to algorithms for solving generalized maxent problems. In the present and the following section, we propose two algorithms for generalized maxent with complete proofs of convergence. Our algorithms cover a wide class of potentials including the basic, box and ℓ_2^2 potential. The ℓ_2-ball potential $\mathrm{U}^{(\sqrt{2})}_{\tilde{\pi}}$ does not fall in this class, but we show that the corresponding maxent problem can be reduced and our algorithms can still be applied.

There are a number of algorithms for finding the basic maxent distribution, especially iterative scaling and its variants (Darroch and Ratcliff, 1972; Della Pietra et al., 1997). The **S**elective-**U**pdate algorithm for **M**aximu**M** **E**n**T**ropy (SUMMET) described in this section modifies one weight λ_j at a time, as explored by Collins, Schapire, and Singer (2002) in a similar setting. This style of coordinate-wise descent is convenient when working with a very large (or infinite) number of features. The original Darroch and Ratcliff algorithm also allows single-coordinate updates. Goodman (2002) observes that this leads to a much faster convergence than with the parallel version. However, updates are performed cyclically over all features, which renders the algorithm less practical with a large number of irrelevant features. Similarly, the sequential-update algorithm of Krishnapuram et al. (2005) requires a visitation schedule that updates each feature weight infinitely many times.

SUMMET differs since the weight to be updated is selected independently in each iteration. Thus, the features whose optimal weights are zero may never be updated. This approach is particularly useful in the context of ℓ_1-regularized maxent which often yields sparse solutions.

As explained in Section 4, the goal of the algorithm is to produce a sequence $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ maximizing the objective function Q in the limit. In this and the next section we assume that the potential U is *decomposable* as defined below:

Definition 16. A potential $\mathrm{U} : \mathbb{R}^n \to (-\infty, \infty]$ is called *decomposable* if it can be written as a sum of coordinate potentials $\mathrm{U}(\mathbf{u}) = \sum_j \mathrm{U}_j(u_j)$, each of which is a closed proper convex function bounded from below.

As a consequence of this definition, the conjugate potential U^* equals the sum of conjugate coordinate potentials U^*_j, by Equation (4), and $\mathrm{U}^*_j(0) = \sup_{u_j}[-\mathrm{U}_j(u_j)]$ is finite for all j.

Input: finite domain $\mathcal{X}$, default estimate q_0
features $f_1, \ldots, f_n$ where $f_j : \mathcal{X} \to [0,1]$, $f_j(\mathcal{X}) \neq \{0\}$ and $f_j(\mathcal{X}) \neq \{1\}$
decomposable potential U where $\operatorname{dom} \mathrm{U}_j \cap [0,1] \neq \{0\}$ and $\operatorname{dom} \mathrm{U}_j \cap [0,1] \neq \{1\}$

Output: $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ maximizing Q

Let $\boldsymbol{\lambda}_1 = \mathbf{0}$

For $t = 1, 2, \ldots$:

- let $(j^\star, \delta^\star) = \arg\max_{(j,\delta)} \Big[-\ln\Big(1 + (e^\delta - 1) q_{\boldsymbol{\lambda}}[f_j]\Big) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \Big]$
- $\lambda_{t+1,j} = \begin{cases} \lambda_{t,j^\star} + \delta^\star & \text{if } j = j^\star \\ \lambda_{t,j} & \text{otherwise} \end{cases}$

Figure 2: **S**elective-**U**pdate algorithm for **M**aximu**M** **E**n**T**ropy (SUMMET).

Throughout this section we assume that values of features f_j lie in the interval $[0,1]$ and that features and coordinate potentials are non-degenerate in the sense that ranges $f_j(\mathcal{X})$ and intersections $\operatorname{dom} \mathrm{U}_j \cap [0,1]$ differ from $\{0\}$ and $\{1\}$. In Appendix G we show that a generalized maxent problem with a decomposable potential can always be reduced to the non-degenerate form.

Our algorithm works by iteratively adjusting the single weight λ_j that maximizes (an approximation of) the change in Q. 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 the objective is

$$\begin{aligned} Q(\boldsymbol{\lambda}') - Q(\boldsymbol{\lambda}) &= -\ln Z_{\boldsymbol{\lambda}'} - \mathrm{U}^*(-\boldsymbol{\lambda}') + \ln Z_{\boldsymbol{\lambda}} + \mathrm{U}^*(-\boldsymbol{\lambda}) \\ &= -\ln(q_{\boldsymbol{\lambda}}[e^{\delta f_j}]) - \textstyle\sum_{j'} [\mathrm{U}_{j'}^*(-\lambda'_{j'}) - \mathrm{U}_{j'}^*(-\lambda_{j'})] && (22) \\ &\geq -\ln(q_{\boldsymbol{\lambda}}[1 + (e^\delta - 1) f_j]) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) && (23) \\ &= -\ln\big(1 + (e^\delta - 1) q_{\boldsymbol{\lambda}}[f_j]\big) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \ . && (24) \end{aligned}$$

Equation (22) uses

$$Z_{\boldsymbol{\lambda}'} = \sum_{x \in \mathcal{X}} q_0(x) e^{\boldsymbol{\lambda} \cdot \mathbf{f}(x) + \delta_j f_j(x)} = Z_{\boldsymbol{\lambda}} \sum_{x \in \mathcal{X}} q_{\boldsymbol{\lambda}}(x) e^{\delta_j f_j(x)} \ . \tag{25}$$

Equation (23) is because $e^{\delta x} \leq 1 + (e^\delta - 1)x$ for $x \in [0,1]$ by convexity.

Let $F_j(\boldsymbol{\lambda}, \delta)$ denote the expression in (24):

$$F_j(\boldsymbol{\lambda}, \delta) = -\ln\big(1 + (e^\delta - 1) q_{\boldsymbol{\lambda}}[f_j]\big) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \ .$$

Our algorithm, shown in Figure 2, on each iteration, maximizes this lower bound over all choices of (j, δ) and for the maximizing j adds the corresponding δ to λ_j. We assume that for each j the maximizing δ is finite. This will be the case if the potential and features are non-degenerate (see Appendix G). Note that $F_j(\boldsymbol{\lambda}, \delta)$ is strictly concave in δ so we can use any of a number of search methods to find the optimal δ.

Solving ℓ_1-Regularized Maxent. For maxent with box constraints (which subsumes the basic maxent), the optimizing δ can be derived explicitly. First note that

$$\begin{aligned} F_j^{(1)}(\boldsymbol{\lambda}, \delta) &= -\ln\big(1 + (e^\delta - 1) q_{\boldsymbol{\lambda}}[f_j]\big) - \mathrm{U}_j^{(1)*}(-\lambda_j - \delta) + \mathrm{U}_j^{(1)*}(-\lambda_j) \\ &= -\ln\big(1 + (e^\delta - 1) q_{\boldsymbol{\lambda}}[f_j]\big) + \delta \tilde{\pi}[f_j] - \beta_j(|\lambda_j + \delta| - |\lambda_j|) \end{aligned}$$

Input: finite domain $\mathcal{X}$, default estimate q_0
examples $x_1,\ldots,x_m \in \mathcal{X}$
features $f_1,\ldots,f_n$ where $f_j : \mathcal{X} \to [0,1]$, $f_j(\mathcal{X}) \neq \{0\}$ and $f_j(\mathcal{X}) \neq \{1\}$
non-negative regularization parameters $\beta_1,\ldots,\beta_n$ where $\beta_j > 0$ if $\tilde{\pi}[f_j] \in \{0,1\}$

Output: $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ minimizing $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$

Let $\boldsymbol{\lambda}_1 = \mathbf{0}$
For $t = 1, 2, \ldots$:

- $(j^\star, \delta^\star) = \arg\max_{(j,\delta)} \Big[-\ln\Big(1 + (e^\delta - 1) q_t[f_j]\Big) + \tilde{\pi}[f_j]\delta - \beta_j(|\lambda_{t,j} + \delta| - |\lambda_{t,j}|) \Big]$

  for each j it suffices to consider the following possibilities (whenever defined)

  $$\delta_+ = \ln\left(\frac{(\tilde{\pi}[f_j] - \beta_j)(1 - q_t[f_j])}{(1 - \tilde{\pi}[f_j] + \beta_j) q_t[f_j]}\right) \quad , \quad \delta_0 = -\lambda_{t,j} \quad , \quad \delta_- = \ln\left(\frac{(\tilde{\pi}[f_j] + \beta_j)(1 - q_t[f_j])}{(1 - \tilde{\pi}[f_j] - \beta_j) q_t[f_j]}\right)$$

  and choose δ_+ if $\lambda_{t,j} + \delta_+ > 0$, δ_- if $\lambda_{t,j} + \delta_- < 0$, and δ_0 otherwise

- $\lambda_{t+1,j} = \begin{cases} \lambda_{t,j^\star} + \delta^\star & \text{if } j = j^\star \\ \lambda_{t,j} & \text{otherwise} \end{cases}$

Figure 3: Selective-update algorithm for ℓ_1-regularized maxent (ℓ_1-SUMMET).

since

$$\mathrm{U}_j^{(1)*}(-\mu_j) = \mathrm{U}_{\tilde{\pi},j}^{(1)*}(\mu_j) - \mu_j \tilde{\pi}[f_j] = \beta_j |\mu_j| - \mu_j \tilde{\pi}[f_j] \quad .$$

The optimum δ can be obtained for each j 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. The complete algorithm, ℓ_1-SUMMET, is shown in Figure 3.

Solving ℓ_2-Regularized Maxent. The ℓ_2-ball potential $\mathrm{U}_{\tilde{\pi}}^{(\sqrt{2})}$ is not decomposable. In order to reduce ℓ_2-regularized maxent to maxent with a decomposable potential, we replace the constraint $\|\tilde{\pi}[\mathbf{f}] - p[\mathbf{f}]\|_2 \leq \beta$ by $\|\tilde{\pi}[\mathbf{f}] - p[\mathbf{f}]\|_2^2 \leq \beta^2$ which yields an equivalent primal:

$$\mathcal{P}' : \quad \min_{p \in \Delta} \mathrm{D}(p \parallel q_0) \text{ subject to } \|\tilde{\pi}[\mathbf{f}] - p[\mathbf{f}]\|_2^2 \leq \beta^2 \quad .$$

If $\beta > 0$ then, by the Lagrange duality and Slater's conditions (Boyd and Vandenberghe, 2004), there exists $\mu \geq 0$ such that the solution of $\mathcal{P}'$ is the same as the solution of

$$\mathcal{P}'' : \quad \min_{p \in \Delta} \big[\mathrm{D}(p \parallel q_0) + \mu \left(\|\tilde{\pi}[\mathbf{f}] - p[\mathbf{f}]\|_2^2 - \beta^2 \right) \big] \quad .$$

The sought-after μ is the one which maximizes the value of $\mathcal{P}''$. Since the value of $\mathcal{P}''$ is concave in μ, we can employ a range of search techniques to find the optimal μ, using SUMMET (or PLUMMET of the next section) with ℓ_2^2 regularization in each iteration.

