Structural Maxent Models

Corinna Cortes CORINNA@GOOGLE.COM
Google Research, 111 8th Avenue, New York, NY 10011

Vitaly Kuznetsov VITALY@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012

Mehryar Mohri MOHRI@CIMS.NYU.EDU
Courant Institute and Google Research, 251 Mercer Street, New York, NY 10012

Umar Syed USYED@GOOGLE.COM
Google Research, 111 8th Avenue, New York, NY 10011

Abstract

We present a new class of density estimation models, *Structural Maxent models*, with feature functions selected from a union of possibly very complex sub-families and yet benefiting from strong learning guarantees. The design of our models is based on a new principle supported by uniform convergence bounds and taking into consideration the complexity of the different sub-families composing the full set of features. We prove new data-dependent learning bounds for our models, expressed in terms of the Rademacher complexities of these sub-families. We also prove a duality theorem, which we use to derive our Structural Maxent algorithm. We give a full description of our algorithm, including the details of its derivation, and report the results of several experiments demonstrating that its performance improves on that of existing L_1-norm regularized Maxent algorithms. We further similarly define *conditional Structural Maxent models* for multi-class classification problems. These are conditional probability models also making use of a union of possibly complex feature subfamilies. We prove a duality theorem for these models as well, which reveals their connection with existing binary and multi-class deep boosting algorithms.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

Maximum entropy models, commonly referred to as Maxent models, are density estimation methods used in a variety of tasks in natural language processing (Berger et al., 1996; Rosenfeld, 1996; Pietra et al., 1997; Malouf, 2002; Manning & Klein, 2003; Ratnaparkhi, 2010) and in many other applications, including species habitat modeling (Phillips et al., 2004; 2006; Dudík et al., 2007; Elith et al., 2011).

Maxent models are based on the density estimation principle advocated by Jaynes (1957), which consists of selecting the distribution that is the closest to the uniform distribution or to some other prior, while ensuring that the average value of each feature matches its empirical value. When closeness is measured in terms of the relative entropy, Maxent models are known to coincide with Gibbs distributions, as in the original Boltzmann models in statistical mechanics.

One key benefit of Maxent models is that they allow the use of diverse features that can be selected and augmented by the user, while in some other popular density estimation techniques such as n-gram modeling in language processing, the features are inherently limited. The richness of the features used in many tasks as well as small sample sizes have motivated the use of regularized Maxent models where the L_1-norm (Kazama & Tsujii, 2003) or the L_2-norm (Chen & Rosenfeld, 2000; Lebanon & Lafferty, 2001) of the parameter vector defining the Gibbs distribution is controlled. This can be shown to be equivalent to the introduction of a Laplacian or Gaussian prior over the parameter vector in a Bayesian interpretation (Williams, 1994; Goodman, 2004). Group sparsity regularizations can also be used with for example L_1- and L_2-norms: the parameter vector is partitioned into blocks with L_2-norm used

within blocks and L_1-norm for combining blocks (Huang & Zhang, 2010).

An extensive theoretical study of these regularizations and the introduction of other more general ones were presented by Dudík et al. (2007). The generalization guarantees for regularized Maxent models depends on the sample size and the complexity of the family of features used. This dependency suggests using relatively simpler feature families, such as threshold functions over the input variables. However, feature functions selected from simpler families could excessively limit the expressiveness of the model.

This paper introduces and studies a new family of density estimation models, *Structural Maxent models*, which offers the flexibility of selecting features out of complex families, while benefiting from strong learning guarantees. Let $H_1, \ldots, H_p$ be p families of feature functions with increasing complexity. H_k could be for example the family of regression trees of depth k or that of monomials of degree k based on the input variables. The main idea behind the design of our model is to allow the use of features from the family of very deep trees or other rich or complex families (that is, H_ks with relatively large k), but to reserve less total model parameter weight for such features than for those chosen from simpler families (H_ks with smaller k). We call our Maxent models *structural* since they exploit the structure of H as a union of H_ks. Note, however, that the main idea behind the design of our models is distinct from that of structural risk minimization (SRM) (Vapnik, 1998): while SRM seeks a single H_k with an optimal trade-off between empirical error and complexity, structural Maxent allocates different model weights to features in different H_ks to achieve an even better trade-off based on multiple H_ks.

In the following, we first define a new Structural Maxent principle for the general scenario where feature functions may be selected from multiple families (Section 2.1). Our new principle takes into consideration the different complexity of each of these families and is supported by data-dependent uniform convergence bounds. This principle guides us to design our new Structural Maxent models, whose regularization depends on the data-dependent complexity of each of the feature families. We study the optimization problem for Structural Maxent models and present a duality theorem showing the equivalence of the primal and dual problems, which can be viewed as a counterpart of the duality theorem for standard Maxent (Pietra et al., 1997) (see also (Dudík et al., 2007)) (Section 2.2). Next, we present data-dependent learning guarantees for our Structural Maxent models in terms of the Rademacher complexity of each of the feature families used (Section 2.3). The amount of total parameter weight apportioned to each family in our models is quantitatively determined by our new Maxent principle and further justified by these learning guarantees. In Section 2.4, we describe in detail our StructMaxent algorithm, including the details of its derivation and its pseudocode. Our algorithm consists of applying coordinate descent to the dual objective function proceeding in the same way as (Dudík et al., 2007) for L_1-norm regularized Maxent. We derive two versions of our algorithm differing only by the definition of the step, which is based on minimizing different upper bounds. The first version of our algorithm uses the same upper bounding technique as Dudík et al. (2007) for L_1-norm regularized Maxent, and, as with the algorithm of Dudík et al. (2007), is subject to several assumptions. The second version of our algorithm does not require any assumption and leads to a simpler analysis, at the cost of only slightly slower convergence. We prove convergence guarantees for both versions of our algorithm.

In Section 3, we further extend our ideas to the scenario of multi-class classification. We present a new conditional Maxent principle for the case where multiple feature families are used (Section 3.1), which leads to the definition of our *conditional Structural Maxent models*. These are conditional probability models that admit as a special case existing conditional Maxent models, or, equivalently, multinomial logistic regression models. We prove a duality theorem showing the equivalence of the primal and dual optimization problems for our conditional Structural Maxent models. This shows that these models precisely coincide with the DeepBoost algorithms of Cortes et al. (2014) and Kuznetsov et al. (2014) in the special case where the surrogate loss function used is the logistic function. Thus, our algorithm benefits from the data-dependent learning guarantees and empirical validation already presented for deep boosting. Conversely, our analysis and new conditional structural Maxent principle provide an alternative justification in support of deep boosting. In Section 4, we report the results of extensive experiments with data from various domains including community crimes, traffic and species habitat modeling. Our results show the advantage of our structural Maxent models for density estimation when compared to existing regularized Maxent models.

