Appearing in Proceedings of the 19th Annual Conference on Learning Theory, 2006.

Maximum Entropy Distribution Estimation with Generalized Regularization

Miroslav Dudík and Robert E. Schapire

Princeton University, Department of Computer Science,
35 Olden Street, Princeton, NJ 08540
{mdudik,schapire}@cs.princeton.edu

Abstract. We present a unified and complete account of maximum entropy distribution 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 can 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. Furthermore, our general approach enables us to use information about the structure of the feature space or about sample selection bias to derive entirely new regularization functions with superior guarantees. We propose an algorithm solving a large and general subclass of generalized maxent problems, including all discussed in the paper, and prove its convergence. Our approach generalizes techniques based on information geometry and Bregman divergences as well as those based more directly on compactness.

1 Introduction

The maximum entropy (maxent) approach to probability distribution estimation was first proposed by Jaynes [1], and has since been used in many areas of computer science and statistical learning, especially natural language processing [2, 3], and more recently in species habitat modeling [4]. In maxent, one is given a set of samples from a target distribution over some space, and a set of known constraints on the distribution. The distribution is then estimated by a distribution of maximum entropy satisfying the given constraints. The constraints are often represented using a set of *features* (real-valued functions) on the space, with the expectation of every feature required to match its empirical average. By convex duality, this turns out to be the unique Gibbs distribution maximizing the likelihood of the samples.

While intuitively appealing, this approach fails to produce good estimates when the number of features is large compared with the number of samples. Conceptually, constraining maxent to match a large number of feature averages exactly forces the algorithm to approximate the empirical distribution too closely. From the dual perspective, the family of Gibbs distributions is too expressive and the algorithm overfits. Common approaches to counter overfitting are regularization [5–8], introduction of a prior [9], feature selection [2, 3], discounting [5, 6] and constraint relaxation [10, 11]. Thus, there are many ways to control overfitting in maxent calling for a general treatment.

In this work, we study a generalized form of maxent. Although mentioned by other authors as *fuzzy maxent* [5–7], we give the first complete theoretical treatment of this very general framework, including fully general performance guarantees, algorithms and convergence proofs. Independently, Altun and Smola [12] 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.

A crucial insight of our general analysis is that maxent relaxations corresponding to tighter constraints on the feature expectations yield better performance guarantees. Applying our analysis to the special case in which such a confidence region is polyhedral allows us to derive novel regularization functions and a corresponding analysis for two cases of particular interest. The first case is when some information about structure of the feature space is available, for example, when some features are known to be squares or products of other "base" features, corresponding to constraints on variances or covariances of the base features. The second case is when the sample selection process is known to be biased. Both of these cases were studied previously [4, 13]. Here, we apply our general framework to derive improved generalization bounds using an entirely new form of regularization. These results improve on bounds for previous forms of regularization by up to a factor of eight — an improvement that would otherwise require a 64-fold increase in the number of training examples.

In the second part, we propose an algorithm solving a large and general subclass of generalized maxent problems. We show convergence of our algorithm using techniques that unify previous approaches and extend them to a more general setting. Specifically, our unified approach generalizes techniques based on information geometry and Bregman divergences [3, 14] as well as those based more directly on compactness [11]. 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.

There are many standard maxent algorithms, such as iterative scaling [3, 15], gradient descent, Newton and quasi-Newton methods [16] and their regularized versions [5, 6, 9, 10, 17]. In this paper, we focus on an algorithm that performs sequential updates of feature weights similarly to boosting and sequential algorithms considered in [11, 14]. Sequential updates are especially desirable when the number of features is very large or when they are produced by a weak learner. When the number of features is small, techniques developed here can be directly applied to derive a parallel update algorithm analogous to the one proposed in [11] for ℓ_1-regularized maxent (details omitted).

Previous Work. There have been many studies of maxent and logistic regression, which is a conditional version of maxent, with ℓ_1-style regularization [9–11, 17, 18], ℓ_2^2-style regularization [5–8] as well as some other types of regularization such as $\ell_1 +$ ℓ_2^2-style [10] and ℓ_2-style regularization [19]. In a recent work, Altun and Smola [12] explore regularized formulations (with duality and performance guarantees) where the entropy is replaced by an arbitrary Bregman or Csiszár divergence and regularization equals a norm raised to a power greater than one. With the exception of [8, 11, 12], previous work does not include guarantees applicable to our case, albeit Krishnapuram et al. [17] and Ng [18] give guarantees for ℓ_1-regularized logistic regression.

2 Preliminaries

The goal is to estimate an unknown *target distribution* π over a *sample space* $\mathcal{X}$ based on *samples* $x_1, \ldots, x_m \in \mathcal{X}$. We assume that samples are independently distributed according to π and denote the *empirical distribution* by $\tilde{\pi}(x) = |\{1 \le i \le m : x_i = x\}|/m$. The structure of the problem is specified by real valued functions $f_1, \ldots, f_n$ on the sample space, called *features* and by a distribution q_0 representing a *default estimate*. The vector of all n features is denoted by $\boldsymbol{f}$ and the image of $\mathcal{X}$ under $\boldsymbol{f}$, the *feature space*, is denoted by $\boldsymbol{f}(\mathcal{X})$. We assume that features capture all the relevant information available for the problem at hand and q_0 is the distribution we would choose if we were given no samples. The distribution q_0 is most often assumed uniform.

Let $p[f]$ denote the expectation of a function $f(x)$ when x is chosen randomly according to distribution p. For a limited number of samples, we expect that $\tilde{\pi}$ will be a poor estimate of π under any reasonable distance measure. 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 that satisfy them, we choose the distribution that minimizes entropy relative to the default estimate q_0. When q_0 is uniform this is the same as maximizing the entropy. Here, as usual, the entropy of a distribution p is defined as $\mathrm{H}(p) = p[\ln(1/p)]$ and the relative entropy, or Kullback-Leibler divergence, as $\mathrm{D}(p \parallel q) = p[\ln(p/q)]$. Thus, the maximum entropy principle chooses the distribution that satisfies the constraints, but imposes as little additional information as possible when compared with q_0.

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

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

where $Z_{\boldsymbol{\lambda}} = \sum_{x\in\mathcal{X}} q_0(x)e^{\boldsymbol{\lambda}\cdot\boldsymbol{f}(x)}$ is a normalizing constant and $\boldsymbol{\lambda} \in \mathbb{R}^n$. It can be proved [3] that the maxent distribution is the maximum likelihood distribution from the closure of the set of Gibbs distributions. Equivalently, it is the distribution that achieves the infimum over all values of $\boldsymbol{\lambda}$ of the empirical log loss $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) = -\frac{1}{m}\sum_{i=1}^{m} \ln q_{\boldsymbol{\lambda}}(x_i)$.