6.1 Convergence

In order to prove convergence of SUMMET, we will measure its progress towards solving the primal and dual. One measure of progress is the difference between the primal evaluated at $q_{\boldsymbol{\lambda}}$ and the dual evaluated at $\boldsymbol{\lambda}$:

$$\begin{aligned} P(q_{\boldsymbol{\lambda}}) - Q(\boldsymbol{\lambda}) &= [\mathrm{D}(q_{\boldsymbol{\lambda}} \parallel q_0) + \mathrm{U}(q_{\boldsymbol{\lambda}}[\mathbf{f}])] - [-\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}^*(-\boldsymbol{\lambda})] \\ &= q_{\boldsymbol{\lambda}}[\boldsymbol{\lambda} \cdot \mathbf{f} - \ln Z_{\boldsymbol{\lambda}}] + \mathrm{U}(q_{\boldsymbol{\lambda}}[\mathbf{f}]) + q_{\boldsymbol{\lambda}}[\ln Z_{\boldsymbol{\lambda}}] + \mathrm{U}^*(-\boldsymbol{\lambda}) \\ &= \mathrm{U}(q_{\boldsymbol{\lambda}}[\mathbf{f}]) + \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot q_{\boldsymbol{\lambda}}[\mathbf{f}] \quad . \end{aligned}$$

By Theorem 1, this difference is non-negative and equals zero exactly when $q_{\boldsymbol{\lambda}}$ solves primal and $\boldsymbol{\lambda}$ solves the dual.

For a decomposable potential, Fenchel's inequality in each coordinate implies that the difference is zero exactly when

$$\mathrm{U}_j(q_{\boldsymbol{\lambda}}[f_j]) + \mathrm{U}_j^*(-\lambda_j) + \lambda_j q_{\boldsymbol{\lambda}}[f_j] = 0$$

for all j. This characterization corresponds to the Kuhn-Tucker conditions (Rockafellar, 1970) in the case when coordinate potentials express equality and inequality constraints.

For many potentials of interest, including equality and inequality constraints, the difference between primal and dual may remain infinite throughout the computation. Therefore, we propose to use an *auxiliary function* as a surrogate for this difference. The auxiliary function is defined, somewhat non-standardly, as follows:

Definition 17. A function $A : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, \infty]$ is called an *auxiliary function* if

$$A(\boldsymbol{\lambda}, \mathbf{a}) = \mathrm{U}(\mathbf{a}) + \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot \mathbf{a} + \mathrm{B}(\mathbf{a} \parallel q_{\boldsymbol{\lambda}}[\mathbf{f}])$$

where $\mathrm{B}(\cdot \parallel \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, \infty]$ satisfies conditions (B1) and (B2).

The interpretation of an auxiliary function as a surrogate for the difference between primal and dual objectives is novel. Unlike previous applications of auxiliary functions (Della Pietra et al., 1997, 2001; Collins et al., 2002), we do not assume that $A(\boldsymbol{\lambda}, \mathbf{a})$ bounds a change in the dual objective and we also make no continuity assumptions. The reason for the former is technical: we need to allow a more flexible relationship between A and a change in the dual objective to accommodate algorithms both with single-coordinate and parallel updates. The absence of continuity assumptions is, however, crucial in order to allow arbitrary (decomposable) potentials. The continuity is replaced by the property (B2). On the other hand, our form of auxiliary function is more restrictive as the only flexibility is in choosing B, which is a function of $q_{\boldsymbol{\lambda}}[\mathbf{f}]$ rather than $q_{\boldsymbol{\lambda}}$.

The auxiliary function is always non-negative since $\mathrm{U}(\mathbf{a}) + \mathrm{U}^*(-\boldsymbol{\lambda}) \geq -\boldsymbol{\lambda} \cdot \mathbf{a}$ by Fenchel's inequality and hence $A(\boldsymbol{\lambda}, \mathbf{a}) \geq \mathrm{B}(\mathbf{a} \parallel q_{\boldsymbol{\lambda}}[\mathbf{f}]) \geq 0$. Moreover, if $A(\boldsymbol{\lambda}, \mathbf{a}) = 0$ then $q_{\boldsymbol{\lambda}}[\mathbf{f}] = \mathbf{a}$ and $A(\boldsymbol{\lambda}, \mathbf{a}) = P(q_{\boldsymbol{\lambda}}) - Q(\boldsymbol{\lambda}) = 0$, that is, by maxent duality, $q_{\boldsymbol{\lambda}}$ solves the primal and $\boldsymbol{\lambda}$ solves the dual.

It turns out, as we show in Lemma 19 below, that the optimality property generalizes to the case when $A(\boldsymbol{\lambda}_t, \mathbf{a}_t) \to 0$ provided that $Q(\boldsymbol{\lambda}_t)$ has a finite limit. In particular, it suffices to find a suitable sequence of $\mathbf{a}_t$'s for $\boldsymbol{\lambda}_t$'s produced by an algorithm to show its convergence. Note that the optimality in the limit trivially holds when $\boldsymbol{\lambda}_t$'s and $\mathbf{a}_t$'s come from a compact set, because $A(\hat{\boldsymbol{\lambda}}, \hat{\mathbf{a}}) = 0$ at a cluster point of $\{(\boldsymbol{\lambda}_t, \mathbf{a}_t)\}$ by the lower semi-continuity of U and U^*.

In a general case, we follow the technique used by Della Pietra, Della Pietra, and Lafferty (1997) for the basic maxent: we consider a cluster point $\hat{q}$ of $\{q_{\boldsymbol{\lambda}_t}\}$ and show that (i) $\hat{q}$ is primal

feasible and (ii) the difference $P(\hat{q}) - Q(\boldsymbol{\lambda}_t)$ approaches zero. In the case of the basic maxent, $A(\boldsymbol{\lambda}, \mathbf{a}) = \mathrm{B}(\tilde{\pi}[\mathbf{f}] \parallel q_{\boldsymbol{\lambda}}[\mathbf{f}])$ whenever finite. Thus, (i) is obtained by (B2), and noting that $P(\hat{q}) - Q(\boldsymbol{\lambda}) = \mathrm{D}(\hat{q} \parallel q_{\boldsymbol{\lambda}})$ yields (ii). For a general potential, however, claims (i) and (ii) seem to require a novel approach. In both steps, we use decomposability and the technical Lemma 18 (proved in Appendix F). Thus compactness or decomposability seem to be crucial in the present approach.

Lemma 18. *Let U_r be a decomposable potential relative to a feasible point r. Let $S = \operatorname{dom} \mathrm{U}_r = \{\mathbf{u} \in \mathbb{R}^n : \mathrm{U}_r(\mathbf{u}) < \infty\}$ and $T_c = \{\boldsymbol{\lambda} \in \mathbb{R}^n : \mathrm{U}_r^*(\boldsymbol{\lambda}) \leq c\}$. Then there exists $\alpha_c \geq 0$ such that $\boldsymbol{\lambda} \cdot \mathbf{u} \leq \alpha_c \|\mathbf{u}\|_1$ for all $\mathbf{u} \in S, \boldsymbol{\lambda} \in T_c$.*

Lemma 19. *Let $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots \in \mathbb{R}^n$, $\mathbf{a}_1, \mathbf{a}_2, \ldots \in \mathbb{R}^n$ be sequences such that $Q(\boldsymbol{\lambda}_t)$ has a finite limit and $A(\boldsymbol{\lambda}_t, \mathbf{a}_t) \to 0$ as $t \to \infty$. Then $\lim_{t\to\infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda})$.*

Proof. Let q_t denote $q_{\boldsymbol{\lambda}_t}$. Distributions q_t come from the compact set Δ, so we can choose a convergent subsequence. We index this subsequence by τ and denote its limit by $\hat{q}$. We assume that the subsequence was chosen in such a manner that values $A(\boldsymbol{\lambda}_\tau, \mathbf{a}_\tau)$ and $Q(\boldsymbol{\lambda}_\tau)$ are finite. We do this without loss of generality because limits of $A(\boldsymbol{\lambda}_\tau, \mathbf{a}_\tau)$ and $Q(\boldsymbol{\lambda}_\tau)$ are finite. We will show that $\lim_{\tau\to\infty} Q(\boldsymbol{\lambda}_\tau) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda})$. The result will then follow since $\lim_{\tau\to\infty} Q(\boldsymbol{\lambda}_\tau) = \lim_{t\to\infty} Q(\boldsymbol{\lambda}_t)$.

As noted earlier, $A(\boldsymbol{\lambda}, \mathbf{a}) \geq \mathrm{B}(\mathbf{a} \parallel q_{\boldsymbol{\lambda}}[\mathbf{f}])$. Since $\mathrm{B}(\mathbf{a}_\tau \parallel q_\tau[\mathbf{f}])$ is non-negative and $A(\boldsymbol{\lambda}_\tau, \mathbf{a}_\tau) \to 0$, we obtain $\mathrm{B}(\mathbf{a}_\tau \parallel q_\tau[\mathbf{f}]) \to 0$. Thus, $\mathbf{a}_\tau \to \hat{q}[\mathbf{f}]$ by the property (B2). Rewriting A in terms of the potential and the conjugate potential relative to an arbitrary feasible point r (which exists by assumption), we obtain

$$A(\boldsymbol{\lambda}_\tau, \mathbf{a}_\tau) = \mathrm{U}_r(r[\mathbf{f}] - \mathbf{a}_\tau) + \mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) - \boldsymbol{\lambda}_\tau \cdot (r[\mathbf{f}] - \mathbf{a}_\tau) + \mathrm{B}(\mathbf{a}_\tau \parallel q_\tau[\mathbf{f}]) \ . \tag{26}$$

Rearrange terms, noting that $A(\boldsymbol{\lambda}_\tau, \mathbf{a}_\tau) \to 0$ and $\mathrm{B}(\mathbf{a}_\tau \parallel q_\tau[\mathbf{f}]) \to 0$:

$$\mathrm{U}_r(r[\mathbf{f}] - \mathbf{a}_\tau) = -\mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) + \boldsymbol{\lambda}_\tau \cdot (r[\mathbf{f}] - \mathbf{a}_\tau) + o(1) \ . \tag{27}$$

We use Equation (27) to prove first the feasibility and then the optimality of $\hat{q}$.

Feasibility. We bound the right hand site of Equation (27) and take limits to show that $\mathrm{U}_r(r[\mathbf{f}] - \hat{q}[\mathbf{f}])$ is also finite. The first term is bounded by Fenchel's inequality:

$$-\mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) \leq -\boldsymbol{\lambda}_\tau \cdot \mathbf{0} + \mathrm{U}_r(\mathbf{0}) = \mathrm{U}_r(\mathbf{0}) \ , \tag{28}$$

which is finite by the feasibility of r. In order to bound $\boldsymbol{\lambda}_\tau \cdot (r[\mathbf{f}] - \mathbf{a}_\tau)$, the second term of Equation (27), we use Lemma 18. First note that $r[\mathbf{f}] - \mathbf{a}_\tau$ is a feasible point of U_r for all τ by Equation (26) and the finiteness of $A(\boldsymbol{\lambda}_\tau, \mathbf{a}_\tau)$. Next, from Equation (11):

$$\mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) = -Q(\boldsymbol{\lambda}_\tau) - \mathrm{D}(r \parallel q_\tau) + \mathrm{D}(r \parallel q_0) \ ,$$

which is bounded above by some constant c independent of τ because $-Q(\boldsymbol{\lambda}_\tau)$ has a finite limit and is thus bounded above, $-\mathrm{D}(r \parallel q_\tau)$ is non-positive, and $\mathrm{D}(r \parallel q_0)$ is a finite constant. Hence by Lemma 18

$$\boldsymbol{\lambda}_\tau \cdot (r[\mathbf{f}] - \mathbf{a}_\tau) \leq \alpha_r \|r[\mathbf{f}] - \mathbf{a}_\tau\|_1 \tag{29}$$

for some constant α_r independent of τ. Plugging Equations (28) and (29) in Equation (27) and taking limits, we obtain by lower semi-continuity of U_r

$$\mathrm{U}_r(r[\mathbf{f}] - \hat{q}[\mathbf{f}]) \leq \mathrm{U}_r(\mathbf{0}) + \alpha_r \|r[\mathbf{f}] - \hat{q}[\mathbf{f}]\|_1 \ .$$

Thus $\hat{q}$ is primal feasible.