2. Structural Maxent models

Let $\mathcal{X}$ denote the input space. We first consider the following problem of density estimation. Assume that a sample $S = (x_1, \ldots, x_m) \in \mathcal{X}^m$ of m points drawn from an unknown distribution $\mathcal{D}$ is given and that we have at our disposal a feature vector $\mathbf{\Phi}(x)$ associated to each point $x \in \mathcal{X}$. Then, the standard density estimation problem consists of using the sample S and the features to find a distribution p that forms a good estimate of $\mathcal{D}$.

We consider the general case of infinite families of feature functions. Note that even standard families of threshold

functions over a finite set of variables have infinite size. Let $H_1, \ldots, H_p$ be $p \geq 1$ families of functions mapping $\mathcal{X}$ to $\mathbb{R}$ with each feature function falling in one of these families. Assume that the H_ks have increasing Rademacher complexities. For example, H_1 may be composed of some simple feature functions ϕ, H_2 the set of all products $\phi\psi$ with $\phi, \psi \in H_1$, and, more generally, H_k may contain all monomials of degree k over functions in H_1. In fact, it is common in some applications where Maxent is used to design new features precisely in this manner, starting with some basic features, here H_1. Similarly, H_k may be the set of regression trees of depth k with node questions based on threshold functions of some set of variables or, more complex functions of these variables as in splines.

2.1. Principle

The key idea behind our new Maxent formulation is to take into consideration the distinct complexity of each family of feature functions H_k. Let $\mathbf{\Phi}$ be the feature mapping from $\mathcal{X}$ to the space $\mathbb{F}$, with $\mathbb{F} = \mathbb{F}_1 \times \cdots \times \mathbb{F}_p$. For any $x \in \mathcal{X}$, $\mathbf{\Phi}(x)$ can be decomposed as $\mathbf{\Phi}(x) = (\mathbf{\Phi}_1(x), \ldots, \mathbf{\Phi}_p(x))$, with $\mathbf{\Phi}_k$ a mapping from $\mathcal{X}$ to $\mathbb{F}_k$ such that $\|\mathbf{\Phi}_k\|_\infty \leq \Lambda$, for all $k \in [1, p]$, and with each of its components $\Phi_{k,j}$ in H_k. By a standard Rademacher complexity bound (Koltchinskii & Panchenko, 2002) and the union bound, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ over the choice of a sample of size m:

$$\left\| \mathop{\mathrm{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}_k(x)] - \mathop{\mathrm{E}}_{x\sim S}[\mathbf{\Phi}_k(x)] \right\|_\infty \leq 2\mathfrak{R}_m(H_k) + \Lambda\sqrt{\frac{\log\frac{2p}{\delta}}{2m}}, \quad (1)$$

for all $k \in [1, p]$. Here, we denote by $\mathrm{E}_{x\sim S}[\mathbf{\Phi}_k(x)]$ the expected value of $\mathbf{\Phi}_k$ with respect to the empirical distribution defined by the sample S. Let p_0 be a distribution over $\mathcal{X}$ with $\mathsf{p}_0[x] > 0$ for all $x \in \mathcal{X}$, typically chosen to be the uniform distribution. In view of (1), our extension of the maximum entropy principle (see Jaynes (1957; 1983)) consists of selecting p as the distribution that is the closest to p_0 and that verifies for all $k \in [1, p]$:

$$\left\| \mathop{\mathrm{E}}_{x\sim\mathsf{p}}[\mathbf{\Phi}_k(x)] - \mathop{\mathrm{E}}_{x\sim S}[\mathbf{\Phi}_k(x)] \right\|_\infty \leq 2\mathfrak{R}_m(H_k) + \beta,$$

where $\beta > 0$ is a parameter. Here, closeness is measured using the relative entropy. Let Δ denote the simplex of distributions over $\mathcal{X}$, then, our structural Maxent principle can be formulated as the following optimization problem:

$$\min_{\mathsf{p}\in\Delta} D(\mathsf{p} \parallel \mathsf{p}_0), \qquad \text{s.t. } \forall k \in [1, p]: \quad (2)$$
$$\left\| \mathop{\mathrm{E}}_{x\sim\mathsf{p}}[\mathbf{\Phi}_k(x)] - \mathop{\mathrm{E}}_{x\sim S}[\mathbf{\Phi}_k(x)] \right\|_\infty \leq 2\mathfrak{R}_m(H_k) + \beta.$$

Since the relative entropy, D, is convex with respect to its arguments and since the constraints are affine, this defines a convex optimization problem. The solution is in fact unique since the relative entropy is strictly convex. Since the empirical distribution is a feasible point, problem (2) is feasible. For any convex set K, let I_K denote the function defined by $I_K(x) = 0$ if $x \in K$, $I_K(x) = +\infty$ otherwise. Then, the problem can be equivalently expressed as $\min_{\mathsf{p}} F(\mathsf{p})$ where

$$F(\mathsf{p}) = D(\mathsf{p} \parallel \mathsf{p}_0) + I_\Delta(\mathsf{p}) + I_C(\mathop{\mathrm{E}}_{\mathsf{p}}[\mathbf{\Phi}]), \quad (3)$$

where Δ is the simplex and where C is the convex set defined by $C = \{\mathbf{u} \colon \|\mathbf{u}_k - \mathrm{E}_S[\mathbf{\Phi}_k]\|_\infty \leq \beta_k, \forall k \in [1, p]\}$, with $\beta_k = 2\mathfrak{R}_m(H_k) + \beta$.

2.2. Dual problem

As for the standard Maxent models with L_1 constraints (Kazama & Tsujii, 2003) (see also (Dudík et al., 2007)), we can derive an equivalent dual problem for (2) or (3) formulated as a regularized maximum likelihood problem over Gibbs distributions. Let G be the function defined for all $\mathbf{w}$ in the dual of $\mathbb{F}$ by

$$G(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \log\left[\frac{\mathsf{p}_\mathbf{w}[x_i]}{\mathsf{p}_0[x_i]}\right] - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1, \quad (4)$$

with $\mathsf{p}_\mathbf{w} = \frac{\mathsf{p}_0[x]e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}}{Z_\mathbf{w}}$ and $Z_\mathbf{w} = \sum_{x\in\mathcal{X}} \mathsf{p}_0[x]e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}$. For simplicity, we assume that the dual of $\mathbb{F}$ is $\mathbb{R}^N$. Then, the following theorem gives a result similar to the duality theorem of (Pietra et al., 1997).