The convex programs corresponding to the two optimization problems are

$$\begin{aligned} \mathcal{P}_{\text{basic}}: &\quad \min_{p\in\Delta} \mathrm{D}(p \parallel q_0) \text{ subject to } p[\boldsymbol{f}] = \tilde{\pi}[\boldsymbol{f}] \\ \mathcal{Q}_{\text{basic}}: &\quad \inf_{\boldsymbol{\lambda}\in\mathbb{R}^n} \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) \end{aligned}$$

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

In general, we use $\mathrm{L}_p(\boldsymbol{\lambda}) = -p[\ln q_{\boldsymbol{\lambda}}]$ to denote the log loss of $q_{\boldsymbol{\lambda}}$ relative to the distribution p. It differs from relative entropy $\mathrm{D}(p \parallel q_{\boldsymbol{\lambda}})$ only by the constant $\mathrm{H}(p)$. We will use the two interchangeably as objective functions.

3 Convex Analysis Background

Throughout this paper we make use of convex analysis. The most relevant concepts are convex conjugacy and Fenchel's duality which we introduce here (see also [20, 21]).

Consider a function $\psi : \mathbb{R}^n \to (-\infty, \infty]$. The *effective domain* of ψ is the set $\operatorname{dom}\psi = \{\boldsymbol{u} \in \mathbb{R}^n \mid \psi(\boldsymbol{u}) < \infty\}$. A point $\boldsymbol{u}$ where $\psi(\boldsymbol{u}) < \infty$ is called *feasible*. The *epigraph* of ψ is the set of points above its graph $\{(\boldsymbol{u}, t) \in \mathbb{R}^n \times \mathbb{R} \mid t \ge \psi(\boldsymbol{u})\}$.

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 $\liminf_{\boldsymbol{u}'\to\boldsymbol{u}} \psi(\boldsymbol{u}') \geq \psi(\boldsymbol{u})$ for all $\boldsymbol{u}$).

If ψ is a closed proper convex function then its *conjugate* is defined by

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

The conjugate provides an alternative description of ψ in terms of tangents to ψ's epigraph. It turns out that ψ^* is also a closed proper convex function and $\psi^{**} = \psi$ (for a proof see Corollary 12.2.1 of [20]). From the definition of conjugate, we obtain *Fenchel's inequality*

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

In this work we use several examples of closed proper convex functions. The first of them is relative entropy, with the second argument fixed, viewed as a function of its first argument and extended to $\mathbb{R}^{\mathcal{X}}$ in the following manner: $\psi(p) = \mathrm{D}(p \parallel q_0)$ if $p \in \Delta$ and equals infinity otherwise. The conjugate of relative entropy is the log partition function $\psi^*(r) = \ln\left(\sum_{x\in\mathcal{X}} q_0(x)e^{r(x)}\right)$.

Relative entropy is also an example of a Bregman divergence which generalizes some common distance measures including the squared Euclidean distance. We use two properties satisfied by any Bregman divergence $\mathrm{B}(\cdot \parallel \cdot)$:

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

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. The conjugate of an indicator function is a *support function*. For $C = \{\boldsymbol{v}\}$, we obtain $\mathrm{I}^*_{\{\boldsymbol{v}\}}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}\cdot\boldsymbol{v}$. For a box $R = [-\beta,\beta]^n$, we obtain a scaled ℓ_1 norm $\mathrm{I}^*_R(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_1$, and for a Euclidean ball $B = \{\boldsymbol{u} \mid \|\boldsymbol{u}\|_2 \leq \beta\}$, a scaled ℓ_2 norm, $\mathrm{I}^*_B(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_2$. If C is a convex hull of closed convex sets C_1, C_2 then

$$\mathrm{I}^*_C(\boldsymbol{\lambda}) = \max\{\mathrm{I}^*_{C_1}(\boldsymbol{\lambda}), \mathrm{I}^*_{C_2}(\boldsymbol{\lambda})\}. \tag{1}$$

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

$$\begin{aligned} &\text{if } \varphi(\boldsymbol{u}) = \psi(\gamma\boldsymbol{u} + \boldsymbol{c}) && \text{then } \varphi^*(\boldsymbol{\lambda}) = \psi^*(\boldsymbol{\lambda}/\gamma) - \boldsymbol{\lambda}\cdot\boldsymbol{c}/\gamma && (2)\\ &\text{if } \varphi(\boldsymbol{u}) = \textstyle\sum_j \varphi_j(u_j) && \text{then } \varphi^*(\boldsymbol{\lambda}) = \textstyle\sum_j \varphi^*_j(\lambda_j) && (3)\end{aligned}$$

where $\gamma \in \mathbb{R}\setminus\{0\}$ and $\boldsymbol{c}\in\mathbb{R}^n$ are constants and u_j, λ_j refer to components of $\boldsymbol{u}, \boldsymbol{\lambda}$.

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 [20] 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* $\boldsymbol{A}$ *a real-valued* $m\times n$ *matrix. Assume that* $\operatorname{dom}\psi^* = \mathbb{R}^n$ *or* $\operatorname{dom}\varphi = \mathbb{R}^m$. *Then*

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

We refer to the minimization over $\boldsymbol{u}$ 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 objective as the dual problem. We call $\boldsymbol{u}$ a primal feasible point if the primal objective is finite at $\boldsymbol{u}$ and analogously define a dual feasible point.

4 Generalized Maximum Entropy

In this paper we study a generalized maxent problem

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

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 $\mathcal{X}$, i.e. $\mathrm{D}(p \parallel q_0)$ is finite for all $p \in \Delta$, and $p_0[\boldsymbol{f}]$ is a feasible point of U for at least one distribution p_0. The latter will typically be satisfied by the empirical distribution.

The definition of generalized maxent captures many cases of interest including the basic maxent, ℓ_1-regularized maxent and ℓ_2^2-regularized maxent. The basic maxent is obtained by using a point indicator potential $\mathrm{U}^{(0)}(\boldsymbol{u}) = \mathrm{I}_{\{\tilde{\pi}[\boldsymbol{f}]\}}(\boldsymbol{u})$, whereas, as shown in [11], ℓ_1-regularized maxent corresponds to box constraints $|\tilde{\pi}[f_j] - p[f_j]| \le \beta$, which can be represented by $\mathrm{U}^{(1)}(\boldsymbol{u}) = \mathrm{I}_C(\boldsymbol{u})$ where $C = \tilde{\pi}[\boldsymbol{f}] + [-\beta, \beta]^n$. Finally, as pointed out in [6, 7], ℓ_2^2-regularized maxent is obtained using the potential $\mathrm{U}^{(2)}(\boldsymbol{u}) = \|\tilde{\pi}[\boldsymbol{f}] - \boldsymbol{u}\|_2^2/(2\alpha)$ which incurs an ℓ_2^2-style penalty for deviating from empirical averages.