Optimality. Since the foregoing holds for any primal feasible r, we can set $r = \hat{q}$ and obtain

$$\begin{aligned} \mathrm{U}_{\hat{q}}(\hat{q}[\mathbf{f}] - \mathbf{a}_\tau) &= -\mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau) + \boldsymbol{\lambda}_\tau \cdot (\hat{q}[\mathbf{f}] - \mathbf{a}_\tau) + o(1) && (30) \\ &\leq -\mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau) + \alpha_{\hat{q}} \|\hat{q}[\mathbf{f}] - \mathbf{a}_\tau\|_1 + o(1) \ . && (31) \end{aligned}$$

Equation (30) follows from Equation (27). Equation (31) follows from Equation (29). Taking limits, we obtain

$$\mathrm{U}_{\hat{q}}(\mathbf{0}) \leq \lim_{\tau \to \infty} \left[-\mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau) \right] \ . \tag{32}$$

Now we are ready to show that $Q(\boldsymbol{\lambda}_\tau)$ maximizes the dual in the limit:

$$\begin{aligned} P(\hat{q}) &= \mathrm{D}(\hat{q} \parallel q_0) + \mathrm{U}_{\hat{q}}(\mathbf{0}) \\ &\leq \mathrm{D}(\hat{q} \parallel q_0) + \lim_{\tau \to \infty} \left[-\mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau) \right] && (33) \\ &= \lim_{\tau \to \infty} \left[\mathrm{D}(\hat{q} \parallel q_0) - \mathrm{D}(\hat{q} \parallel q_\tau) - \mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau) \right] && (34) \\ &= \lim_{\tau \to \infty} Q(\boldsymbol{\lambda}_\tau) \ . && (35) \end{aligned}$$

Equation (33) follows from Equation (32). Equation (34) follows from the continuity of relative entropy since $q_\tau \to \hat{q}$. Equation (35) follows from Equation (11). Finally, combining Equations (33–35), we obtain by maxent duality that $\hat{q}$ minimizes the primal and λ_τ maximizes the dual as $\tau \to \infty$. ■

Theorem 20. SUMMET *produces a sequence* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ *for which*

$$\lim_{t \to \infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}) \ .$$

Proof. It suffices to show that $Q(\boldsymbol{\lambda}_t)$ has a finite limit and present an auxiliary function A and a sequence $\mathbf{a}_1, \mathbf{a}_2, \ldots$ for which $A(\boldsymbol{\lambda}_t, \mathbf{a}_t) \to 0$.

Note that $Q(\boldsymbol{\lambda}_1) = Q(\mathbf{0}) = -\mathrm{U}^*(\mathbf{0})$ is finite by the decomposability of the potential, and Q is bounded above by the feasibility of the primal. Let $F_{t,j} = \max_\delta F_j(\boldsymbol{\lambda}_t, \delta)$. Note that $F_{t,j}$ is non-negative since $F_j(\boldsymbol{\lambda}_t, 0) = 0$. Since $F_{t,j}$ bounds change in the objective from below, the dual objective $Q(\boldsymbol{\lambda}_t)$ is non-decreasing and thus has a finite limit.

In each step

$$Q(\boldsymbol{\lambda}_{t+1}) - Q(\boldsymbol{\lambda}_t) \geq F_{t,j} \geq 0 \ .$$

Since Q has a finite limit, differences $Q(\boldsymbol{\lambda}_{t+1}) - Q(\boldsymbol{\lambda}_t)$ converge to zero and thus $F_{t,j} \to 0$. We use $F_{t,j}$ to define an auxiliary function. To begin, we rewrite $F_{t,j}$ using Fenchel's duality:

$$\begin{aligned} F_{t,j} &= \max_\delta \Big[-\ln(1 + (e^\delta - 1) q_t[f_j]) - \mathrm{U}^*_j(-\lambda_{t,j} - \delta) + \mathrm{U}^*_j(-\lambda_{t,j}) \Big] \\ &= \max_\delta \Big[-\ln \left\{ \big(1 - q_t[f_j]\big) e^{0 \cdot \delta} + q_t[f_j] e^{1 \cdot \delta} \right\} - \mathrm{U}'^*_j(-\delta) \Big] + \mathrm{U}^*_j(-\lambda_{t,j}) && (36) \\ &= \min_{a', a} \Big[\mathrm{D}\big((a', a) \parallel (1 - q_t[f_j], q_t[f_j])\big) + \mathrm{U}'_j(0 \cdot a' + 1 \cdot a) \Big] + \mathrm{U}^*_j(-\lambda_{t,j}) && (37) \\ &= \min_{0 \leq a \leq 1} \Big[\mathrm{D}(a \parallel q_t[f_j]) + \mathrm{U}_j(a) + \lambda_{t,j} \cdot a \Big] + \mathrm{U}^*_j(-\lambda_{t,j}) \ . && (38) \end{aligned}$$

Input: finite domain $\mathcal{X}$, default estimate q_0
features $f_1, \ldots, f_n$ where $f_j : \mathcal{X} \to [0,1]$, $f_j(\mathcal{X}) \neq \{0\}$ and $\sum_j f_j(x) \leq 1$ for all $x \in \mathcal{X}$
decomposable potential U where $\operatorname{dom} \mathrm{U}_j \cap [0,1] \neq \{0\}$

Output: $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ maximizing $Q(\boldsymbol{\lambda})$

Let $\boldsymbol{\lambda}_1 = \mathbf{0}$

For $t = 1, 2, \ldots$:

- for each j, let
  $\delta_j = \arg\max_{\delta} \left[-q_t[f_j](e^{\delta} - 1) - \mathrm{U}_j^*(-\lambda_{t,j} - \delta) + \mathrm{U}_j^*(-\lambda_{t,j}) \right]$
- update $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \boldsymbol{\delta}$

Figure 4: **P**aralle**L**-**U**pdate algorithm for **M**aximu**M** **E**n**T**ropy (PLUMMET).

In Equation (36), we rearranged terms inside the logarithm so they would take the form of a partition function. We write $\mathrm{U}_j'^*(u)$ for $\mathrm{U}_j^*(u - \lambda_{t,j})$. In Equation (37), we applied Theorem 1, noting that the conjugate of the log partition function is the relative entropy (see Section 3). The value of relative entropy $\mathrm{D}((a', a) \parallel (1 - q_t[f_j], q_t[f_j]))$ is infinite whenever (a', a) is not a probability distribution, so it suffices to consider pairs where $0 \leq a \leq 1$ and $a' = 1 - a$. In Equation (38), we use $\mathrm{D}(a \parallel q_t[f_j])$ as a shorthand for $\mathrm{D}((1-a, a) \parallel (1 - q_t[f_j], q_t[f_j]))$. We use Equation (3) to convert U_j' into U_j:

$$\mathrm{U}_j'(0 \cdot a' + 1 \cdot a) = \mathrm{U}_j'(a) = \mathrm{U}_j(a) + \lambda_{t,j} \cdot a \ .$$

The minimum in Equation (38) is always attained because a comes from a compact set and the minimized expression is lower semi-continuous in a. We use $a_{t,j}$ to denote a value attaining this minimum. Thus

$$F_{t,j} = \mathrm{U}_j(a_{t,j}) + \mathrm{U}_j^*(-\lambda_{t,j}) + \lambda_{t,j} a_{t,j} + \mathrm{D}(a_{t,j} \parallel q_t[f_j]) \ .$$

Note that $\mathrm{D}(a \parallel b)$ satisfies conditions (B1) and (B2) hence the sum $\mathrm{B}(\mathbf{a} \parallel \mathbf{b}) = \sum_j \mathrm{D}(a_j \parallel b_j)$ also satisfies (B1) and (B2). We use this to derive the auxiliary function

$$A(\boldsymbol{\lambda}, \mathbf{a}) = \textstyle\sum_j \left[\mathrm{U}_j(a_j) + \mathrm{U}_j^*(-\lambda_j) + \lambda_j a_j + \mathrm{D}(a_j \parallel q_{\boldsymbol{\lambda}}[f_j]) \right] \ .$$

Now $A(\boldsymbol{\lambda}_t, \mathbf{a}_t) = \sum_j F_{t,j} \to 0$, and the result follows by Lemma 19. ■

7. 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 maximize Q using an algorithm that updates all features simultaneously on every iteration. In this section, we describe a variant of generalized iterative scaling (Darroch and Ratcliff, 1972) applicable to generalized maxent with an arbitrary decomposable potential and prove its convergence. Note that gradient-based or Newton methods may be faster in practice.

Throughout this section, we make the assumption (without loss of generality) that, for all $x \in \mathcal{X}$, $f_j(x) \geq 0$ and $\sum_j f_j(x) \leq 1$ and features and coordinate potentials are non-degenerate in the sense that feature ranges $f_j(\mathcal{X})$ and intersections $\operatorname{dom} \mathrm{U}_j \cap [0,1]$ differ from $\{0\}$. Note that this differs from the notion of degeneracy in SUMMET.

Input: finite domain $\mathcal{X}$, default estimate q_0
examples $x_1, \ldots, x_m \in X$
features $f_1, \ldots, f_n$ where $f_j : \mathcal{X} \to [0,1]$, $f_j(\mathcal{X}) \neq \{0\}$ and $\sum_j f_j(x) \leq 1$ for all $x \in \mathcal{X}$
nonnegative regularization parameters $\beta_1, \ldots, \beta_n$ where $\beta_j > 0$ if $\tilde{\pi}[f_j] = 0$

Output: $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ minimizing $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$

Let $\boldsymbol{\lambda}_1 = \mathbf{0}$

For $t = 1, 2, \ldots$:

- for each j, let

  $$\delta_j = \arg\max_{\delta} \left[-q_t[f_j](e^{\delta} - 1) + \delta \tilde{\pi}[f_j] - \beta_j(|\lambda_{t,j} + \delta| - |\lambda_{t,j}|) \right]$$

  it suffices to consider the following possibilities (whenever defined)

  $$\delta_+ = \ln\left(\frac{\tilde{\pi}[f_j] - \beta_j}{q_t[f_j]}\right) \quad , \quad \delta_0 = -\lambda_{t,j} \quad , \quad \delta_- = \ln\left(\frac{\tilde{\pi}[f_j] + \beta_j}{q_t[f_j]}\right)$$

  and choose δ_+ if $\lambda_{t,j} + \delta_+ > 0$, δ_- if $\lambda_{t,j} + \delta_- < 0$, and δ_0 otherwise

- update $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \boldsymbol{\delta}$

Figure 5: Parallel-update algorithm for ℓ_1-regularized maxent (ℓ_1-PLUMMET).