Theorem 1. *Problems (2) and (3) are equivalent to the dual optimization problem* $\sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w})$*:*

$$\sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p}). \quad (5)$$

Furthermore, let $\mathsf{p}^* = \operatorname{argmin}_{\mathsf{p}} F(\mathsf{p})$*, then, for any* $\epsilon > 0$ *and any* $\mathbf{w}$ *such that* $|G(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w})| < \epsilon$*, the following inequality holds:* $D(\mathsf{p}^* \parallel \mathsf{p}_\mathbf{w}) \leq \epsilon$.

The proof is given in Appendix A (Theorem 1). In view of the theorem, if $\mathbf{w}$ is an ϵ-solution of the dual optimization problem, then $D(\mathsf{p}^* \parallel \mathsf{p}_\mathbf{w}) \leq \epsilon$, which, by Pinsker's inequality implies that $\mathsf{p}_\mathbf{w}$ is $\sqrt{2\epsilon}$-close in L_1-norm to the optimal solution of the primal: $\|\mathsf{p}^* - \mathsf{p}_\mathbf{w}\|_1 \leq \sqrt{2\epsilon}$. Thus, the solution of our structural Maxent problem can be determined by solving the dual problem, which can be written equivalently as follows:

$$\inf_{\mathbf{w}\in\mathbb{R}^N} \beta\|\mathbf{w}\|_1 + 2\sum_{k=1}^{p} \mathfrak{R}_m(H_k)\|\mathbf{w}_k\|_1 - \frac{1}{m}\sum_{i=1}^{m} \log \mathsf{p}_\mathbf{w}[x_i]. \quad (6)$$

The difference in problem (6) with respect to the common L_1-regularized Maxent problem is the remarkable new second term, which is defined by the Rademacher complexities of the H_ks. This term penalizes more the norm of a weight vector $\mathbf{w}_k$ associated to a feature vector selected from a complex hypothesis set H_k.

2.3. Generalization bound

Let $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ denote the log-loss of the distribution $\mathsf{p}_{\mathbf{w}}$ with respect to a distribution $\mathcal{D}$, $\mathcal{L}_{\mathcal{D}}(\mathbf{w}) = \mathrm{E}_{x\sim\mathcal{D}}[-\log \mathsf{p}_{\mathbf{w}}[x]]$, and similarly $\mathcal{L}_S(\mathbf{w})$ its log-loss with respect to the empirical distribution defined by a sample S.

Theorem 2. *Fix $\delta > 0$. Let $\widehat{\mathbf{w}}$ be a solution to the optimization (6) with $\beta = \Lambda\sqrt{\frac{\log\frac{2p}{\delta}}{2m}}$. Then, with probability at least $1-\delta$ over the sample S, $\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}})$ is bounded by*

$$\inf_{\mathbf{w}} \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + 2\sum_{k=1}^{p} \|\mathbf{w}_k\|_1 \left[2\mathfrak{R}_m(H_k) + \Lambda\sqrt{\frac{\log\frac{2p}{\delta}}{2m}}\right].$$

Proof. Using the definition of $\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ and $\mathcal{L}_S(\mathbf{w})$, Hölder's inequality, and inequality (1), with probability at least $1-\delta$, the following holds:

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) \\
&\quad= \widehat{\mathbf{w}} \cdot [\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] - \mathop{\mathrm{E}}_{\mathcal{D}}[\mathbf{\Phi}]] \\
&\quad= \sum_{k=1}^{p} \widehat{\mathbf{w}}_k \cdot [\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}_k] - \mathop{\mathrm{E}}_{\mathcal{D}}[\mathbf{\Phi}_k]] \\
&\quad\leq \sum_{k=1}^{p} \|\widehat{\mathbf{w}}_k\|_1 \, \|\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}_k] - \mathop{\mathrm{E}}_{\mathcal{D}}[\mathbf{\Phi}_k]\|_\infty \\
&\quad\leq \sum_{k=1}^{p} \|\widehat{\mathbf{w}}_k\|_1 (\beta + 2\mathfrak{R}_m(H_k)).
\end{aligned}$$

Thus, since $\widehat{\mathbf{w}}$ is a minimizer, we can write for any $\mathbf{w}$

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \\
&= \mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \\
&\leq \sum_{k=1}^{p} \|\widehat{\mathbf{w}}_k\|_1 (\beta + 2\mathfrak{R}_m(H_k)) + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \\
&\leq \sum_{k=1}^{p} \|\mathbf{w}_k\|_1 (\beta + 2\mathfrak{R}_m(H_k)) + \mathcal{L}_S(\mathbf{w}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}) \\
&\leq 2\sum_{k=1}^{p} \|\mathbf{w}_k\|_1 (\beta + 2\mathfrak{R}_m(H_k)),
\end{aligned}$$

where we used in the last step the left inequality counterpart of inequality (1). This concludes the proof. □

This bound suggests that learning with feature functions selected from highly complex families (relatively large $\mathfrak{R}_m(H_k)$) can benefit from favorable learning guarantees so long as the total weight assigned to these features by the model is relatively small.

2.4. Algorithm

Our algorithm consists of applying coordinate descent to the objective function of (6). Ignoring the constant term $-\frac{1}{m}\sum_{i=1}^{m} \log \mathsf{p}_0[x_i]$, the optimization problem (6) can be rewritten as $\inf_{\mathbf{w}} F(\mathbf{w})$ with

$$F(\mathbf{w}) = \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 - \mathbf{w}\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] + \log\Big[\sum_{x\in\mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}\Big],$$

and $\beta_k = \beta + 2\mathfrak{R}_m(H_k)$ for $k \in [1,p]$. Function F is not differentiable but it is convex and admits a subdifferential at any point. For notational simplicity, we will assume that each $\mathbb{F}_k$ is finite.

2.4.1. Direction

Let $\mathbf{w}_{t-1}$ denote the weight vector defined after $(t-1)$ iterations. At each iteration $t \in [1,T]$, the direction $\mathbf{e}_{k,j}$, $(k,j) \in [1,p]\times[1,N_k]$, with N_k being the size of $\mathbb{F}_k$, considered by coordinate descent is $\delta F(\mathbf{w}_{t-1}, \mathbf{e}_{k,j})$.