To simplify the exposition, we use the notation $\mathrm{U}_{p_0}(\boldsymbol{u}) = \mathrm{U}(p_0[\boldsymbol{f}] - \boldsymbol{u})$ for a potential centered at p_0. Thus the basic maxent potential $\mathrm{U}^{(0)}(\boldsymbol{u}) = \mathrm{I}_{\{\tilde{\pi}[\boldsymbol{f}]\}}(\boldsymbol{u})$ could have been specified by defining $\mathrm{U}_{\tilde{\pi}}^{(0)}(\boldsymbol{u}) = \mathrm{I}_{\{\boldsymbol{0}\}}(\boldsymbol{u})$ and similarly the box potential by defining $\mathrm{U}_{\tilde{\pi}}^{(1)}(\boldsymbol{u}) = \mathrm{I}_{[-\beta,\beta]^n}(\boldsymbol{u})$ and the ℓ_2^2 penalty by defining $\mathrm{U}_{\tilde{\pi}}^{(2)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_2^2/(2\alpha)$.

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

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

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 $\boldsymbol{F}_{jx} = f_j(x)$ and use Fenchel's duality:

$$\begin{aligned}
\min_{p \in \Delta} \left[\mathrm{D}(p \parallel q_0) + \mathrm{U}(p[\boldsymbol{f}])\right] &= \min_{p \in \Delta} \left[\mathrm{D}(p \parallel q_0) + \mathrm{U}(\boldsymbol{F}p)\right] \\
&= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left[-\ln\left(\textstyle\sum_{x \in \mathcal{X}} q_0(x) \exp\bigl\{(\boldsymbol{F}^\top \boldsymbol{\lambda})_x\bigr\}\right) - \mathrm{U}^*(-\boldsymbol{\lambda})\right] &\quad (4) \\
&= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left[-\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}^*(-\boldsymbol{\lambda})\right]. &\quad (5)
\end{aligned}$$

In Eq. (4), we apply Theorem 1. We use $(\boldsymbol{F}^\top \boldsymbol{\lambda})_x$ to denote the entry of $\boldsymbol{F}^\top \boldsymbol{\lambda}$ indexed by x. In Eq. (5), we note that $(\boldsymbol{F}^\top \boldsymbol{\lambda})_x = \boldsymbol{\lambda} \cdot \boldsymbol{f}(x)$ and thus the expression inside the logarithm equals the normalization constant of $q_{\boldsymbol{\lambda}}$. The dual objective will be referred to as Q:

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

We could rewrite Q in terms of the conjugate of a centered potential. By Eq. (2)

$$\mathrm{U}_{p_0}^*(\boldsymbol{\lambda}) = \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot p_0[\boldsymbol{f}], \qquad (6)$$

hence the dual objective can be rewritten as

$$Q(\boldsymbol{\lambda}) = -\mathrm{D}(p_0 \parallel q_{\boldsymbol{\lambda}}) + \mathrm{D}(p_0 \parallel q_0) - \mathrm{U}^*_{p_0}(\boldsymbol{\lambda}).$$

For any fixed distribution p_0, $\mathrm{D}(p_0 \parallel q_0)$ is a finite constant, so maximizing $Q(\boldsymbol{\lambda})$ is equivalent to minimizing $\mathrm{D}(p_0 \parallel q_{\boldsymbol{\lambda}}) + \mathrm{U}^*_{p_0}(\boldsymbol{\lambda})$, or $\mathrm{L}_{p_0}(\boldsymbol{\lambda}) + \mathrm{U}^*_{p_0}(\boldsymbol{\lambda})$. Using $p_0 = \tilde{\pi}$ we obtain a dual analogous to $\mathcal{Q}_{\text{basic}}$:

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

Note that a minimizing $\boldsymbol{\lambda}$ does not depend on a particular choice of p_0. In particular, a minimizer of $\mathcal{Q}$ is also a 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 $\mathcal{Q}$ has two terms. The first of them is the empirical log loss and the second one can be viewed as a regularization term penalizing "complex" solutions. From a Bayesian perspective, $\mathrm{U}^*_{\tilde{\pi}}$ corresponds to negative log of the prior. Thus, minimizing $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda})$ is equivalent to maximizing the posterior.

In case of the basic maxent, we obtain $\mathrm{U}^{(0)*}_{\tilde{\pi}}(\boldsymbol{\lambda}) = \mathrm{I}^*_{\{\mathbf{0}\}}(\boldsymbol{\lambda}) = 0$ and thus recover the basic dual. For the box potential, we obtain $\mathrm{U}^{(1)*}_{\tilde{\pi}}(\boldsymbol{\lambda}) = \mathrm{I}^*_{[-\beta,\beta]^n}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_1$ which corresponds to an ℓ_1-style regularization and a Laplace prior. For the ℓ_2^2 potential, we obtain $\mathrm{U}^{(2)*}_{\tilde{\pi}}(\boldsymbol{\lambda}) = \alpha\|\boldsymbol{\lambda}\|_2^2/2$ which corresponds to an ℓ_2^2-style regularization and a Gaussian prior.

Results of this section are summarized in the following theorem:

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

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

Moreover, if $\lim_{t\to\infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^n} Q(\boldsymbol{\lambda})$ *then the sequence of* $q_t = q_{\boldsymbol{\lambda}_t}$ *has a limit and*

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

Sketch of proof. Eq. (i) is a consequence of Fenchel's duality as was shown earlier. It remains to prove Eq. (ii). Let p_0 be the minimizer of P. Centering primal and dual objectives at p_0, we obtain by Eq. (i) and the assumption

$$\mathrm{D}(p_0 \parallel q_0) + \mathrm{U}_{p_0}(\mathbf{0}) = \lim_{t\to\infty} \big[-\mathrm{D}(p_0 \parallel q_t) + \mathrm{D}(p_0 \parallel q_0) - \mathrm{U}^*_{p_0}(\boldsymbol{\lambda}_t)\big].$$

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

$$\mathrm{U}_{p_0}(\mathbf{0}) + \mathrm{U}^*_{p_0}(\boldsymbol{\lambda}_t) = -\mathrm{D}(p_0 \parallel q_t) + o(1).$$

The left-hand side is by Fenchel's inequality nonnegative, so $\mathrm{D}(p_0 \parallel q_t) \to 0$ by property (B1). Therefore, by property (B2), every convergent subsequence of $q_1, q_2, \ldots$ has the limit p_0. Since the q_t's come from the compact set Δ, we obtain $q_t \to p_0$. □

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 the algorithm in Section 6.

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 $\mathrm{L}_\pi(\boldsymbol{\lambda}^*)$ when $q_{\hat{\boldsymbol{\lambda}}}$ maximizes the dual objective Q and $q_{\boldsymbol{\lambda}^*}$ is any Gibbs distribution. In particular, bounds hold for the Gibbs distribution minimizing the true loss. 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 $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}^*})$.

The crux of our method is the lemma below. Even though its proof is remarkably simple, it is sufficiently general to cover all cases of interest.