Similarly to SUMMET, our **P**aralle**L**-**U**pdate algorithm for **M**aximu**M** **E**n**T**ropy (PLUMMET) is based on an approximation of the change in the objective function Q, in this case the following, where $\boldsymbol{\lambda}' = \boldsymbol{\lambda} + \boldsymbol{\delta}$:

$$\begin{aligned} Q(\boldsymbol{\lambda}') - Q(\boldsymbol{\lambda}) &= -\ln Z_{\boldsymbol{\lambda}'} - \mathrm{U}^*(-\boldsymbol{\lambda}') + \ln Z_{\boldsymbol{\lambda}} + \mathrm{U}^*(-\boldsymbol{\lambda}) \\ &= -\ln q_{\boldsymbol{\lambda}}\left[e^{\boldsymbol{\delta}\cdot\mathbf{f}}\right] - \mathrm{U}^*(-\boldsymbol{\lambda} - \boldsymbol{\delta}) + \mathrm{U}^*(-\boldsymbol{\lambda}) \qquad (39) \\ &\geq \sum_j \left[-q_{\boldsymbol{\lambda}}[f_j](e^{\delta_j} - 1) - \mathrm{U}_j^*(-\lambda_j - \delta_j) + \mathrm{U}_j^*(-\lambda_j) \right] \quad . \qquad (40) \end{aligned}$$

Equation (39) uses Equation (25). For Equation (40), note first that if $x_j \in \mathbb{R}$ and $p_j \geq 0$ with $\sum_j p_j \leq 1$ then

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

(See Collins, Schapire, and Singer, 2002, for a proof.) Thus,

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

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

PLUMMET, shown in Figure 4, on each iteration, maximizes Equation (40) over all choices of the δ_j's. For the basic potential $\mathrm{U}^{(0)}$, this algorithm reduces to generalized iterative scaling of Darroch and Ratcliff (1972). For ℓ_1-style regularization, the maximizing $\boldsymbol{\delta}$ can be calculated explicitly (see algorithm ℓ_1-PLUMMET in Figure 5). Again, it turns out that all the components of the maximizing $\boldsymbol{\delta}$ are finite as long as features and potentials are non-degenerate (see Appendix G). As before, we can prove the convergence of PLUMMET, and thus also of ℓ_1-PLUMMET.

Theorem 21. PLUMMET *produces a sequence* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ *for which*

$$\lim_{t\to\infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}) \quad .$$

Proof. The proof mostly follows the same lines as the proof of Theorem 20. Here we sketch the main differences.

Let q_t denote $q_{\boldsymbol{\lambda}_t}$ and F_t denote the lower bound on the change in the objective:

$$F_t = \sup_{\boldsymbol{\delta}} \sum_j \Big[-q_t[f_j](e^{\delta_j} - 1) - \mathrm{U}_j^*(-\lambda_{t,j} - \delta_j) + \mathrm{U}_j^*(-\lambda_{t,j}) \Big] \quad .$$

As before, $Q(\boldsymbol{\lambda}_t)$ has a finite limit and $F_t \to 0$. We can rewrite F_t using Fenchel's duality:

$$F_t = \sup_{\boldsymbol{\delta}} \sum_j \Big[-q_t[f_j](e^{\delta_j} - 1) - \mathrm{U}_j'^*(-\delta_j) \Big] + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \tag{41}$$

$$= \inf_{\mathbf{a}\geq\mathbf{0}} \sum_j \Big[\widetilde{\mathrm{D}}(a_j \parallel q_t[f_j]) + \mathrm{U}_j'(a_j) \Big] + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \tag{42}$$

$$= \inf_{\mathbf{a}\geq\mathbf{0}} \Big[\widetilde{\mathrm{D}}(\mathbf{a} \parallel q_t[\mathbf{f}]) + \mathrm{U}(\mathbf{a}) + \boldsymbol{\lambda}_t \cdot \mathbf{a} + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \Big] \quad . \tag{43}$$

In Equation (41) we write $\mathrm{U}_j'^*(u)$ for $\mathrm{U}_j^*(u - \lambda_{t,j})$. In Equation (42) we use Theorem 1, noting that the conjugate of $u_0(e^u - 1)$ is the unnormalized relative entropy. In Equation (43) we convert U_j' back into U_j and take the sum over j. Note that $\widetilde{\mathrm{D}}(\mathbf{a} \parallel q_t[\mathbf{f}])$ increases without bound if $\|\mathbf{a}\|_\infty \to \infty$ and, by Fenchel's inequality,

$$\mathrm{U}(\mathbf{a}) + \boldsymbol{\lambda}_t \cdot \mathbf{a} + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \geq 0$$

so in Equation (43) it suffices to take an infimum over the $\mathbf{a}$'s of a bounded norm, that is, over a compact set. By lower semi-continuity we thus obtain that the infimum is attained at some point $\mathbf{a}_t$ and

$$F_t = \widetilde{\mathrm{D}}(\mathbf{a}_t \parallel q_t[\mathbf{f}]) + \mathrm{U}(\mathbf{a}_t) + \mathrm{U}^*(-\boldsymbol{\lambda}_t) + \boldsymbol{\lambda}_t \cdot \mathbf{a}_t \quad .$$

Since $\widetilde{\mathrm{D}}(\mathbf{a} \parallel \mathbf{b})$ satisfies conditions (B1) and (B2), we obtain that

$$A(\boldsymbol{\lambda}, \mathbf{a}) = \widetilde{\mathrm{D}}(\mathbf{a} \parallel q_{\boldsymbol{\lambda}}[\mathbf{f}]) + \mathrm{U}(\mathbf{a}) + \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot \mathbf{a}$$

is an auxiliary function. Noting that $A(\boldsymbol{\lambda}_t, \mathbf{a}_t) = F_t \to 0$ and using Lemma 19 yields the result. ■

8. Species Distribution Modeling Experiments

In this section we study how generalized maxent can be applied to the problem of modeling geographic distributions of species. This is a critical topic in ecology and conservation biology: to protect a threatened species, one first needs to know its environmental requirements, that is, its *ecological niche* (Hutchinson, 1957). A model of the ecological niche can further be used to predict the set of locations with sufficient conditions for the species to persist, that is, the *potential distribution* of the species (Anderson and Martínez-Meyer, 2004; Phillips et al., 2006), or the set of locations where conditions may become suitable under future climate conditions (Hannah et al., 2005). Ecological niche models are also useful for predicting the spread of invasive species and infectious diseases (Welk et al., 2002; Peterson and Shaw, 2003), as well as understanding ecological processes such as speciation (Graham et al., 2006).

As mentioned earlier, the input for species distribution modeling typically consists of a list of georeferenced occurrence localities as well as data on a number of environmental variables which have been measured or estimated across a geographic region of interest. The most basic goal is to predict which areas within the region are within the species' potential distribution. The potential distribution can be used to estimate the species' *realized distribution*, for example by removing areas where the species is known to be absent because of deforestation or other habitat destruction. Although a species' realized distribution may exhibit some spatial correlation, the potential distribution does not, so considering spatial correlation is not necessarily desirable during species distribution modeling.

It is often the case that only *presence* data is available indicating the occurrence of the species. Natural history museum and herbarium collections constitute the richest source of occurrence localities (Ponder et al., 2001; Stockwell and Peterson, 2002). Their collections typically have no information about the *failure* to observe the species at any given location; in addition, many locations have not been surveyed. In the lingo of machine learning, this means that we have only positive examples and no negative examples from which to learn. Moreover, the number of sightings (training examples) will often be very small by machine learning standards, for example, a hundred, ten, or even less. Thus, species distribution modeling is an example of a scientifically important problem which presents a challenging area for study by the machine learning community.

To explore the utility of generalized maxent and effects of regularization, we used ℓ_1-regularized maxent to model distributions of bird species, based on occurrence records in the North American Breeding Bird Survey (Sauer et al., 2001), an extensive data set consisting of thousands of occurrence localities for North American birds and used previously for species distribution modeling (Peterson, 2001). A preliminary version of these experiments and others was evaluated by Phillips, Dudík, and Schapire (2004).

In modeling species distributions from presence-only data, sample selection bias may hinder accurate prediction. Sample selection bias refers to the fact that observations are typically more likely in places that are easier to access, such as areas close to towns, roads, airports, or waterways. The impact of sample selection bias on maxent models, and various ways of coping with it are explored by Dudík, Schapire, and Phillips (2005). Here, we assume that the bias is not significant.

A comprehensive comparison of maxent and other species distribution modeling techniques was carried out by Elith et al. (2006) on a different data set than analyzed here. In that comparison, maxent is in the group of the best-performing methods. Here, we do not perform comparison with other approaches. We use species modeling as a setting to explore various aspects of ℓ_1-regularized maxent.

From the North American Breeding Bird Survey, we selected four species with a varying number of occurrence records: Hutton's Vireo (198 occurrences), Blue-headed Vireo (973 occurrences), Yellow-throated Vireo (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 (USGS, 2001), and the remaining four were interpolated from weather station readings (New et al., 1999). Each coverage is defined over a 386×286 grid, of which 58,065 points have data for all coverages. In addition to threshold features derived from

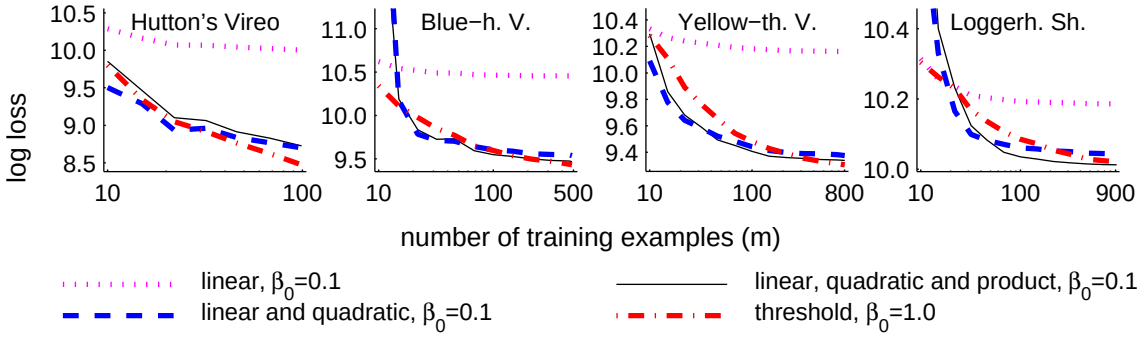

Figure 6: Learning curves. Log loss averaged over 10 partitions as a function of the number of training examples. Numbers of training examples are plotted on a logarithmic scale.

all environmental variables, we also used raw environmental variables (*linear* features), squares of environmental variables (*quadratic* features), and products of pairs of environmental variables (*product* features). Maxent with linear features finds the distribution of maximum entropy that matches empirical means of environmental variables; maxent with linear and quadratic features matches empirical means and variances; and maxent with linear, quadratic, and product features matches empirical means, variances, and covariances.

Recall that threshold features derived from a particular environmental variable are binary features equal to one if the variable is greater than a specified threshold and equal to zero otherwise. Formally, we consider a continuum of threshold features for each variable. In practice, it suffices to consider a single threshold between each pair of consecutive values appearing in the sample space; thus, in our data set we consider up to 58,064 threshold features for each variable. Given enough data, threshold features across all variables can model arbitrary additive responses in the exponent of the Gibbs distribution. Because of their expressivity, we expect that the danger of overfitting will be the most severe and regularization necessary.

In our experiments, we used ℓ_1-SUMMET of Section 6. All features are scaled to the interval $[0,1]$. Motivated by Corollary 7, we reduced the β_j's to a single regularization parameter β_0 by using $\beta_j = \beta_0 \tilde{\sigma}[f_j]/\sqrt{m}$. According to the bounds of Section 5.2, we expect that β_0 will depend on the number and complexity of features. Therefore, we expect that different values of β_0 will be optimal for different combinations of the feature types.