If $w_{t-1,k,j} \neq 0$, then F admits a directional derivative along $\mathbf{e}_{k,j}$ given by

$$F'(\mathbf{w}_{t-1}, \mathbf{e}_{k,j}) = \beta_k \operatorname{sgn}(w_{t-1,k,j}) + \epsilon_{t-1,k,j}.$$

where $\epsilon_{t-1,k,j} = \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] - \mathrm{E}_S[\Phi_{k,j}]$. If $w_{t-1,k,j} = 0$, F admits right and left directional derivatives along $\mathbf{e}_{k,j}$: $F'_+(\mathbf{w}_{t-1}, \mathbf{e}_{k,j}) = \beta_k + \epsilon_{t-1,k,j}$ and $F'_-(\mathbf{w}_{t-1}, \mathbf{e}_{k,j}) = -\beta_k + \epsilon_{t-1,k,j}$. Thus, in summary, we can write,

$$\delta F(\mathbf{w}_{t-1}, \mathbf{e}_{k,j}) = \begin{cases} \beta_k \operatorname{sgn}(w_{t-1,k,j}) + \epsilon_{t-1,k,j} & \text{if } (w_{t-1,k,j} \neq 0) \\ 0 & \text{else if } \left|\epsilon_{t-1,k,j}\right| \leq \beta_k \\ -\beta_k \operatorname{sgn}(\epsilon_{t-1,k,j}) + \epsilon_{t-1,k,j} & \text{otherwise.} \end{cases}$$

The coordinate descent algorithm selects the direction $\mathbf{e}_{k,j}$ with the largest numeric value of $\delta F(\mathbf{w}_{t-1}, \mathbf{e}_{k,j})$.

2.4.2. Two alternative step sizes

Given the direction $\mathbf{e}_{k,j}$, the optimal step value η is given by $\operatorname{argmin}_\eta F(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_{k,j})$. η can be found via a line search or other numerical methods. We can also derive a closed-form solution for the step by minimizing an upper bound on $F(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_{k,j})$. We present closed-form solutions based on two different but related upper bounds. The full argument is given in Appendix C. In what follows, we highlight the main steps of these derivations.

Observe that

$$\begin{aligned}
&F(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1}) \qquad (7)\\
&= \beta_k(|w_{k,j} + \eta| - |w_{k,j}|) - \eta \mathop{\mathrm{E}}_{S}[\Phi_{k,j}] + \log\Big[\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}]\Big].
\end{aligned}$$

Step size – StructMaxent1

Since $\Phi_{k,j} \in [-\Lambda, +\Lambda]$, by the convexity of $x \mapsto e^{\eta x}$, we can write

$$e^{\eta\Phi_{k,j}} \leq \tfrac{\Lambda - \Phi_{k,j}}{2\Lambda} e^{-\eta\Lambda} + \tfrac{\Phi_{k,j} + \Lambda}{2\Lambda} e^{\eta\Lambda}.$$

Taking the expectation and the log yields

$$\log \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta \Phi_{k,j}}] \leq \log \left[\frac{\overline{\Phi}^{+}_{t-1,k,j} e^{\eta \Lambda} - \overline{\Phi}^{-}_{t-1,k,j} e^{-\eta \Lambda}}{2\Lambda} \right],$$

where we used the following notation:

$$\overline{\Phi}^{s}_{t-1,k,j} = \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + s\Lambda \qquad \overline{\Phi}^{s}_{k,j} = \mathop{\mathrm{E}}_{S}[\Phi_{k,j}] + s\Lambda,$$

for all $(k, j) \in [1, p] \times [1, N_k]$ and $s \in \{-1, +1\}$.

Plugging back this inequality in (7) and minimizing the resulting upper bound on $F(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1})$ leads to the closed-form expression for the step size η. The full pseudocode of the algorithm using this closed-form solution for the step size, StructMaxent1, is given in Appendix B.

Note that the following condition on Φ, Λ and β_k must be satisfied:

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\}) \wedge (|\mathop{\mathrm{E}}_{S}[\Phi_{k,j}]| < \Lambda - \beta_k). \quad (8)$$

This first version of StructMaxent uses the same upper bounding technique as the L_1-norm regularized Maxent algorithm of Dudík et al. (2007), which is subject to the same type of conditions as (8).

STEP SIZE – STRUCTMAXENT2

An alternative method is based on a somewhat looser upper bound for $\log \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta \Phi_{k,j}}]$ using Hoeffding's lemma. A key advantage is that the analysis is no more subject to conditions such as (8). Additionally, the StructMaxent2 algorithm is much simpler. The price to pay is a slightly slower convergence but, as pointed out in Section 4, both algorithms exhibit an exponential convergence and lead to the same results. By Hoeffding's lemma, since $\Phi_{k,j} \in [-\Lambda, +\Lambda]$, we can write

$$\log \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta \Phi_{k,j}}] \leq \eta \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + \frac{\eta^2 \Lambda^2}{2}.$$

Combining this inequality with (7) and minimizing the resulting upper bound leads to an alternative closed-form solution for the step size η. Figure 1 shows the pseudocode of our algorithm using the closed-form solution for the step size just presented.

2.5. Convergence analysis

The following result gives convergence guarantees for both versions of our StructMaxent algorithm.

Theorem 3. *Let $(\mathbf{w}_t)_t$ be the sequence of parameter vectors generated by StructMaxent1 or StructMaxent2. Then, $(\mathbf{w}_t)_t$ converges to the optimal solution $\mathbf{w}^*$ of (6).*

STRUCTMAXENT2$(S = (x_1, \ldots, x_m))$

for $t \leftarrow 1$ **to** T **do**
  for $k \leftarrow 1$ **to** p **and** $j \leftarrow 1$ **to** N_k **do**
    if $(w_{t-1,k,j} \neq 0)$ **then**
      $d_{k,j} \leftarrow \beta_k \operatorname{sgn}(w_{t-1,k,j}) + \epsilon_{t-1,k,j}$
    elseif $|\epsilon_{t-1,k,j}| \leq \beta_k$ **then**
      $d_{k,j} \leftarrow 0$
    else $d_{k,j} \leftarrow -\beta_k \operatorname{sgn}(\epsilon_{t-1,k,j}) + \epsilon_{t-1,k,j}$
  $(k, j) \leftarrow \operatorname*{argmax}_{(k,j) \in [1,p] \times [1,N_k]} |d_{k,j}|$
  if $(|w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j}| \leq \beta_k)$ **then**
    $\eta \leftarrow -w_{t-1,k,j}$
  elseif $(w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j} > \beta_k)$ **then**
    $\eta \leftarrow \frac{1}{\Lambda^2}[-\beta_k - \epsilon_{t-1,k,j}]$
  else $\eta \leftarrow \frac{1}{\Lambda^2}[\beta_k - \epsilon_{t-1,k,j}]$
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \eta \mathbf{e}_{k,j}$
  $\mathsf{p}_{\mathbf{w}_t} \leftarrow \frac{\mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}{\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}$
return $\mathsf{p}_{\mathbf{w}_t}$

Figure 1. Pseudocode of the StructMaxent2 algorithm. For all $(k, j) \in [1, p] \times [1, N_k]$, $\beta_k = 2\mathfrak{R}_m(H_k) + \beta$ and $\epsilon_{t-1,k,j} = \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] - \mathrm{E}_S[\Phi_{k,j}]$. Note that no technical assumption such as those of (8) are required here for the closed-form expressions of the step size.