Lemma 1. *Let $\hat{\boldsymbol{\lambda}}$ maximize Q. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^*}$*

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

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

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

$$\begin{aligned}\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + \mathrm{U}^*_\pi(\boldsymbol{\lambda}^*) - \mathrm{U}^*_\pi(\hat{\boldsymbol{\lambda}}) \\ &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + (\boldsymbol{\lambda}^* - \hat{\boldsymbol{\lambda}}) \cdot \pi[\boldsymbol{f}] + \mathrm{U}^*(-\boldsymbol{\lambda}^*) - \mathrm{U}^*(-\hat{\boldsymbol{\lambda}}). \end{aligned} \tag{7}$$

Eq. (7) follows from Eq. (6). Now Eq. (i) can be obtained by applying Fenchel's inequality to the second term of Eq. (7):

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

Eq. (ii) also follows from (7) by centering the conjugate potential at $\tilde{\pi}$. □

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 this case, the right hand side of Lemma 1.i will be infinite unless $\pi[\boldsymbol{f}]$ lies in C. In order to apply Lemma 1.i, we ensure that $\pi[\boldsymbol{f}] \in C$ with high probability. Therefore, we choose C as a confidence region for $\pi[\boldsymbol{f}]$. If $\pi[\boldsymbol{f}] \in C$ then for any Gibbs distribution $q_{\boldsymbol{\lambda}^*}$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + \mathrm{I}^*_C(\boldsymbol{\lambda}^*) + \mathrm{I}^*_C(-\boldsymbol{\lambda}^*). \tag{8}$$

For a fixed $\boldsymbol{\lambda}^*$ and non-empty C, $\mathrm{I}^*_C(\boldsymbol{\lambda}^*) + \mathrm{I}^*_C(-\boldsymbol{\lambda}^*)$ is always nonnegative and proportional to the size of C's projection onto a line parallel with $\boldsymbol{\lambda}^*$. 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 Eq. (8). We assume that confidence regions are obtained by scaling some symmetric prototype C_0 and shifting it to empirical averages.

Theorem 3. *Assume that $\tilde{\pi}[\boldsymbol{f}] - \pi[\boldsymbol{f}] \in \beta C_0$ where C_0 is a closed convex set symmetric around the origin, $\beta > 0$ and βC_0 denotes $\{\beta \boldsymbol{u} \mid \boldsymbol{u} \in C_0\}$. Let $\hat{\boldsymbol{\lambda}}$ minimize the regularized log loss $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta \mathrm{I}^*_{C_0}(\boldsymbol{\lambda})$. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^*}$*

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

5.1 Maxent with Polyhedral Regularization

We now apply the foregoing general results to some specific cases of interest. To begin, we consider potentials which are indicator functions of polytopes. The simplest case is the box indicator $\mathrm{U}^{(1)}$, for which Dudík, Phillips and Schapire [11] give generalization bounds. However, when additional knowledge about structure of the feature space is available or when samples are biased, other polytopes yield tighter confidence regions and hence better performance guarantees, as we now show.

Feature Space Derived Potential. When values of $\boldsymbol{f}(x)$ lie inside a polytope with a possibly very large number of facets then a symmetrized version of this polytope can be used as a prototype for the confidence region. For example, suppose that values $\boldsymbol{f}(x)$ lie inside the polytope $\{\boldsymbol{u} \mid a_{\bar{\jmath}} \leq \boldsymbol{\mu}_{\bar{\jmath}} \cdot \boldsymbol{u} \leq b_{\bar{\jmath}}$ for $\bar{\jmath} = 1, \ldots, \bar{n}\}$ where $\boldsymbol{\mu}_{\bar{\jmath}} \in \mathbb{R}^n, a_{\bar{\jmath}}, b_{\bar{\jmath}} \in \mathbb{R}$ are constants. Then the following holds:

Theorem 4. *Let $\boldsymbol{\mu}_{\bar{\jmath}}, a_{\bar{\jmath}}, b_{\bar{\jmath}}$ be as above. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta \mathrm{I}^*_{C_0}(\boldsymbol{\lambda})$ with $\beta = \sqrt{\ln(2\bar{n}/\delta)/(2m)}$ and $C_0 = \{\boldsymbol{u} \mid |\boldsymbol{\mu}_{\bar{\jmath}} \cdot \boldsymbol{u}| \leq b_{\bar{\jmath}} - a_{\bar{\jmath}}$ for all $\bar{\jmath}\}$. Then with probability at least $1 - \delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^*}$,*

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_{\pi}(\boldsymbol{\lambda}^*) + \mathrm{I}^*_{C_0}(\boldsymbol{\lambda})\sqrt{2\ln(2\bar{n}/\delta)/m}.$$

Proof. By Hoeffding's inequality, for a fixed $\bar{\jmath}$, the probability that $|\boldsymbol{\mu}_{\bar{\jmath}} \cdot (\tilde{\pi}[\boldsymbol{f}] - \pi[\boldsymbol{f}])|$ exceeds $\beta(b_{\bar{\jmath}} - a_{\bar{\jmath}})$ is at most $2e^{-2\beta^2 m} = \delta/\bar{n}$. By the union bound, the probability of this happening for any $\bar{\jmath}$ is at most δ. Thus, $\tilde{\pi}[\boldsymbol{f}] - \pi[\boldsymbol{f}] \in \beta C_0$ with probability at least $1 - \delta$ and the claim follows from Theorem 3. □

This performance bound decreases as $1/\sqrt{m}$ with an increasing number of samples and grows only logarithmically with the number of facets of the bounding polytope. Thus, bounding polytopes can have a very large number of facets and still yield good bounds for moderate sample sizes. When deciding between several polytopes based on this bound, the increase in the number of facets should be weighed against the decrease in the regularization $\mathrm{I}^*_{C_0}$ as will be demonstrated in examples below.

Means and Variances. As a specific application, consider a set of $n = 2K$ features indexed as f_k, f_{kk}, $1 \leq k \leq K$, such that $0 \leq f_k(x) \leq 1$ and $f_{kk}(x) = f_k^2(x)$. Note that the best Gibbs distribution is the one that matches f_k's true means and variances. These types of features were successfully used with box constraints in habitat modeling [4]. Box constraints yield the guarantee

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_{\pi}(\boldsymbol{\lambda}^*) + \|\boldsymbol{\lambda}^*\|_1\sqrt{2\ln(4K/\delta)/m}.$$

Noting that $t - 1/4 \leq t^2 \leq t$ for $t \in [0, 1]$, it is possible to obtain a tighter polytope

$$C_0 = \{\boldsymbol{u} \mid |\boldsymbol{u}_k| \leq 1, |\boldsymbol{u}_{kk}| \leq 1, |\boldsymbol{u}_k - \boldsymbol{u}_{kk}| \leq 1/4 \text{ for all } k\}$$

and the guarantee

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_{\pi}(\boldsymbol{\lambda}^*) + \mathrm{I}^*_{C_0}(\boldsymbol{\lambda}^*)\sqrt{2\ln(6K/\delta)/m}.$$

In this case, it is possible to derive $\mathrm{I}^*_{C_0}$ explicitly:

$$\mathrm{I}^*_{C_0}(\boldsymbol{\lambda}) = \textstyle\sum_k (7|\lambda_k + \lambda_{kk}| + |\lambda_k| + |\lambda_{kk}|)/8.$$

Note that $\mathrm{I}^*_{C_0}(\boldsymbol{\lambda})$ may be up to eight times smaller than $\|\boldsymbol{\lambda}\|_1$ while the relative increase of the bound due to an increase in $\bar{n}$ is close to 1 for moderate sizes of K. Thus, the bound may decrease up to eight times for moderate K. Such improvement would require a 64-fold increase in the number of training samples using ℓ_1-regularization.