On each training set, we ran maxent with four different subsets of the feature types: linear (L); linear and quadratic (LQ); linear, quadratic and product (LQP); and threshold (T). 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_0 and plotted the log loss for fixed numbers of training examples as functions of β_0. These curves are referred to as sensitivity curves.

In addition to these curves, we show how Gibbs distributions returned by maxent can be interpreted in terms of contribution of individual environmental variables to the exponent. The corresponding plots are called feature profiles. We give examples of feature profiles returned by maxent with and without regularization.

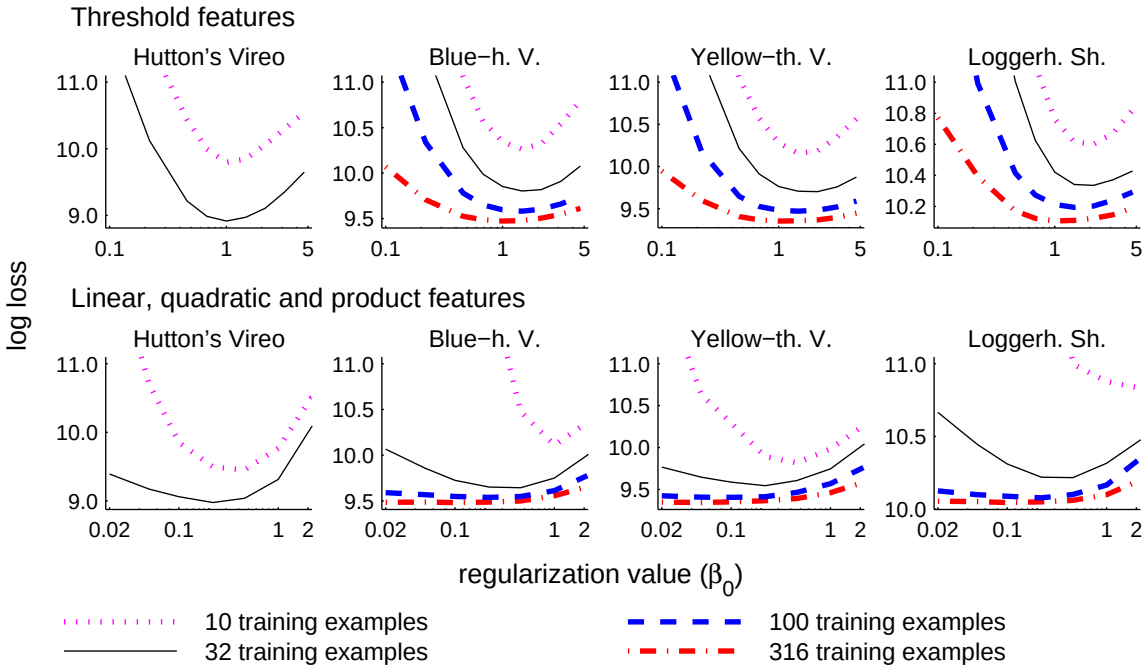

Figure 7: Sensitivity curves. Log loss averaged over 10 partitions as a function of β_0 for a varying number of training examples. For a fixed value of β_0, maxent finds better solutions (with smaller log loss) as the number of examples grows. Values of β_0 are plotted on a log scale.

Figure 6 shows learning curves for the four studied species. We set $\beta_0 = 0.1$ in L, LQ and LQP runs and $\beta_0 = 1.0$ in T runs. This choice is justified by the sensitivity curve experiments described below. In all cases, the performance improves as more samples become available. This is especially striking in the case of threshold features. In the absence of regularization, maxent would exactly fit the training data with delta functions around sample values of the environmental variables which would result in severe overfitting even when the number of training examples is large. As the learning curves show, regularized maxent does not exhibit this behavior.

Note the heavy overfitting of LQ and LQP features on the smallest sample sizes of Blue-headed Vireo and Loggerhead Shrike. A more detailed analysis of the sensitivity curves suggests that this overfitting could be alleviated by using larger values of β_0, resulting in curves qualitatively similar to those of other species. Similarly, performance of linear features, especially for larger feature sizes, could be somewhat improved using smaller regularization values.

Figure 7 shows the sensitivity of maxent to the regularization value β_0 for LQP and T versions of maxent. Results for L and LQ versions are similar to those for the LQP version. Note the remarkably consistent minimum at $\beta_0 \approx 1.0$ for threshold feature curves across different species, especially for larger sample sizes. It suggests that for the purposes of ℓ_1 regularization, $\tilde{\sigma}[f_j]/\sqrt{m}$ are good estimates of $|\tilde{\pi}[f_j] - \pi[f_j]|$ for threshold features. For LQP runs, the minima are much less pronounced as the number of samples increases and do not appear at the same value of β_0 across different species nor for different sizes of the same species. Benefits of regularization in LQP runs

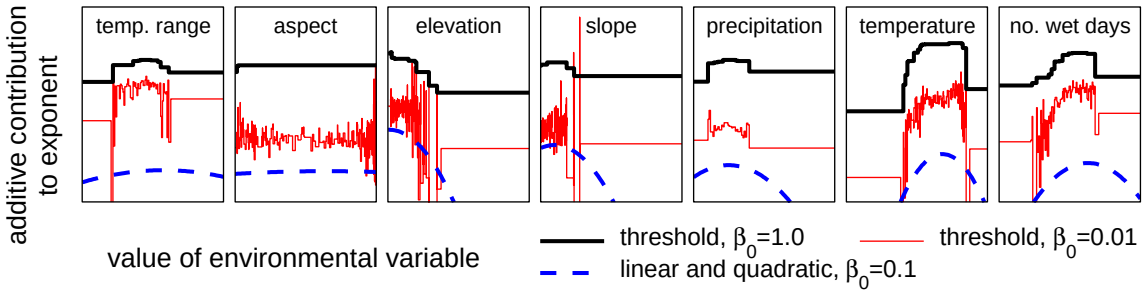

Figure 8: Feature profiles learned on the first partition of the Yellow-throated Vireo. For every environmental variable, its additive contribution to the exponent of the Gibbs distribution is given as a function of its value. Profiles have been shifted for clarity. This corresponds to adding a constant in the exponent, which has no effect on the resulting models since constants in the exponent cancel out with the normalization factor.

diminish as the number of training examples increases (this is even more so for LQ and L runs, not presented here). One possible explanation is that the relatively small number of features (compared with threshold features) prevents overfitting for large training sets.

To derive feature profiles, recall that maxent with a uniform default distribution returns the Gibbs distribution $q_{\boldsymbol{\lambda}}(x) = e^{\boldsymbol{\lambda}\cdot\mathbf{f}(x)}/Z_{\boldsymbol{\lambda}}$ minimizing the regularized log loss. For L, LQ, and T runs, the exponent is additive in contributions of individual environmental variables. Plotting this contribution as a function of the corresponding environmental variable we obtain feature profiles for the respective variables. Note that adding a constant to a profile has no impact on the resulting distribution as constants in the exponent cancel out with $Z_{\boldsymbol{\lambda}}$. For L models profiles are linear functions, for LQ models profiles are quadratic functions, and for T models profiles can be arbitrary piecewise constant functions. These profiles provide an easier to understand characterization of the distribution than the vector $\boldsymbol{\lambda}$.

Figure 8 shows feature profiles for an LQ run on the first partition of the Yellow-throated Vireo and two T runs with different values of β_0. The value of $\beta_0 = 0.01$ only prevents components of $\boldsymbol{\lambda}$ from becoming extremely large, but it does little to prevent heavy overfitting with numerous peaks capturing single training examples. Raising β_0 to 1.0 completely eliminates these peaks. This is especially prominent for the aspect variable where the regularized T as well as the LQ model show no dependence while the insufficiently regularized T model overfits heavily. Note the rough agreement between LQ profiles and regularized T profiles. Peaks in these profiles can be interpreted as intervals of environmental conditions favored by a species.[3]

3. Such interpretations should be made with caution as the objective of maxent is based solely on the predictive performance. In the extreme case, consider two identical environmental variables, only one of which has a causal effect on the species. Maxent has no knowledge which of the two variables is truly relevant, and may easily pick the wrong one, leaving the profile of the relevant one flat. Thus, interpretability is affected by correlations between variables.

9. Conclusion and Future Research Directions

The maximum entropy principle is a widely used method of density estimation. When the number of features is large, overfitting needs to be prevented by measures such as feature selection, regularization, discounting, or introduction of priors. In this work, we have provided a unified and complete account of maxent with generalized regularization. We have proved general performance guarantees and proposed versions of iterative scaling that incorporate regularization.

We have carried out analysis of several regularization types and presented scenarios in which these regularizations may be useful. In our experiments, we have shown how the principled ℓ_1 regularization facilitates learning. Further empirical study is needed to verify whether the theory derived for other regularization types corresponds to their performance. Generalizing the present analysis could help design task-specific regularization functions based on some prior information (for example properties of the feature space such as diameters with respect to various norms). Note that the quality of a regularization function can be assessed from two different perspectives: performance over test data and running time. The tradeoff between statistical guarantees and computational efficiency remains open for future research. In particular, convergence rates of algorithms presented in this paper are not known.

We have explored one direction of generalizing maxent: replacing equality constraints by an arbitrary convex potential in the primal or, equivalently, adding a convex regularization term to the maximum likelihood estimation in the dual. An alternative line of generalizations arises by replacing relative entropy in the primal objective by an arbitrary Bregman or Csiszár divergence along the lines of Altun and Smola (2006), and Collins, Schapire, and Singer (2002). Modified duality results and modified algorithms apply in the new setting, but performance guarantees do not directly translate to the case when divergences are derived from samples. Divergences of this kind are used in many cases of interest such as logistic regression (a conditional version of maxent) and boosting. In future work, we would like to generalize performance guarantees to these settings.

Finally, we have demonstrated the utility of generalized maxent in a novel application to species distribution modeling. We believe it is a scientifically important area that deserves the attention of the machine learning community while presenting some interesting challenges. Even though maxent fits the problem of species distribution modeling cleanly and effectively, there are many other techniques that could be used such as Markov random fields or mixture models. We leave the question of alternative machine learning approaches to species distribution modeling open for future research.

Acknowledgments

R. Schapire and M. Dudík received support through NSF grant CCR-0325463. We would like to thank anonymous referees for numerous suggestions that helped improve the quality of the paper.

Appendix A. Proof of Corollary 7

Proof of Corollary 7. Let

$$\beta'_j = \sqrt{\frac{\ln(4n/\delta)}{3m}} \cdot \sqrt{6\sigma^2[f_j] + \frac{\ln(4n/\delta)}{3m}} + \frac{\ln(4n/\delta)}{3m} \quad .$$

We will show that $|\tilde{\pi}[f_j]-\pi[f_j]|>\beta'_j$ with probability at most $\delta/2n$, and also $\beta'_j \geq \beta_j$ with probability at most $\delta/2n$. Then by the union bound, we obtain that

$$|\tilde{\pi}[f_j]-\pi[f_j]| \leq \beta'_j \leq \beta_j$$

for all j with probability at least $1-\delta$.