The full proof of this result is given in Appendix D. We also remark that if the step size is determined via a line search, then our algorithms benefit from an exponential convergence rate (Luo & Tseng, 1992).

3. Conditional Structural Maxent Models

In this section, we extend the analysis presented in the previous section to that of conditional Maxent models, also known as multinomial logistic regression.

We consider a multi-class classification problem with $c \geq 1$ classes and denote by $\mathcal{Y} = \{1, \ldots, c\}$ the output space and by $\mathcal{D}$ the distribution over $\mathcal{X} \times \mathcal{Y}$ from which pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn i.i.d. As in standard supervised learning problems, the learner receives a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ and, as in Section 2.1, we assume the learner has access to a feature function $\mathbf{\Phi} \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{F}$ that can be decomposed as $\mathbf{\Phi} = (\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_p)$, where for any $k \in [1, p]$, $\mathbf{\Phi}_k$ is a mapping from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{F}_k$ with elements in H_k.

3.1. Principle

Our conditional Structural Maxent, or structural logistic regression models can be defined in a way similar to that of density estimation. Here, the problem consists of learning a conditional probability $\mathsf{p}[\cdot|x]$ for all $x \in \mathcal{X}$. As in the density estimation case, we will denote by $\mathsf{p}_0[\cdot|x]$ a conditional

probability, often chosen to be the uniform distribution.

As in the density estimation case, the definition of our algorithm is guided by a generalization bound expressed in terms of the Rademacher complexity of the families H_k. For any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ over the choice of a sample of size m, for all $k \in [1, p]$:

$$\left\| \mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathcal{D}[\cdot|x]}}[\boldsymbol{\Phi}_k(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\widehat{\mathsf{p}}[\cdot|x]}}[\boldsymbol{\Phi}_k(x,y)] \right\|_\infty \leq \beta_k, \quad (9)$$

where we denote by E_S the expectation over the empirical distribution defined by sample S and $\beta_k = 2\mathfrak{R}_m(H_k) + \sqrt{\frac{\log 2\frac{p}{\delta}}{2m}}$. Let $\widehat{\mathsf{p}}$ denote the empirical distribution of the input points. Our structural conditional Maxent model is defined by searching the conditional probabilities $\mathsf{p}[\cdot|x]$ that are as close as possible to $\mathsf{p}_0[\cdot|x]$, while ensuring an inequality similar to (9), where closeness is defined via the conditional relative entropy based on $\widehat{\mathsf{p}}$ (Cover & Thomas, 2006). This leads to the following convex optimization problem:

$$\begin{aligned} \min_{\mathsf{p}[\cdot|x]\in\Delta} & \sum_{x\in\mathcal{X}} \widehat{\mathsf{p}}[x]\, D\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) \quad (10)\\ \text{s.t.} & \left\| \mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}\Big[\mathop{\mathrm{E}}_{y\sim\mathsf{p}[\cdot|x]}[\boldsymbol{\Phi}_k(x,y)]\Big] - \mathop{\mathrm{E}}_{(x,y)\sim S}[\boldsymbol{\Phi}_k(x,y)] \right\|_\infty \\ & \leq 2\mathfrak{R}_m(H_k) + \beta, \quad \forall k \in [1,p], \end{aligned}$$

where we again denote by Δ the simplex in $\mathcal{Y}$. Let p denote the vector of conditional probabilities $(\mathsf{p}[\cdot|x_i])_{i\in[1,m]}$. Then, this problem can be equivalently expressed as $\min_{\mathsf{p}} \widetilde{F}(\mathsf{p})$ where

$$\widetilde{F}(\mathsf{p}) = \mathop{\mathrm{E}}_{x\sim\widehat{p}}\Big[D\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) + I_\Delta\big(\mathsf{p}[\cdot|x]\big)\Big] + I_C\Big(\mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\boldsymbol{\Phi}]\Big), \quad (11)$$

where Δ is the simplex and C the convex set defined by $C = \{\mathbf{u} \colon \|\mathbf{u}_k - \mathrm{E}_S[\boldsymbol{\Phi}_k]\|_\infty \leq \beta_k, \forall k \in [1,p]\}$, with $\beta_k = 2\mathfrak{R}_m(H_k) + \beta$.

3.2. Dual problem

Let $\widetilde{G}$ be the function defined for all $\mathbf{w} \in \mathbb{R}^N$ by

$$\widetilde{G}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \log\left[\frac{\mathsf{p}_\mathbf{w}[y_i|x_i]}{\mathsf{p}_0[y_i|x_i]}\right] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1, \quad (12)$$

with, for all $x \in \mathcal{X}$, $\mathsf{p}_\mathbf{w}[y|x] = \frac{\mathsf{p}_0[x]e^{\mathbf{w}\cdot\boldsymbol{\Phi}(x)}}{Z_\mathbf{w}(x)}$ and $Z_\mathbf{w}(x) = \sum_{x\in\mathcal{X}} \mathsf{p}_0[x]e^{\mathbf{w}\cdot\boldsymbol{\Phi}(x)}$. Then, the following theorem gives a result similar to the duality theorem presented in the non-conditional case.

Theorem 4. *Problem (10) is equivalent to dual optimization problem* $\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})$*:*

$$\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w}) = \min_{\mathsf{p}} \widetilde{F}(\mathsf{p}). \quad (13)$$

Furthermore, let $\mathsf{p}^* = \operatorname{argmin}_{\mathsf{p}} \widetilde{F}(\mathsf{p})$*. Then, for any* $\epsilon > 0$ *and any* $\mathbf{w}$ *such that* $|\widetilde{G}(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})| < \epsilon$*, we have* $\mathrm{E}_{x\sim\widehat{p}}\left[D\big(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big)\right] \leq \epsilon$.