Means, Variances and Covariances. In this example, we expand the feature set to include also covariance terms $f_{kl}(x) = f_k(x)f_l(x)$ where $1 \le k < l \le K$. In this case the box can be restricted to a much tighter set

$$\begin{aligned} C_0 = \{\boldsymbol{u} \mid\ & |\boldsymbol{u}_k| \le 1, |\boldsymbol{u}_{kk}| \le 1, |\boldsymbol{u}_k - \boldsymbol{u}_{kk}| \le 1/4 \text{ for all } k, \\ & |\boldsymbol{u}_{kl}| \le 1, |\boldsymbol{u}_k - \boldsymbol{u}_{kl}| \le 1, |\boldsymbol{u}_l - \boldsymbol{u}_{kl}| \le 1, \\ & |\boldsymbol{u}_k + \boldsymbol{u}_l - 2\boldsymbol{u}_{kl}| \le 1, |\boldsymbol{u}_{kk} + \boldsymbol{u}_{ll} - 2\boldsymbol{u}_{kl}| \le 1 \text{ for all } k < l\}. \end{aligned}$$

Note that $\bar{n}$ increases approximately fivefold from $K(K+3)/2$ to $5K(K+1/5)/2$ resulting in only slight relative increase of the bound for moderate K. This is outweighed by the decrease of the bound due to a tighter confidence region.

Debiased Potential. In previous work [13], we considered the problem of using maxent when the data was sampled in a biased manner. Here we show how superior bounds can be obtained using our generalized maxent framework.

In previous examples, confidence regions were symmetric sets centered at empirical averages of features. Here, however, asymmetric regions are more appropriate. We assume now that samples do not come from the target distribution π, but from the *biased distribution* πs, where $s \in \Delta$ is the *sampling distribution* and $\pi s(x) = \pi(x)s(x)/\sum_{x' \in \mathcal{X}} \pi(x')s(x')$ corresponds to the probability of observing x given that it is sampled by both π and s. We assume that s is known and strictly positive. Further, let $s_{\min} = \min_x s(x)$ and $s_{\max} = \max_x s(x)$. We use the following theorem to derive confidence intervals of true feature expectations from biased samples.

Theorem 5 (Theorem 2 of [13]). $\pi s[\boldsymbol{f}/s]/\pi s[1/s] = \pi[\boldsymbol{f}]$.

Earlier [13] we derived, for example by Hoeffding's inequality, confidence intervals $[c_j, d_j]$ for $\pi s[f_j/s]$ and an interval $[c_0, d_0]$ for $\pi s[1/s]$. We converted these into box constraints for $\pi[\boldsymbol{f}]$ by Theorem 5. However, Theorem 5 can also be used to obtain a tighter confidence region:

$$\begin{aligned} C &= \bigcup_{c_0 \le t \le d_0} \{\boldsymbol{u} \mid c_j/t \le u_j \le d_j/t \text{ for all } j\} \\ &= \underset{t=c_0,d_0}{\text{convex hull}} \{\boldsymbol{u} \mid c_j/t \le u_j \le d_j/t \text{ for all } j\}. \end{aligned} \tag{9}$$

Eq. (1) can be used to obtain an explicit form of I^*_C. Working out Hoeffding's bounds, applying Lemma 1.i and converting $\mathrm{I}^*_C(\boldsymbol{\lambda}) + \mathrm{I}^*_C(-\boldsymbol{\lambda})$ into a sample independent form under the assumption that terms on the left-hand side of Theorem 5 lie in their confidence intervals, we obtain the following guarantee (the proof is omitted):

Theorem 6. *Assume that features $f_1, \ldots, f_n$ are bounded in $[0,1]$. Let s be as above and let $\widetilde{\pi s}$ denote the empirical distribution of samples drawn from πs. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\ln Z_{\boldsymbol{\lambda}} + \mathrm{I}^*_C(-\boldsymbol{\lambda})$ where*

$$\mathrm{I}^*_C(-\boldsymbol{\lambda}) = \max_{t=c_0,d_0} \left[\frac{-\boldsymbol{\lambda} \cdot \widetilde{\pi s}[\boldsymbol{f}/s] + \beta \|\boldsymbol{\lambda}\|_1}{t} \right]$$

with $\beta = \sqrt{\ln(2(n+1)/\delta)/(2m)}/s_{\min}$, $c_0 = \max\{1/s_{\max},\ \widetilde{\pi s}[1/s] - \beta\}$, $d_0 = \widetilde{\pi s}[1/s] + \beta$. Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^*}$,

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + \frac{\|\boldsymbol{\lambda}^*\|_1 + |\boldsymbol{\lambda}^* \cdot \pi[\boldsymbol{f}]|}{\sqrt{m}} \cdot \frac{\pi[s]}{s_{\min}} \cdot \left(\alpha + \alpha^2 \frac{s_{\max}}{s_{\min}\sqrt{m}} \right) \tag{10}$$

where $\alpha = \sqrt{2\ln(2(n+1)/\delta)}$.

This bound shares many of the favorable properties of the bound of Theorem 4. In particular, it decreases as the square root of the number of samples and grows only logarithmically with the number of features. It also increases with the level of correlation between the sampling and target distributions as measured by the ratio $\pi[s]/s_{\min}$. Intuitively, this dependence should not be surprising, because high values of $\pi[s]/s_{\min}$ mean that π puts more weight on points with larger bias. As a result, it is more difficult to disambiguate effects of s and π on the sampling process.

When using box constraints, as in [13], we obtain an analogous bound with the term $|\boldsymbol{\lambda}^* \cdot \pi[\boldsymbol{f}]|$ in Eq. (10) replaced by a larger term $\sum_j |\lambda^*_j| \pi[f_j]$. Improvement in the guarantee due to the new regularization will be the most significant when $\boldsymbol{\lambda}^*$ and $\pi[\boldsymbol{f}]$ are close to orthogonal. This is true for almost all directions of $\boldsymbol{\lambda}^*$ as the dimension of the feature space increases.

5.2 Maxent with ℓ_2-regularization

In some cases, tighter performance guarantees can be obtained by using non-polyhedral confidence regions. In this section we consider confidence regions which take the shape of a Euclidean ball. We use an ℓ_2 version of Hoeffding's inequality and apply Theorem 3 to obtain performance guarantees (the proof is omitted).