Consider a fixed j and let $\varepsilon = \ln(4n/\delta)/3m$. Thus,

$$\begin{aligned}
\beta'_j &= \sqrt{\varepsilon}\left(\sqrt{6\sigma^2[f_j]+\varepsilon}+\sqrt{\varepsilon}\right) \\
\beta_j &= \sqrt{6\varepsilon}\sqrt{\tilde{\sigma}^2[f_j]+\sqrt{\frac{\ln(2n/\delta)}{2m}}+\frac{\varepsilon}{6}}+\varepsilon \\
&= \sqrt{\varepsilon}\left(\sqrt{6\left[\tilde{\sigma}^2[f_j]+\sqrt{\ln(2n/\delta)/(2m)}\right]+\varepsilon}+\sqrt{\varepsilon}\right) \quad .
\end{aligned}$$

By Bernstein's inequality (Bernstein, 1946)

$$\begin{aligned}
\Pr\left[|\tilde{\pi}[f_j]-\pi[f_j]|>\beta'_j\right] &\leq 2\exp\left\{-\frac{3m\beta'^2_j}{6\sigma^2[f_j]+2\beta'_j}\right\} \\
&= 2\exp\left\{-\frac{3m\varepsilon\left(6\sigma^2[f_j]+\varepsilon+2\sqrt{\varepsilon}\sqrt{6\sigma^2[f_j]+\varepsilon}+\varepsilon\right)}{6\sigma^2[f_j]+2\sqrt{\varepsilon}\sqrt{6\sigma^2[f_j]+\varepsilon}+2\varepsilon}\right\} \\
&= 2\exp\{-3m\varepsilon\} = 2\exp\{-\ln(4n/\delta)\} = \delta/2n \quad .
\end{aligned}$$

To bound the probability that $\beta'_j \geq \beta_j$, it suffices to bound the probability of

$$\sigma^2[f_j] \geq \tilde{\sigma}^2[f_j]+\sqrt{\frac{\ln(2n/\delta)}{2m}} \quad .$$

We will use McDiarmid's inequality (McDiarmid, 1989) for the function

$$s(y_1,y_2,\ldots,y_m) = \frac{\sum_{i=1}^m y_i^2}{m-1} - \frac{\left(\sum_{i=1}^m y_i\right)^2}{m(m-1)} \quad .$$

Note that $\tilde{\sigma}^2[f_j] = s(f_j(x_1), f_j(x_2), \ldots, f_j(x_m))$ and $\mathrm{E}\left[\tilde{\sigma}^2[f_j]\right] = \sigma^2[f_j]$. By a simple case analysis,

$$\sup_{y_1,\ldots,y_m,y'_i\in[0,1]} |s(y_1,\ldots,y_m)-s(y_1,\ldots,y_{i-1},y'_i,y_{i+1},\ldots,y_m)| \leq \frac{1}{m}$$

for all i. Thus,

$$\begin{aligned}
\Pr\left[\sigma^2[f_j] \geq \tilde{\sigma}^2[f_j]+\sqrt{\frac{\ln(2n/\delta)}{2m}}\right] &\leq \exp\left\{\frac{-2\cdot\left[\ln(2n/\delta)/2m\right]}{m\cdot(1/m)^2}\right\} \\
&= \exp\{-\ln(2n/\delta)\} = \delta/2n \quad .
\end{aligned}$$

Hence also $\beta'_j \geq \beta_j$ with probability at most $\delta/2n$. ■

Appendix B. Derivation of $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$

It suffices to derive a single coordinate potential $\mathrm{U}_{\tilde{\pi},j}^{(\approx 1)}$:

$$\begin{aligned}
\mathrm{U}_{\tilde{\pi},j}^{(\approx 1)}(u_j) &= \sup_{\lambda_j}\left[u_j\lambda_j - \alpha_j\beta_j \ln\left(\frac{e^{\lambda_j/\alpha_j}+e^{-\lambda_j/\alpha_j}}{2}\right)\right] \\
&= \alpha_j\beta_j \sup_{\lambda_j}\left[u_j\cdot\frac{\lambda_j}{\alpha_j\beta_j} - \ln\left(\frac{1}{2}\exp\left\{\beta_j\cdot\frac{\lambda_j}{\alpha_j\beta_j}\right\}+\frac{1}{2}\exp\left\{-\beta_j\cdot\frac{\lambda_j}{\alpha_j\beta_j}\right\}\right)\right] \\
&= \alpha_j\beta_j \sup_{\lambda'_j:=\lambda_j/\alpha_j\beta_j}\left[u_j\lambda'_j - \ln\left(\frac{1}{2}e^{\beta_j\lambda'_j}+\frac{1}{2}e^{-\beta_j\lambda'_j}\right)\right] \qquad (44)\\
&= \alpha_j\beta_j\,\mathrm{D}\left(\left(\frac{1+u_j/\beta_j}{2},\frac{1-u_j/\beta_j}{2}\right)\middle\|\left(\frac{1}{2},\frac{1}{2}\right)\right) \qquad (45)
\end{aligned}$$

Equation (44) follows by change of variables. Note that the supremum in Equation (44) takes form of a dual objective in a basic maxent over a two-sample space, say $\mathcal{X}=\{0,1\}$, with a single feature $f(0)=\beta_j, f(1)=-\beta_j$, and the empirical expectation $\tilde{\pi}[f]=u_j$. Thus, by maxent duality, the value of the supremum equals $\mathrm{D}(p\,\|\,(1/2,1/2))$, where p comes from a closure of the set of Gibbs distribution and $p[f]=u_j$. However, the only distribution on $\mathcal{X}$ that satisfies the expectation constraint is

$$p(0)=\frac{1+u_j/\beta_j}{2}\quad,\qquad p(1)=\frac{1-u_j/\beta_j}{2}\quad.$$

Thus, we obtain Equation (45).

Appendix C. Proof of Lemma 10

Before we proceed with the proof of Lemma 10, we show how the trace of the feature covariance matrix can be bounded in terms of the ℓ_2 diameter of the feature space.

Lemma 22. *Let $D_2=\sup_{x,x'\in\mathcal{X}}\|\mathbf{f}(x)-\mathbf{f}(x')\|_2$ be the ℓ_2 diameter of $\mathbf{f}(\mathcal{X})$ and let $\mathbf{\Sigma}=\mathrm{E}\left[(\mathbf{f}(X)-\pi[\mathbf{f}])(\mathbf{f}(X)-\pi[\mathbf{f}])^\top\right]$, where X is distributed according to π, denote the feature covariance matrix. Then $\operatorname{tr}\mathbf{\Sigma}\le D_2^2/2$.*

Proof. Consider independent random variables X,X' distributed according to π. Let $\mathbf{f},\mathbf{f}'$ denote the random variables $\mathbf{f}(X)$ and $\mathbf{f}(X')$. Then

$$\begin{aligned}
\mathrm{E}\left[\|\mathbf{f}-\mathbf{f}'\|_2^2\right] &= \mathrm{E}\left[\mathbf{f}\cdot\mathbf{f}\right]-2\mathrm{E}\left[\mathbf{f}\right]\cdot\mathrm{E}\left[\mathbf{f}'\right]+\mathrm{E}\left[\mathbf{f}'\cdot\mathbf{f}'\right] \\
&= 2\mathrm{E}\left[\mathbf{f}\cdot\mathbf{f}\right]-2\mathrm{E}\left[\mathbf{f}\right]\cdot\mathrm{E}\left[\mathbf{f}\right] \\
&= 2\textstyle\sum_j\left[\mathrm{E}\left[f_j^2\right]-(\mathrm{E}\left[f_j\right])^2\right]=2\operatorname{tr}\mathbf{\Sigma}\quad.
\end{aligned}$$

Since $\|\mathbf{f}-\mathbf{f}'\|_2\le D_2$, we obtain $\operatorname{tr}\mathbf{\Sigma}\le D_2^2/2$. ■

Proof of Lemma 10. Consider independent samples $X_1,\ldots,X_m$ distributed according to π and the random variable $\mathbf{v}(X_1,\ldots,X_m)=\sum_i(\mathbf{f}(X_i)-\pi[\mathbf{f}])=m(\tilde{\pi}[\mathbf{f}]-\pi[\mathbf{f}])$. We will bound $\mathrm{E}\left[\|\mathbf{v}\|_2\right]$ and use McDiarmid's inequality (McDiarmid, 1989) to show that

$$\Pr\left[\|\mathbf{v}\|_2-\mathrm{E}\left[\|\mathbf{v}\|_2\right]\ge D_2\sqrt{m\ln(1/\delta)/2}\right]\le\delta\quad. \qquad (46)$$

By Jensen's inequality and Lemma 22, we obtain

$$\mathrm{E}\left[\|\mathbf{v}\|_2\right] \leq \sqrt{\mathrm{E}\left[\|\mathbf{v}\|_2^2\right]} = \sqrt{m \operatorname{tr} \boldsymbol{\Sigma}} \leq D_2 \sqrt{m/2} \ .$$

Now, by the triangle inequality,

$$\begin{aligned} \sup_{X_1,\ldots,X_m,X_i'} & \left| \left\| \mathbf{v}(X_1,\ldots,X_m) \right\|_2 - \left\| \mathbf{v}(X_1,\ldots,X_{i-1},X_i',X_{i+1},\ldots,X_m) \right\|_2 \right| \\ & \leq \sup_{X_i,X_i'} \|\mathbf{f}(X_i) - \mathbf{f}(X_i')\|_2 \leq D_2 \ , \end{aligned}$$

and Equation (46) follows by McDiarmid's inequality. ■

Appendix D. Proof of Lemma 12

Proof of Lemma 12. Let

$$\begin{aligned} \boldsymbol{\lambda}^{\star\star} &= \arg\min_{\boldsymbol{\lambda}} \left[\mathrm{L}_{\pi}(\boldsymbol{\lambda}) + \frac{\alpha \|\boldsymbol{\lambda}\|_2^2}{2} \right] \\ \hat{\boldsymbol{\lambda}} &= \arg\min_{\boldsymbol{\lambda}} \left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \frac{\alpha \|\boldsymbol{\lambda}\|_2^2}{2} \right] \ . \end{aligned}$$

As the first step, we show that

$$\left\| \boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}} \right\|_2 \leq \frac{\|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2}{\alpha} \ . \tag{47}$$

Assume that $\boldsymbol{\lambda}^{\star\star} \neq \hat{\boldsymbol{\lambda}}$ (otherwise Equation (47) holds). Let $g(\boldsymbol{\lambda})$ denote $\ln Z_{\boldsymbol{\lambda}}$. This is the cumulant or log partition function of the family of Gibbs distributions. It is well known (and not difficult to show by calculus) that this function is convex in $\boldsymbol{\lambda}$. By the convexity of $g(\boldsymbol{\lambda})$ and $\alpha\|\boldsymbol{\lambda}\|_2^2/2$, the gradients of

$$\begin{aligned} \mathrm{L}_{\pi}(\boldsymbol{\lambda}) + \frac{\alpha \|\boldsymbol{\lambda}\|_2^2}{2} &= \ln Z_{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \cdot \pi[\mathbf{f}] + \frac{\alpha \|\boldsymbol{\lambda}\|_2^2}{2} \\ \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \frac{\alpha \|\boldsymbol{\lambda}\|_2^2}{2} &= \ln Z_{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \cdot \tilde{\pi}[\mathbf{f}] + \frac{\alpha \|\boldsymbol{\lambda}\|_2^2}{2} \end{aligned}$$

at their respective minima must equal zero:

$$\begin{aligned} \nabla g(\boldsymbol{\lambda}^{\star\star}) - \pi[\mathbf{f}] + \alpha \boldsymbol{\lambda}^{\star\star} &= 0 \\ \nabla g(\hat{\boldsymbol{\lambda}}) - \tilde{\pi}[\mathbf{f}] + \alpha \hat{\boldsymbol{\lambda}} &= 0 \ . \end{aligned}$$