The proof of the theorem is given in Appendix A (Theorem 4). As in the non-conditional case, the theorem suggests that the solution of our structural conditional Maxent problem can be determined by solving the dual problem, which can be written equivalently as follows:

$$\inf_{\mathbf{w}} \beta\|\mathbf{w}\|_1 + 2\sum_{k=1}^p \mathfrak{R}_m(H_k)\|\mathbf{w}_k\|_1 - \frac{1}{m}\sum_{i=1}^m \log\big[\mathsf{p}_\mathbf{w}[y_i|x_i]\big]. \quad (14)$$

We note that problem (14) coincides with the optimization problem presented by Cortes et al. (2014) and Kuznetsov et al. (2014) for the DeepBoost algorithms in the particular case where the logistic function is used as a convex surrogate loss function. Our analysis and derivation starting from the conditional Maxent principle provide an alternative justification for these algorithms. In return, our conditional StructMaxent algorithm benefits from the learning guarantees already given by deep boosting.

4. Experiments

This section reports the results of our experiments with the StructMaxent algorithm. We have fully implemented both StructMaxent1 and StructMaxent2 with diverse feature families and will make the software used in our experiments available as open-source. We do not report empirical results for the conditional StructMaxent algorithm since, as already pointed out, the conditional algorithm coincides with the deep boosting algorithms that have been already extensively studied by Cortes et al. (2014).

Our StructMaxent algorithm can be applied with a variety of different families of feature functions H_k mapping the input space $\mathcal{X}$ to $\mathbb{R}$. In our experiments, we used the union of two broad classes of feature maps: monomial features H^{mono} and tree features H^{trees}. Let $H_1^{\text{mono}} = \{\psi_1, \ldots, \psi_d\}$ denote the set of raw features mapping $\mathcal{X}$ to $\mathbb{R}$. We define $H_k^{\text{mono}} = \{\psi_{j_1} \cdots \psi_{j_k} \colon j_r \in [1,d]\}$ as the set of all monomials of degree k derived from H_1^{mono} and H^{mono} as the union of all these families: $H^{\text{mono}} = \cup_{k=1}^\infty H_k^{\text{mono}}$. Similarly, we denote by H_k^{trees} the set of all binary decision trees with k internal nodes, where the node questions are threshold functions $1_{\psi_j(x)\leq\theta}$ with $j \in [1,d]$ and $\theta \in \mathbb{R}$, and where the leaves are labeled with zero or one and define H^{trees} as the union of these families: $H^{\text{trees}} = \cup_{k=1}^\infty H_k^{\text{trees}}$. Note that our monomial and tree features are strict generalizations of the product features and threshold features used by Phillips et al. (2004; 2006) and Dudík et al. (2007).

The hypothesis set $H = H^{\text{mono}} \cup H^{\text{trees}}$ is infinite and can be very large even for a bounded monomial degree and for

Table 1. Experimental comparison of Maxent algorithms.

Dataset	Maxent	L_1-Maxent	StructMaxent
b. variegatus	19.95 (0.54)	15.67 (0.33)	**13.36 (0.28)**
arson	5.93 (0.02)	5.75 (0.01)	**5.68 (0.02)**
rapes	6.42 (0.02)	6.22 (0.01)	**6.16 (0.01)**
burglary	6.04 (0.01)	5.85 (0.01)	**5.78 (0.01)**
larceny	5.83 (0.01)	5.65 (0.01)	**5.58 (0.01)**
traffic	14.72 (1.11)	13.85 (0.24)	**13.00 (1.01)**

Dataset	Maxent	L_1-Maxent	StructMaxent
m. minutus	17.41 (0.87)	12.20 (0.78)	**10.25 (0.30)**
murders	5.38 (0.02)	5.23 (0.02)	**5.17 (0.02)**
robberies	5.14 (0.01)	5.00 (0.01)	**4.94 (0.01)**
assault	6.65 (0.02)	6.35 (0.01)	**6.30 (0.01)**
auto theft	5.93 (0.02)	5.75 (0.01)	**5.68 (0.02)**

trees restricted to the finite training sample. Thus, exhaustively searching for the best monomial in H^{mono} and the best decision tree in H^{trees} is not tractable. Instead, we used the following greedy procedure: given the best monomial $m(x)$ of degree k (starting with $k = 0$), we find the best monomial of degree $k+1$ that can be obtained from $m(x)$ by multiplying it with one of $\psi \in H_1^{\text{mono}}$. Similarly, given the best decision tree of size k, we find the best decision tree of size $k+1$ that can be obtained by splitting exactly one leaf of the given tree. Let $\Phi_1, \ldots, \Phi_{t-1}$ be features chosen by the algorithm after the first $t-1$ iterations. Then, on the t-th iteration, the algorithm selects a feature from $\{\Phi_1, \ldots, \Phi_{t-1}, t^*, m^*\}$, where t^* is the tree and m^* the monomial obtained by the procedure just described.

Note that StructMaxent requires the knowledge of the Rademacher complexities $\mathfrak{R}_m(H_k)$, which in certain cases can be estimated from the data. For simplicity, in our experiments, we used the following upper bounds

$$\begin{aligned} \mathfrak{R}_m(H_k^{\text{mono}}) &\leq \sqrt{\frac{2k \log d}{m}} \\ \mathfrak{R}_m(H_k^{\text{trees}}) &\leq \sqrt{\frac{(4k+2)\log_2(d+2)\log(m+1)}{m}}, \end{aligned} \tag{15}$$

and we redefined β_k as $\beta_k = \lambda B_k + \beta$, where B_k is the appropriate choice of an upper bound in (15) and where the parameter λ is introduced to control the balance between the magnitude of B_k and β. The proof of these upper bounds is given in Appendix E.

In our experiments, we determined both parameters λ and β through a validation procedure. Specifically, we optimized over $\lambda, \beta \in \{0.0001, 0.001, 0.01, 0.1, 0.5, 1, 2\}$. Remarkably, in all of our experiments the best value of λ was 0.1. We compared StructMaxent with L_1-regularized Maxent (Kazama & Tsujii, 2003; Phillips et al., 2004; 2006; Dudík et al., 2007) which is the special case of StructMaxent with $\lambda = 0$. L_1-regularized Maxent is the only algorithm used for experiments in (Phillips et al., 2004; 2006; Dudík et al., 2007). The parameter β of L_1-Maxent was set in the same way as for the StructMaxent algorithm. We compared the performance of StructMaxent to that of L_1-Maxent and to standard Maxent ($\lambda = \beta = 0$) which served as our baseline. Following Phillips et al. (2004), we ran each algorithm for 500 rounds, or until the change in the objective on a single round fell below 10^{-5}.