Theorem 7. *Let $D_2 = \sup_{x,x' \in \mathcal{X}} \|\boldsymbol{f}(x) - \boldsymbol{f}(x')\|_2$ be the ℓ_2 diameter of $\boldsymbol{f}(\mathcal{X})$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\|\boldsymbol{\lambda}\|_2$ with $\beta = D_2\big[1 + (2+\sqrt{2})\sqrt{\ln(1/\delta)}\big]/\sqrt{2m}$. Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^*}$,*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^*) + \|\boldsymbol{\lambda}^*\|_2 D_2 \big[\sqrt{2} + 2(1+\sqrt{2})\sqrt{\ln(1/\delta)}\big]/\sqrt{m}.$$

Unlike results of the previous section, 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 by Theorem 4:

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

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 D_2 may increase as $\Omega(\sqrt{n})$. Thus the guarantee for ℓ_2-regularized maxent 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 [18].

5.3 Maxent with ℓ_2^2-regularization

So far we have considered potentials that take the form of an indicator function. In this section we present a result for the ℓ_2^2 potential $\mathrm{U}_{\tilde{\pi}}^{(2)}(\boldsymbol{u}) = \|\boldsymbol{u}\|_2^2/(2\alpha)$ which grows continuously with increasing distance from empirical averages. In addition to probabilistic guarantees (which we do not discuss in this section), it is possible to derive guarantees on the expected performance. However, these guarantees require an *a priori* bound on $\|\boldsymbol{\lambda}^\star\|_2$ and thus are not entirely uniform.

Expectation guarantees can be simply obtained by taking an expectation in Lemma 1.i and bounding the trace of the feature covariance matrix by $D_2^2/2$. Instead, we use a stability bound on $\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}\|_2$ along the lines of [8], then apply Lemma 1.ii and only then bound the trace. This results in tighter guarantees (also tighter than those in [8]). Optimizing α under the condition $\|\boldsymbol{\lambda}^\star\|_2 \le L_2$ then yields the following:

Theorem 8. *Let D_2 be the ℓ_2 diameter of $\boldsymbol{f}(\mathcal{X})$ and let $L_2 > 0$. Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ with $\alpha = D_2/(L_2\sqrt{m})$. Then for all $\boldsymbol{\lambda}^\star$ such that $\|\boldsymbol{\lambda}^\star\| \le L_2$*

$$\mathrm{E}\big[\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})\big] \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + L_2 D_2/\sqrt{m}.$$

Expectation guarantees can also be obtained for regularization types of the form $\mathrm{U}_{\tilde{\pi}}^*(\boldsymbol{\lambda}) = \beta \mathrm{I}_{C_0}^*(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2$. Using that $\mathrm{U}_{\tilde{\pi}}(\boldsymbol{u}) \le \min\{\mathrm{I}_{\beta C_0}(\boldsymbol{u}), \|\boldsymbol{u}\|_2^2/(2\alpha)\}$, the expectation is derived by distinguishing whether $\tilde{\pi}[\boldsymbol{f}] - \pi[\boldsymbol{f}]$ lies in βC_0 or not (with a small probability δ). The resulting guarantee contains one term proportional to $\mathrm{I}_{C_0}^*(\boldsymbol{\lambda}^\star)/\sqrt{m}$ and another proportional to $L_2 D_2/\sqrt{m}$ with δ controlling the tradeoff.

6 A Sequential-update Algorithm and Convergence Proof

In this section, we present an algorithm for the generalized maxent and proof of convergence. The algorithm covers a wide class of potentials including the basic, box and ℓ_2^2 potential. Polyhedral and ℓ_2-ball potentials do not fall in this class, but the corresponding maxent problems can be transformed and our algorithm can still be applied.

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. We assume that the potential U is *decomposable* in the sense that it can be written as a sum of coordinate potentials $\mathrm{U}(\boldsymbol{u}) = \sum_j \mathrm{U}_j(u_j)$, each of which is a closed proper convex functions bounded from below. The conjugate potential U^* then equals the sum of conjugate coordinate potentials U_j^* (see Eq. (3)) and $\mathrm{U}_j^*(0) = \sup_{u_j}\left[-\mathrm{U}_j(u_j)\right]$ is finite for all j.

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 $\{1\}$.

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. By decomposability and convexity, we can bound the change in the objective (analogously to [11]):

$$Q(\boldsymbol{\lambda}') - Q(\boldsymbol{\lambda}) \geq -\ln\bigl(1 + (e^{\delta} - 1) q_{\boldsymbol{\lambda}}[f_j]\bigr) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j). \qquad (11)$$

Our algorithm starts with $\boldsymbol{\lambda}_1 = \mathbf{0}$ and then, on each iteration, maximizes this lower bound over all choices of (j, δ). For the maximizing j, it adds the corresponding δ to λ_j. This is repeated until convergence. We assume that for each j the maximizing δ is finite. This will be the case if the potential and features are non-degenerate.

For maxent with box constraints, the minimizing δ can be derived explicitly yielding the algorithm of [11]. For a general potential note that (11) is strictly concave in δ so we can use any of a number of search methods to find the optimal δ.

Reductions from Non-decomposable Potentials. Polyhedral and ℓ_2-ball potentials are not decomposable. When a polyhedral potential is represented as an intersection of halfspaces $\boldsymbol{\mu}_{\bar{\jmath}} \cdot \boldsymbol{u} \geq a_{\bar{\jmath}}$, it suffices to use transformed features $\bar{f}_{\bar{\jmath}}(x) = \boldsymbol{\mu}_{\bar{\jmath}} \cdot \boldsymbol{f}(x)$ with coordinate potentials corresponding to inequality constraints. Note that the debiased potential polytope (9) is not described in this form. However, it is not too difficult to obtain such a representation. It turns out that this representation uses $O(n^2)$ halfspaces and is thus polynomial in the original problem size.

In case of an ℓ_2-ball potential, we replace the constraint $\|\tilde{\pi}[\boldsymbol{f}] - p[\boldsymbol{f}]\|_2 \leq \beta$ by $\|\tilde{\pi}[\boldsymbol{f}] - p[\boldsymbol{f}]\|_2^2 \leq \beta^2$ which yields an equivalent primal $\mathcal{P}'$. If $\beta > 0$ then, by the Lagrange duality and Slater's conditions [21], we know there exists $\mu \geq 0$ such that the solution of $\mathcal{P}'$ is the same as the solution of

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

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 our algorithm to solve an instance of ℓ_2^2-regularized maxent in each iteration.

Convergence. In order to prove convergence of our algorithm, 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 at $\boldsymbol{\lambda}$:

$$P(q_{\boldsymbol{\lambda}}) - Q(\boldsymbol{\lambda}) = \mathrm{U}(q_{\boldsymbol{\lambda}}[\boldsymbol{f}]) + \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot q_{\boldsymbol{\lambda}}[\boldsymbol{f}].$$

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

However, for many potentials of interest, including equality and inequality constraints, the difference between primal and dual may remain infinite throughout the computation. Therefore, we introduce an *auxiliary function*, defined, somewhat non-standardly, as a surrogate for this difference.