Taking the difference yields

$$\alpha(\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) = -(\nabla g(\boldsymbol{\lambda}^{\star\star}) - \nabla g(\hat{\boldsymbol{\lambda}})) + (\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]) \ .$$

Multiplying both sides by $(\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}})$, we obtain

$$\begin{aligned} \alpha \|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2^2 &= -(\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\nabla g(\boldsymbol{\lambda}^{\star\star}) - \nabla g(\hat{\boldsymbol{\lambda}})) + (\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]) \\ &\leq (\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]) \qquad (48) \\ &\leq \|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2 \|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2 \ . \qquad (49) \end{aligned}$$

Equation (48) follows because by convexity of $g(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$

$$(\nabla g(\boldsymbol{\lambda}_2) - \nabla g(\boldsymbol{\lambda}_1)) \cdot (\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1) \geq 0 \ .$$

Equation (49) follows by the Cauchy-Schwartz inequality. Dividing (49) by $\alpha \|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2$ we obtain Equation (47). Now, by Lemma 3.13, the Cauchy-Schwartz inequality, Equation (47) and the optimality of $\boldsymbol{\lambda}^{\star\star}$ we obtain

$$\begin{aligned}
\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^{\star\star}) + (\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]) + \mathrm{U}_{\tilde{\pi}}^{(2)*}(\boldsymbol{\lambda}^{\star\star}) - \mathrm{U}_{\tilde{\pi}}^{(2)*}(\hat{\boldsymbol{\lambda}}) \\
&\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^{\star\star}) + \|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2 \|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2 + \frac{\alpha \|\boldsymbol{\lambda}^{\star\star}\|_2^2}{2} - \frac{\alpha \|\hat{\boldsymbol{\lambda}}\|_2^2}{2} \\
&\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^{\star\star}) + \frac{\|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2^2}{\alpha} + \frac{\alpha \|\boldsymbol{\lambda}^{\star\star}\|_2^2}{2} \\
&\leq \mathrm{L}_\pi(\boldsymbol{\lambda}^{\star}) + \frac{\|\pi[\mathbf{f}] - \tilde{\pi}[\mathbf{f}]\|_2^2}{\alpha} + \frac{\alpha \|\boldsymbol{\lambda}^{\star}\|_2^2}{2} \ .
\end{aligned}$$

■

Appendix E. Derivation of $\mathrm{U}_{\tilde{\pi}}^{(1+2)}$

It suffices to derive a single coordinate potential $\mathrm{U}_{\tilde{\pi},j}^{(1+2)}$:

$$\begin{aligned}
\mathrm{U}_{\tilde{\pi},j}^{(1+2)}(u_j) &= \sup_{\lambda_j} \left(u_j \lambda_j - \beta |\lambda_j| - \frac{\alpha \lambda_j^2}{2} \right) \\
&= \sup_{\lambda_j : u_j \lambda_j = |u_j||\lambda_j|} \left(\frac{\alpha}{2} \cdot |\lambda_j| \cdot \left[\frac{2\,(|u_j| - \beta)}{\alpha} - |\lambda_j| \right] \right) \ . \qquad (50)
\end{aligned}$$

In Equation (50) we note that for each pair $\pm\lambda_j$, it suffices to consider the value whose sign agrees with u_j. Next, if $|u_j| \leq \beta$ then the bracketed expression is non-positive, hence the supremum is attained at $\lambda_j = 0$ and its value equals 0. For $|u_j| > \beta$, the supremum is attained when $|\lambda_j| = (|u_j| - \beta)/\alpha$, in which case its value equals $(|u_j| - \beta)^2/(2\alpha)$.

Appendix F. Proof of Lemma 18

We will first prove a single coordinate version of Lemma 18 and then turn to the general case.

Lemma 23. *Let $\psi : \mathbb{R} \to (-\infty, \infty]$ be a proper closed convex function. Let $S = \mathrm{dom}\,\psi = \{u \in \mathbb{R} : \psi(u) < \infty\}$ and $T_c = \{v \in \mathbb{R} : \psi^*(v) \leq c\}$. Then there exists $\alpha_c \geq 0$ such that $uv \leq \alpha_c |u|$ for all $u \in S, v \in T_c$.*

Proof. Inequality $uv \leq \alpha_c |u|$ holds for an arbitrary α_c if $u = 0$. We determine α_c separately for cases $u \in S_+ = S \cap (0, \infty)$ and $u \in S_- = S \cap (-\infty, 0)$ and choose the maximum.

Assume $S_+ \neq \emptyset$ and pick an arbitrary $u_+ \in S_+$. Then for any $v \in T_c$ by Fenchel's inequality

$$u_+ v \leq \psi(u_+) + \psi^*(v) \leq \psi(u_+) + c$$

and thus

$$v \leq \frac{\psi(u_+) + c}{u_+} \ .$$

Now for any $u \in S_+$

$$uv \leq u \cdot \frac{\psi(u_+)+c}{u_+} \leq |u| \cdot \left| \frac{\psi(u_+)+c}{u_+} \right| \quad .$$

Similarly, if $S_- \neq \emptyset$ then we can choose an arbitrary $u_- \in S_-$ and obtain for all $u \in S_-$

$$uv \leq |u| \cdot \left| \frac{\psi(u_-)+c}{u_-} \right|.$$

To complete the proof we choose

$$\alpha_c = \max \left\{ \left| \frac{\psi(u_+)+c}{u_+} \right|, \left| \frac{\psi(u_-)+c}{u_-} \right| \right\}$$

setting the respective terms to 0 if S_+ or S_- is empty. ■

Proof of Lemma 18. Assume that $\mathrm{U}_r(\mathbf{u}) < \infty$ and thus by decomposability $\mathrm{U}_{r,j}(u_j) < \infty$ for all j. Also assume that $\mathrm{U}_r^*(\boldsymbol{\lambda}) = \sum_j \mathrm{U}_{r,j}^*(\lambda_j) < c$. By Fenchel's inequality $\mathrm{U}_{r,j}^*(\lambda_j) \geq -\mathrm{U}_{r,j}(0)$ which is finite by the feasibility of r. Since the sum of $\mathrm{U}_{r,j}^*(\lambda_j)$ is bounded above by c and individual functions are bounded below by constants, they must also be bounded above by some constants c_j. By Lemma 23 applied to coordinate potentials, we obtain that $u_j\lambda_j \leq \alpha_j|u_j|$ for some constants $\alpha_1, \ldots, \alpha_n$. The conclusion follows by taking $\alpha_c = \max_j \alpha_j$. ■

Appendix G. Ensuring Finite Updates

In this appendix, we discuss how to ensure that features and coordinate potentials are non-degenerate in SUMMET and PLUMMET, and show that non-degeneracy implies that updates in both algorithms are always finite.

G.1 Non-degeneracy in SUMMET

In SUMMET, we assume that $f_j(\mathcal{X}) \subseteq [0,1]$. In context of this algorithm, a feature f_j is degenerate if $f_j(\mathcal{X}) = \{0\}$ or $f_j(\mathcal{X}) = \{1\}$ and a coordinate potential U_j is degenerate if $\operatorname{dom}\mathrm{U}_j \cap [0,1] = \{0\}$ or $\operatorname{dom}\mathrm{U}_j \cap [0,1] = \{1\}$. In order to obtain non-degenerate features and coordinate potentials, it suffices to preprocess the sample space $\mathcal{X}$ and the feature set as follows:

1. For all j: if $\operatorname{dom}\mathrm{U}_j \cap [0,1] = \{0\}$ then $\mathcal{X} \leftarrow \{x \in \mathcal{X} : f_j(x) = 0\}$.
2. For all j: if $\operatorname{dom}\mathrm{U}_j \cap [0,1] = \{1\}$ then $\mathcal{X} \leftarrow \{x \in \mathcal{X} : f_j(x) = 1\}$.
3. For all j: if $f_j(x) = 0$ for all $x \in \mathcal{X}$ then remove feature f_j.
4. For all j: if $f_j(x) = 1$ for all $x \in \mathcal{X}$ then remove feature f_j.

Whenever U_j is degenerate, steps 1–2 guarantee that f_j will be eventually removed in steps 3–4. While f_j could be removed immediately in steps 1–2, note that steps 3–4 are still necessary since features may be degenerate even when potentials are not. Also note that steps 1–2 must precede steps 3–4 since restricting $\mathcal{X}$ may introduce new degenerate features.

The preprocessing described above yields an equivalent form of the primal. By restricting the sample space in steps 1–2, we effectively eliminate distributions that are nonzero outside the restricted sample set. Note that those distributions are infeasible because their feature means lie outside $\operatorname{dom}\mathrm{U}$. In steps 3–4, we simply remove constant terms of the potential function.

Theorem 24. *Let $\boldsymbol{\lambda}$ and $Q(\boldsymbol{\lambda})$ be finite and f_j, U_j non-degenerate. Then $F_j(\boldsymbol{\lambda}, \delta)$ is maximized by a finite δ.*

Proof. We will show that $F_j(\boldsymbol{\lambda}, \delta) \to -\infty$ if $\delta \to \pm\infty$. Thus, it suffices to consider δ from a compact interval and the result follows by upper semi-continuity of F_j. First, consider the case $\delta \to \infty$. Let r be an arbitrary feasible distribution. Rewrite $F_j(\boldsymbol{\lambda}, \delta)$ as follows:

$$\begin{aligned} F_j(\boldsymbol{\lambda}, \delta) &= -\ln\big(1 + (e^{\delta} - 1) q_{\boldsymbol{\lambda}}[f_j]\big) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \\ &= -\ln\left\{e^{\delta}\left[e^{-\delta}(1 - q_{\boldsymbol{\lambda}}[f_j]) + q_{\boldsymbol{\lambda}}[f_j]\right]\right\} + \delta r[f_j] - \mathrm{U}_{r,j}^*(\lambda_j + \delta) + \mathrm{U}_{r,j}^*(\lambda_j) \\ &= -\ln\left[e^{-\delta}(1 - q_{\boldsymbol{\lambda}}[f_j]) + q_{\boldsymbol{\lambda}}[f_j]\right] - \delta(1 - r[f_j]) - \mathrm{U}_{r,j}^*(\lambda_j + \delta) + \mathrm{U}_{r,j}^*(\lambda_j). \end{aligned} \tag{51}$$

Suppose that $r[f_j] < 1$. Then $F_j(\boldsymbol{\lambda}, \delta) \to -\infty$: the first term of (51) is bounded above by $-\ln(q_{\boldsymbol{\lambda}}[f_j])$ which is finite by non-degeneracy of f_j; the second term decreases without bound; the third term is bounded above by $\mathrm{U}_{r,j}(0)$ by Fenchel's inequality; and the fourth term is a finite constant because $Q(\boldsymbol{\lambda})$ is finite. In case $r[f_j] = 1$, the second term equals zero, but the third term decreases without bound because by non-degeneracy of U_j there exists $\varepsilon > 0$ such that $\mathrm{U}_{r,j}(\varepsilon) = \mathrm{U}_j(1 - \varepsilon) < \infty$ and hence by Fenchel's inequality $-\mathrm{U}_{r,j}^*(\lambda_j + \delta) \leq -(\lambda_j + \delta)\varepsilon + \mathrm{U}_{r,j}(\varepsilon)$.