For our experiments, we used a number of different datasets related to species habitat modeling, traffic modeling, and to communities and crimes, which we describe in the following subsections.

4.1. Species habitat modeling

Species habitat modeling is among the most prominent applications of Maxent models (Phillips et al., 2004; 2006; Dudík et al., 2007; Elith et al., 2011). For our experiments, we used two data sets from (Phillips et al., 2006) which are accessible from `http://www.cs.princeton.edu/~schapire/maxent`. These datasets consist of a set of 648,658 geographical locations in South America which constitute the input space $\mathcal{X}$. For each of these locations, we have a set of environmental variables such as temperature and precipitation, which constitute our raw features $\psi_1, \ldots, \psi_d$. For each location, we are also given the number of times a particular kind of species has been observed, which defines the sample S used as an input to Maxent algorithms. The first dataset consists of 116 observations of *bradypus variegatus* and the second dataset has 88 observations of *microryzomys minutus*. The task consists of estimating the geographical distribution of each species.

For each species, we first randomly split the sample S into a training set S_1 (70%) and a test set S_2 (30%). We trained all algorithms on S_1 and used the error on S_2 to find the optimal value of the parameters λ and β. Next we again split the sample into a training set S_1' (70%) and a test set S_2' (30%). Using the best parameter values found on the previous step, each algorithm was trained on S_1' and the log-loss on S_2' was recorded. We repeated the last step 10 times and reported the average log-loss over these 10 runs. This experimental setup matches that of (Phillips et al., 2004;

2006), with the exception of the validation step, which is omitted in both of these references.

4.2. Minnesota traffic modeling

We also experimented with a Minnesota traffic dataset (Kwon, 2004) which is accessible from http://www.d.umn.edu/~tkwon/TMCdata/TMCarchive.html. This dataset contains traffic volume and velocity records collected by 769 road sensors over the period of 31 days with an interval of 30 seconds between observations. The input space $\mathcal{X}$ is the set of sensor locations and the task consists of estimating traffic density at each particular location based on the historical data. More precisely, we chose 11 times $t_1, \ldots, t_{11}$ on the 31st day starting at midnight and separated by one hour. For computational efficiency, instead of using the whole history, we defined the raw features $\psi_1, \ldots, \psi_d$ as the historical volume and velocity averages for each of the past 30 days, augmented with data from the past hour collected every 10 minutes. For any time t_l, there are between 1,000 to 20,000 cars observed by all of 769 sensors. For each time t_l, we randomly selected 70% of observations for training and the remaining observations were reserved for testing. Data from the first time t_1 was used to determine the best parameters λ and β. The parameter values were then fixed and used for training on the remaining ten times $t_2, \ldots, t_{11}$. We report log-loss on the test set averaged over these 10 time values.

4.3. Communities and crime

We used the UCI Communities and Crime dataset as another test case for StructMaxent algorithm. This dataset contains demographic and crime statistics for 2,215 US communities. Each community represents a point x in the input space $\mathcal{X}$ and we define base features ψ_j to be various demographic statistics. The goal is to model the likelihood of certain types of crimes based on the demographic statistics available. The data set includes 8 different types of crime: murder, rape, robbery, burglary, assault, larceny, auto theft and arson. For each type of crime, there are between 17,000 to 5,000,000 records. To speed up our experiments, we randomly sub-sampled 11 training sets of size 5,000 for each type of crime (with the remaining data used for testing). As before, the first training and test sets is used to determine the best values for the parameters λ and β, and we report the averaged log-loss on the test set when training with λ and β fixed at these values.

4.4. Results and discussion

The results of our experiments are presented in Table 1. They show that StructMaxent provides an improvement over L_1-Maxent that is comparable to the improvement of L_1-Maxent over standard Maxent models. All of our results are statistically significant using paired t-test at 1% level. Furthermore, StructMaxent outperforms other algorithms on each of the individual runs. Our experiments indicate that the performances of StructMaxent1 and StructMaxent2 are comparable – both better than that of L_1-regularized and non-regularized Maxent in terms of log-loss. Note that for the species habitat modeling experiments, Phillips et al. (2004; 2006) report only AUC (Area Under the ROC curve) values, which measure ranking quality, instead of the log-loss optimized for density estimation. They do so by treating the absence of any recorded observation at a particular location as a negative label. For completeness, we also report AUC results for our experiments in Table 2. StructMaxent outperforms other methods in terms of AUC as well.

Table 2. Average AUC values.

	b. variegatus	m. minutus
Maxent	0.810 ± 0.020	0.879 ± 0.141
L_1-Maxent	0.817 ± 0.027	0.972 ± 0.026
StructMaxent	$\mathbf{0.873 \pm 0.027}$	$\mathbf{0.984 \pm 0.006}$

The convergence of StructMaxent2 is somewhat slower than that of StructMaxent1, but both exhibit exponential convergence. Finally, note that the running time of our StructMaxent algorithms is similar to that of L_1-Maxent.

5. Conclusion

We presented a new family of density estimation models, Structural Maxent models, which benefit from strong data-dependent learning guarantees and can be used even with complex feature families. Our experiments demonstrated the empirical advantage of these models. We also introduced a new family of conditional probability models for multi-class classification, structural conditional Maxent models, and showed them to coincide with deep boosting when using the logistic function as a surrogate loss. Our conditional structural Maxent principle provide additional support in favor of this family of algorithms.

As with standard Maxent models (Lafferty, 1999), our structural Maxent models can be generalized by using an arbitrary Bregman divergence (Bregman, 1967) in place of the (unnormalized) relative entropy. Much of our analysis and theoretical guarantees straightforwardly extend to cover this generalization, modulo some additional assumptions on the properties of the Bregman divergence used.

Acknowledgments

We thank Tian Jiang for several discussions about this work and early experiments, and Slobodan Vucetic for facilitating our access to the Minnesota traffic dataset. This work was partly funded by the NSF award IIS-1117591 and the NSERC PGS D3 award.

References

Berger, Adam L., Pietra, Stephen Della, and Pietra, Vincent J. Della. A maximum entropy approach to natural language processing. *Comp. Linguistics*, 22(1), 1996.

Boyd, Stephen P. and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

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

Chen, Stanley F. and Rosenfeld, Ronald. A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, 8(1), 2000.