Definition 1. *A function* $A : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, \infty]$ *is called an* auxiliary function *if*

$$A(\boldsymbol{\lambda}, \boldsymbol{a}) = \mathrm{U}(\boldsymbol{a}) + \mathrm{U}^*(-\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot \boldsymbol{a} + \mathrm{B}(\boldsymbol{a} \parallel q_{\boldsymbol{\lambda}}[\boldsymbol{f}]) \tag{12}$$

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

Unlike the previous applications of auxiliary functions [3, 14], we do not assume that $A(\boldsymbol{\lambda}, \boldsymbol{a})$ bounds a change in the dual objective and we also make no continuity assumptions. However, an auxiliary function is always non-negative since by Fenchel's inequality $\mathrm{U}(\boldsymbol{a}) + \mathrm{U}^*(-\boldsymbol{\lambda}) \geq -\boldsymbol{\lambda} \cdot \boldsymbol{a}$ and hence $A(\boldsymbol{\lambda}, \boldsymbol{a}) \geq \mathrm{B}(\boldsymbol{a} \parallel q_{\boldsymbol{\lambda}}[\boldsymbol{f}]) \geq 0$. Moreover, if $A(\boldsymbol{\lambda}, \boldsymbol{a}) = 0$ then $q_{\boldsymbol{\lambda}}[\boldsymbol{f}] = \boldsymbol{a}$ and $A(\boldsymbol{\lambda}, \boldsymbol{a}) = P(q_{\boldsymbol{\lambda}}) - Q(\boldsymbol{\lambda}) = 0$, i.e. by maxent duality, $q_{\boldsymbol{\lambda}}$ solves the primal and $\boldsymbol{\lambda}$ solves the dual.

It turns out, as we show in Lemma 3 below, that the optimality property generalizes to the case when $A(\boldsymbol{\lambda}_t, \boldsymbol{a}_t) \to 0$ provided that $Q(\boldsymbol{\lambda}_t)$ has a finite limit. In particular, it suffices to find a suitable sequence of $\boldsymbol{a}_t$'s for $\boldsymbol{\lambda}_t$'s produced by our algorithm to show its convergence. Note that the optimality in the limit trivially holds when $\boldsymbol{\lambda}_t$'s and $\boldsymbol{a}_t$'s come from a compact set, because $A(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{a}}) = 0$ at a cluster point of $\{(\boldsymbol{\lambda}_t, \boldsymbol{a}_t)\}$ by the lower semi-continuity of U and U^*.

In the general case, we follow the technique used by [3] 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 case of the basic maxent, $A(\boldsymbol{\lambda}, \boldsymbol{a}) = \mathrm{B}(\tilde{\pi}[\boldsymbol{f}] \parallel q_{\boldsymbol{\lambda}}[\boldsymbol{f}])$ 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 2 (the proof is omitted).

Lemma 2. *Let* U_{p_0} *be a decomposable potential centered at a feasible point* p_0. *Let* $S = \operatorname{dom} \mathrm{U}_{p_0} = \{\boldsymbol{u} \in \mathbb{R}^n \mid \mathrm{U}_{p_0}(\boldsymbol{u}) < \infty\}$ *and* $T_c = \{\boldsymbol{\lambda} \in \mathbb{R}^n \mid \mathrm{U}^*_{p_0}(\boldsymbol{\lambda}) \leq c\}$. *Then there exists* $\alpha_c \geq 0$ *such that* $\boldsymbol{\lambda} \cdot \boldsymbol{u} \leq \alpha_c \|\boldsymbol{u}\|_1$ *for all* $\boldsymbol{u} \in S, \boldsymbol{\lambda} \in T_c$.

Lemma 3. *Let* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots \in \mathbb{R}^n$, $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots \in \mathbb{R}^n$ *be sequences such that* $Q(\boldsymbol{\lambda}_t)$ *has a finite limit and* $A(\boldsymbol{\lambda}_t, \boldsymbol{a}_t) \to 0$ *as* $t \to \infty$. *Then* $\lim_{t\to\infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda})$.

Sketch of proof. Let q_t denote $q_{\boldsymbol{\lambda}_t}$. Consider a convergent subsequence of q_t's, index it by τ and denote its limit by $\hat{q}$. As noted earlier, $A(\boldsymbol{\lambda}, \boldsymbol{a}) \geq \mathrm{B}(\boldsymbol{a} \parallel q_{\boldsymbol{\lambda}}[\boldsymbol{f}])$. Since $A(\boldsymbol{\lambda}_\tau, \boldsymbol{a}_\tau) \to 0$, we obtain that $\mathrm{B}(\boldsymbol{a}_\tau \parallel q_\tau[\boldsymbol{f}]) \to 0$ and thus $\boldsymbol{a}_\tau \to \hat{q}[\boldsymbol{f}]$. Rewrite Eq. (12) in terms of potentials and conjugate potentials centered at an arbitrary feasible point p_0 (which must exist by assumption), denoting terms with zero limits by $o(1)$:

$$\mathrm{U}_{p_0}(p_0[\boldsymbol{f}] - \boldsymbol{a}_\tau) = -\mathrm{U}^*_{p_0}(\boldsymbol{\lambda}_\tau) + \boldsymbol{\lambda}_\tau \cdot (p_0[\boldsymbol{f}] - \boldsymbol{a}_\tau) + o(1). \tag{13}$$

We use Eq. (13) to show first the feasibility and then the optimality of $\hat{q}$.

Feasibility. We bound the right hand side of Eq. (13). The first term $-\mathrm{U}^*_{p_0}(\boldsymbol{\lambda}_\tau)$ is, by Fenchel's inequality, bounded above by $\mathrm{U}_{p_0}(\mathbf{0})$. The second term $\boldsymbol{\lambda}_\tau \cdot (p_0[\boldsymbol{f}] - \boldsymbol{a}_\tau)$ can be bounded above by Lemma 2. Taking limits yields feasibility.

Optimality. Since $\hat{q}$ is feasible, we can set p_0 equal to $\hat{q}$ in Eq. (13). Using Lemma 2 and taking limits we obtain that $\mathrm{U}_{\hat{q}}(\mathbf{0}) \leq \lim_{\tau\to\infty}[-\mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau)]$. Adding $\mathrm{D}(\hat{q} \parallel q_0)$ to both sides yields $P(\hat{q}) \leq \lim_{\tau\to\infty} Q(\boldsymbol{\lambda}_\tau)$ which by the maxent duality implies the optimality of $\hat{q}$. □

Theorem 9. *The sequential-update algorithm produces a sequence* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ *for which* $\lim_{t\to\infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda})$.