Now consider $\delta \to -\infty$ and rewrite $F_j(\boldsymbol{\lambda}, \delta)$ as follows:

$$F_j(\boldsymbol{\lambda}, \delta) = -\ln\big((1 - q_{\boldsymbol{\lambda}}[f_j]) + e^{\delta} q_{\boldsymbol{\lambda}}[f_j]\big) + \delta r[f_j] - \mathrm{U}_{r,j}^*(\lambda_j + \delta) + \mathrm{U}_{r,j}^*(\lambda_j) \ .$$

Assuming that $r[f_j] > 0$, the second term decreases without bound and the remaining terms are bounded above. If $r[f_j] = 0$ then the third term decreases without bound because by non-degeneracy of U_j there exists $\varepsilon > 0$ such that $\mathrm{U}_{r,j}(-\varepsilon) = \mathrm{U}_j(\varepsilon) < \infty$ and thus by Fenchel's inequality $-\mathrm{U}_{r,j}^*(\lambda_j + \delta) \leq (\lambda_j + \delta)\varepsilon + \mathrm{U}_{r,j}(-\varepsilon)$. ■

Corollary 25. *Updates of* SUMMET *are always finite.*

Proof. We proceed by induction. In the first step, both $\boldsymbol{\lambda}_1$ and $Q(\boldsymbol{\lambda}_1)$ are finite (see proof of Theorem 20). Now suppose that in step t, $\boldsymbol{\lambda}_t$ and $Q(\boldsymbol{\lambda}_t)$ are finite. Then by Theorem 24, all considered coordinate updates will be finite, so $\boldsymbol{\lambda}_{t+1}$ will be finite too. Since $Q(\boldsymbol{\lambda}_{t+1}) \geq Q(\boldsymbol{\lambda}_t)$ and $Q(\boldsymbol{\lambda})$ is bounded above (see proof of Theorem 20), we obtain that $Q(\boldsymbol{\lambda}_{t+1})$ is finite. ■

G.2 Non-degeneracy in PLUMMET

In this case, we assume that $f_j(x) \geq 0$ and $\sum_j f_j(x) \leq 1$ for all $x \in \mathcal{X}$. We call a feature f_j degenerate if $f_j(\mathcal{X}) = \{0\}$ and a coordinate potential U_j degenerate if $\operatorname{dom}\mathrm{U}_j \cap [0,1] = \{0\}$. To obtain non-degenerate features and coordinate potentials, it suffices to preprocess the sample space $\mathcal{X}$ and the feature set as follows:

1. For all j: if $\operatorname{dom} U_j \cap [0,1] = \{0\}$ then $\mathcal{X} \leftarrow \{x \in \mathcal{X} : f_j(x) = 0\}$.
2. For all j: if $f_j(x) = 0$ for all $x \in \mathcal{X}$ then remove feature f_j.

Similarly to SUMMET, this preprocessing derives an equivalent form of the primal. Using analogous reasoning as in Theorem 24, we show below that non-degeneracy implies finite updates in PLUMMET.

In each iteration of the algorithm we determine updates δ_j by maximizing

$$\begin{aligned} F_j(\boldsymbol{\lambda},\delta) &= -q_{\boldsymbol{\lambda}}[f_j](e^{\delta}-1) - \mathrm{U}_j^*(-\lambda_j-\delta) + \mathrm{U}_j^*(-\lambda_j) \\ &= -q_{\boldsymbol{\lambda}}[f_j](e^{\delta}-1) + \delta r[f_j] - \mathrm{U}_{r,j}^*(\lambda_j+\delta) + \mathrm{U}_{r,j}^*(\lambda_j) \ . \end{aligned}$$

It suffices to prove that $F_j(\boldsymbol{\lambda},\delta) \to -\infty$ if $\delta \to \pm\infty$ given that $Q(\boldsymbol{\lambda})$ and λ_j are finite and f_j, U_j are non-degenerate.

First, we rewrite F_j as follows:

$$F_j(\boldsymbol{\lambda},\delta) = -e^{\delta}\left[q_{\boldsymbol{\lambda}}[f_j] - e^{-\delta}q_{\boldsymbol{\lambda}}[f_j] - e^{-\delta}\delta r[f_j]\right] - \mathrm{U}_{r,j}^*(\lambda_j+\delta) + \mathrm{U}_{r,j}^*(\lambda_j) \ .$$

If $\delta \to \infty$ then the expression in the brackets approaches $q_{\boldsymbol{\lambda}}[f_j]$, which is positive by non-degeneracy of f_j. Thus the first term decreases without bound while the second and third terms are bounded from above. Next, rewrite F_j as

$$F_j(\boldsymbol{\lambda},\delta) = \delta\left[r[f_j] - \frac{e^{\delta}}{\delta}q_{\boldsymbol{\lambda}}[f_j] + \frac{1}{\delta}q_{\boldsymbol{\lambda}}[f_j]\right] - \mathrm{U}_{r,j}^*(\lambda_j+\delta) + \mathrm{U}_{r,j}^*(\lambda_j) \ .$$

If $\delta \to -\infty$ then the expression in the brackets approaches $r[f_j]$. Thus, if $r[f_j] > 0$ then the first term decreases without bound and the other two terms are bounded above. If $r[f_j] = 0$ then the first term approaches $q_{\boldsymbol{\lambda}}[f_j]$ and the second term decreases without bound because, by non-degeneracy of U_j, there exists $\varepsilon > 0$ such that $\mathrm{U}_{r,j}(-\varepsilon) = \mathrm{U}_j(\varepsilon) < \infty$ and hence by Fenchel's inequality $-\mathrm{U}_{r,j}^*(\lambda_j + \delta) \le (\lambda_j+\delta)\varepsilon + \mathrm{U}_{r,j}(-\varepsilon)$.

References

Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the Nineteenth Annual Conference on Learning Theory*, 2006.

R. P. Anderson and E. Martínez-Meyer. Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (*Heteromys*) of Ecuador. *Biological Conservation*, 116:167–179, 2004.

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

S. N. Bernstein. *Theory of Probability*. Gostekhizdad, 1946.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(1):200–217, 1967.

N. Cesa-Bianchi, A. Krogh, and M. K. Warmuth. Bounds on approximate steepest descent for likelihood maximization in exponential families. *IEEE Transactions on Information Theory*, 40 (4):1215–1220, July 1994.

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.

M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1):253–285, 2002.

J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

O. Dekel, S. Shalev-Shwartz, and Y. Singer. Smooth ε-insensitive regression by loss symmetrization. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, pages 433–447. Springer, 2003.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):1–13, April 1997.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Duality and auxiliary functions for Bregman distances. Technical Report CMU-CS-01-109, School of Computer Science, Carnegie Mellon University, 2001.

L. Devroye. Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, 12:72–79, 1982.

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, March 2003.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 (3):425–455, August 1994.

M. Dudík and R. E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Proceedings of the Nineteenth Annual Conference on Learning Theory*, pages 123–138. Springer-Verlag, 2006.

M. Dudík, R. E. Schapire, and S. J. Phillips. Correcting sample selection bias in maximum entropy density estimation. In *Advances in Neural Information Processing Systems 18*, pages 323–330. MIT Press, 2005.

J. Elith. Quantitative methods for modeling species habitat: Comparative performance and an application to Australian plants. In Scott Ferson and Mark Burgman, editors, *Quantitative Methods for Conservation Biology*, pages 39–58. Springer-Verlag, New York, 2002.

J. Elith, C. H. Graham, and the NCEAS Species Distribution Modelling Group. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

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

J. Goodman. Sequential conditional generalized iterative scaling. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, July 2002.

J. Goodman. Exponential priors for maximum entropy models. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2004.

C. H. Graham, C. Moritz, and S. E. Williams. Habitat history improves prediction of biodiversity in rainforest fauna. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):632–636, January 2006.

L. Hannah, G. Midgley, G. Hughes, and B. Bomhard. The view from the Cape: Extinction risk, protected areas, and climate change. *BioScience*, 55(3), March 2005.

A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

G. E. Hutchinson. Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22:415–427, 1957.

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

B. M. Jedynak and S. Khudanpur. Maximum likelihood set for estimating a probability mass function. *Neural Computation*, 17:1508–1530, 2005.

J. Kazama and J. Tsujii. Evaluation and extension of maximum entropy models with inequality constraints. In *Conference on Empirical Methods in Natural Language Processing*, pages 137–144, 2003.

S. P. Khudanpur. A method of maximum entropy estimation with relaxed constraints. In *Proceedings of the Johns Hopkins University Language Modeling Workshop*, pages 1–17, 1995.

B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, June 2005.

R. Lau. Adaptive statistical language modeling. Master's thesis, MIT Department of Electrical Engineering and Computer Science, May 1994.

J. R. Leathwick, J. Elith, M. P. Francis, T. Hastie, and P. Taylor. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, 321:267–281, 2006.

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

G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. Technical Report CMU-CS-01-144, CMU School of Computer Science, October 2001.

R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning*, pages 49–55, 2002.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, 1989.

G. G. Moisen and T. S. Frescino. Comparing five modeling techniques for predicting forest characteristics. *Ecological Modeling*, 157:209–225, 2002.

M. New, M. Hulme, and P. 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.

W. I. Newman. Extension to the maximum entropy method. *IEEE Transactions on Information Theory*, IT-23(1):89–93, January 1977.

A. Y. Ng. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 615–622, 2004.

A. T. Peterson. Predicting species' geographic distributions based on ecological niche modeling. *The Condor*, 103:599–605, 2001.

A. T. Peterson and J. Shaw. *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: Ecological niche models, predicted geographic distribution, and climate change effects. *International Journal of Parasitology*, 33:919–931, 2003.

S. J. Phillips, R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3–4):231–259, 2006.

S. J. Phillips, M. Dudík, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 655–662, 2004.

W. F. Ponder, G. A. Carter, P. Flemons, and R. R. Chapman. Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, 15:648–657, 2001.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228, 1996.

S. Rosset and E. Segal. Boosting density estimation. In *Advances in Neural Information Processing Systems 15*, pages 641–648. MIT Press, 2003.

R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani. On the convergence of bound optimization algorithms. In *Uncertainty in Artificial Intelligence 19*, pages 509–516, 2003.

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.

R. E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

D. R. B. Stockwell and A. T. Peterson. Controlling bias in biodiversity data. In J. Michael Scott, Patricia J. Heglund, Michael L. Morrison, Jonathan B. Haufler, Martin G. Raphael, William A. Wall, and Fred B. Samson, editors, *Predicting Species Occurrences: Issues of Accuracy and Scale*, pages 537–546. Island Press, Washington, DC, 2002.

R. Tibshirani. Regression shrikage and selection via the lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

USGS. HYDRO 1k, elevation derivative database. Available at http://edcdaac.usgs.gov/gtopo30/hydro/, 2001. United States Geological Survey, Sioux Falls, South Dakota.

E. Welk, K. Schubert, and M. H. Hoffmann. Present and potential distribution of invasive mustard (*Alliara petiolata*) in North America. *Diversity and Distributions*, 8:219–233, 2002.

M. Welling, R. S. Zemel, and G. E. Hinton. Self supervised boosting. In *Advances in Neural Information Processing Systems 15*, pages 665–672. MIT Press, 2003.

P. M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7 (1):117–143, 1995. ISSN 0899-7667.

T. Zhang. Class-size independent generalization analysis of some discriminative multi-category classification. In *Advances in Neural Information Processing Systems 17*, pages 1625–1632, 2005.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320, 2005.