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

Note that $Q(\boldsymbol{\lambda}_1) = Q(\mathbf{0}) = -\mathrm{U}^*(\mathbf{0})$ is finite by decomposability and Q is bounded above by feasibility of the primal. For each j let $F_{t,j}$ denote the maximum over δ of the lower bound (11) in step t. Note that $F_{t,j}$ is nonnegative since the bound is zero when $\delta = 0$. Thus $Q(\boldsymbol{\lambda}_t)$ is nondecreasing and hence has a finite limit.

In each step, $Q(\boldsymbol{\lambda}_{t+1}) - Q(\boldsymbol{\lambda}_t) \geq F_{t,j} \geq 0$. Since $Q(\boldsymbol{\lambda}_t)$ has a finite limit, we obtain $F_{t,j} \to 0$. We will use $F_{t,j}$ to construct A. Rewrite $F_{t,j}$ using Fenchel's duality:

$$\begin{aligned}
F_{t,j} &= \max_{\delta}\Big[-\ln(1 + (e^{\delta} - 1)q_t[f_j]) - \mathrm{U}_j^*(-\lambda_{t,j} - \delta) + \mathrm{U}_j^*(-\lambda_{t,j})\Big] \\
&= \max_{\delta}\Big[-\ln\big\{(1 - q_t[f_j])e^{0\cdot\delta} + q_t[f_j]e^{1\cdot\delta}\big\} - \mathrm{U}_j'^*(-\delta)\Big] + \mathrm{U}_j^*(-\lambda_{t,j}) && (14)\\
&= \min_{\bar{a},a}\Big[\mathrm{D}\big((\bar{a}, a) \parallel (1 - q_t[f_j], q_t[f_j])\big) + \mathrm{U}_j'(0\cdot\bar{a} + 1\cdot a)\Big] + \mathrm{U}_j^*(-\lambda_{t,j}) && (15)\\
&= \min_{0\leq a\leq 1}\Big[\mathrm{D}(a \parallel q_t[f_j]) + \mathrm{U}_j(a) + a\cdot\lambda_{t,j}\Big] + \mathrm{U}_j^*(-\lambda_{t,j}). && (16)
\end{aligned}$$

In Eq. (14), we write $\mathrm{U}_j'^*(u)$ for $\mathrm{U}_j^*(u - \lambda_{t,j})$. In Eq. (15), we applied Theorem 1, noting that the conjugate of the log partition function is the relative entropy. The value of relative entropy $\mathrm{D}((\bar{a}, a) \parallel (1 - q_t[f_j], q_t[f_j]))$ is infinite whenever $(\bar{a}, a)$ is not a probability distribution, so it suffices to consider pairs where $0 \leq a \leq 1$ and $\bar{a} = 1 - a$. In Eq. (16), we use $\mathrm{D}(a \parallel q_t[f_j])$ as a shorthand for $\mathrm{D}((1 - a, a) \parallel (1 - q_t[f_j], q_t[f_j]))$ and use Eq. (2) to convert U_j' into U_j.

The minimum in Eq. (16) is always attained because a comes from a compact set. Let $a_{t,j}$ denote a value attaining this minimum. We define the auxiliary function $A(\boldsymbol{\lambda}, \boldsymbol{a})$ as the sum over j of Eq. (16) (evaluated at $a = a_j$ and with $\lambda_{t,j}$ replaced by λ_j). Now $A(\boldsymbol{\lambda}_t, \boldsymbol{a}_t) = \sum_j F_{t,j} \to 0$ and the result follows by Lemma 3. □

7 Conclusion and Future Work

In this work, we have explored one direction of generalizing maxent: replacing equality constraints in the primal by an arbitrary convex potential or, equivalently, adding a convex regularization term to the maximum likelihood estimation in the dual. In our unified approach, we derived performance guarantees for many existing and novel regularization types and presented an algorithm covering a wide range of potentials.

As the next step, we would like to explore whether theoretical superiority of the new regularization types results in improved performance on real-world data. If this turns out to be the case, we would like to investigate strategies for obtaining tighter confidence regions and hence better performing regularizations using sample-derived statistics or properties of the feature space.

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 [12, 14]. Analogous duality results as well as a modified algorithm 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), boosting or linear regression. In the future, we would like to generalize performance guarantees to this setting.

Finally, the convergence rate of the present algorithm and a possible tradeoff between statistical guarantees and computational efficiency of different regularizations is open for future research.

Acknowledgements: We would like to thank Steven Phillips for many helpful discussions. Authors received support through NSF grant CCR-0325463.

References

1. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106** (1957) 620–630
2. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J.: A maximum entropy approach to natural language processing. Computational Linguistics **22**(1) (1996) 39–71
3. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(4) (1997) 1–13
4. Phillips, S.J., Dudík, M., Schapire, R.E.: A ME approach to species distribution modeling. In: Proceedings of the Twenty-First International Conference on Machine Learning. (2004)
5. Lau, R.: Adaptive statistical language modeling. Master's thesis, MIT Department of Electrical Engineering and Computer Science (1994)
6. Chen, S.F., Rosenfeld, R.: A survey of smoothing techniques for ME models. IEEE Transactions on Speech and Audio Processing **8**(1) (2000) 37–50
7. Lebanon, G., Lafferty, J.: Boosting and maximum likelihood for exponential models. Technical Report CMU-CS-01-144, CMU School of Computer Science (2001)
8. Zhang, T.: Class-size independent generalization analysis of some discriminative multi-category classification. In: Advances in Neural Information Processing Systems 17. (2005)
9. Goodman, J.: Exponential priors for maximum entropy models. In: Conference of the North American Chapter of the Association for Computational Linguistics. (2004)
10. Kazama, J., Tsujii, J.: Evaluation and extension of ME models with inequality constraints. In: Conference on Empirical Methods in Natural Language Processing. (2003) 137–144
11. Dudík, M., Phillips, S.J., Schapire, R.E.: Performance guarantees for regularized maximum entropy density estimation. In: COLT 2004. (2004) 472–486
12. Altun, Y., Smola, A.: Unifying divergence minimization and statistical inference via convex duality. In: COLT 2006. (2006)
13. Dudík, M., Schapire, R.E., Phillips, S.J.: Correcting sample selection bias in ME density estimation. In: Advances in Neural Information Processing Systems 18. (2006)
14. Collins, M., Schapire, R.E., Singer, Y.: Logistic regression, AdaBoost and Bregman distances. Machine Learning **48**(1) (2002) 253–285
15. Darroch, J.N., Ratcliff, D.: Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics **43**(5) (1972) 1470–1480
16. Malouf, R.: A comparison of algorithms for maximum entropy parameter estimation. In: Proceedings of the Sixth Conference on Natural Language Learning. (2002) 49–55
17. Krishnapuram, B., Carin, L., Figueiredo, M.A.T., Hartemink, A.J.: Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6) (2005) 957–968
18. Ng, A.Y.: Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In: Proceedings of the Twenty-First International Conference on Machine Learning. (2004)
19. Newman, W.: Extension to the ME method. IEEE Trans. on Inf. Th. **IT-23**(1) (1977) 89–93
20. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970)
21. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)