Cortes, Corinna, Mohri, Mehryar, and Syed, Umar. Deep boosting. In *Proceedings of ICML*, 2014.

Cover, Thomas M. and Thomas, Joy M. *Elements of Information Theory*. Wiley-Interscience, 2006.

Dudík, Miroslav, Phillips, Steven J., and Schapire, Robert E. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8, 2007.

Elith, Jane, Phillips, Steven J., Hastie, Trevor, Dudík, Miroslav, Chee, Yung En, and Yates, Colin J. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 1, 2011.

Goodman, Joshua. Exponential priors for maximum entropy models. In *Proceedings of HLT-NAACL*, 2004.

Huang, Junzhou and Zhang, Tong. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 08 2010.

Jaynes, E. T. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Jaynes, E. T. *Papers on probability, statistics, and statistical physics*. Synthese library. D. Reidel Pub. Co., 1983.

Kazama, Jun'ichi and Tsujii, Jun'ichi. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of EMNLP*, pp. 137–144, 2003.

Koltchinskii, Vladmir and Panchenko, Dmitry. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30, 2002.

Kuznetsov, Vitaly, Mohri, Mehryar, and Syed, Umar. Multi-class deep boosting. In *Proceedings of NIPS*, 2014.

Kwon, Taek M. TMC traffic data automation for Mn/DOT's traffic monitoring program. *Univ. of Minnesota*, Report no. Mn/DOT 2004-29, 2004.

Lafferty, John D. Additive models, boosting, and inference for generalized divergences. In *Proceedings of COLT*, pp. 125–133, 1999.

Lebanon, Guy and Lafferty, John D. Boosting and maximum likelihood for exponential models. In *Proceedings of NIPS*, pp. 447–454, 2001.

Luo, Zhi-Quan and Tseng, Paul. On the convergence of coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7 – 35, 1992.

Malouf, Robert. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of CoNLL-2002*, pp. 49–55, 2002.

Manning, Christopher D. and Klein, Dan. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of HLT-NAACL*, 2003.

Mansour, Yishay. Pessimistic decision tree pruning based on tree size. In *Proceedings of ICML*, 1997.

Phillips, Steven J., Dudík, Miroslav, and Schapire, Robert E. A maximum entropy approach to species distribution modeling. In *Proceedings of ICML*, 2004.

Phillips, Steven J., Anderson, Robet P., and Schapire, Robert E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 2006.

Pietra, Stephen Della, Pietra, Vincent J. Della, and Lafferty, John D. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4), 1997.

Ratnaparkhi, Adwait. Maximum entropy models for natural language processing. In *Encyclopedia of Machine Learning*, pp. 647–651. Springer, 2010.

Rockafellar, R. Tyrrell. *Convex analysis*. Princeton University Press, 1997.

Rosenfeld, Ronald. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech & Language*, 10(3):187–228, 1996.

Vapnik, Vladimir N. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Williams, Peter M. Bayesian regularisation and pruning using a Laplace prior. *Neural Computation*, 7:117–143, 1994.

A. Duality

The following is a version of the Fenchel duality theorem (see (Rockafellar, 1997)).

Theorem 5. *Let X and Y be Banach spaces, and $f\colon X \to \mathbb{R} \cup \{+\infty\}$ and $g\colon Y \to \mathbb{R} \cup \{+\infty\}$ convex functions. Let $A\colon X \to Y$ be a bounded linear map. If g is continuous at some point $y \in A\,\mathrm{dom}(f)$, then the following holds:*

$$\inf_{x \in X} (f(x) + g(Ax)) = \sup_{y^* \in Y^*} (-f^*(A^* y^*) - g^*(-y^*)), \tag{16}$$

where f^* and g^* are conjugate functions of f and g respectively, and A^* the adjoint of A. Furthermore, the supremum in (16) is attained if it is finite.

The following lemma gives the expression of the conjugate function of the (extended) relative entropy, which is a standard result (Boyd & Vandenberghe, 2004).

Lemma 6 (Conjugate function of the relative entropy)**.** *Let $f\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ be defined by $f(\mathsf{p}) = D(\mathsf{p} \parallel \mathsf{p}_0)$ if $\mathsf{p} \in \Delta$ and $f(\mathsf{p}) = +\infty$ elsewhere. Then, the conjugate function of f is the function $f^*\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ defined for all $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$ by*

$$f^*(\mathsf{q}) = \log\Big(\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathsf{q}[x]}\Big) = \log\Big(\mathop{\mathrm{E}}_{x \sim \mathsf{p}_0}[e^{\mathsf{q}[x]}]\Big).$$

Proof. By definition of f, for any $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$, we can write

$$\sup_{\mathsf{p} \in \mathbb{R}^{\mathcal{X}}} \big(\langle \mathsf{p}, \mathsf{q}\rangle - D(\mathsf{p} \parallel \mathsf{p}_0)\big) = \sup_{\mathsf{p} \in \Delta} \big(\langle \mathsf{p}, \mathsf{q}\rangle - D(\mathsf{p} \parallel \mathsf{p}_0)\big).$$

Fix $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$ and let $\bar{\mathsf{q}} \in \Delta$ be defined for all $x \in \mathcal{X}$ by

$$\bar{\mathsf{q}}[x] = \frac{\mathsf{p}_0[x] e^{\mathsf{q}[x]}}{\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathsf{q}[x]}} = \frac{\mathsf{p}_0[x] e^{\mathsf{q}[x]}}{\mathrm{E}_{p_0}[e^{\mathsf{q}}]}. \tag{17}$$

Then, the following holds for all $\mathsf{p} \in \Delta$:

$$\begin{aligned}
\langle \mathsf{p}, \mathsf{q}\rangle - D(\mathsf{p} \parallel \mathsf{p}_0) &= \mathop{\mathrm{E}}_{\mathsf{p}}[\log(e^{\mathsf{q}})] - \mathop{\mathrm{E}}_{\mathsf{p}}\left[\log \frac{\mathsf{p}}{\mathsf{p}_0}\right] \\
&= \mathop{\mathrm{E}}_{\mathsf{p}}\left[\log \frac{\mathsf{p}_0 e^{\mathsf{q}}}{p}\right] \\
&= -D(\mathsf{p} \parallel \bar{\mathsf{q}}) + \log \mathop{\mathrm{E}}_{\mathsf{p}_0}[e^{\mathsf{q}}].
\end{aligned}$$

Since $D(\mathsf{p} \parallel \bar{\mathsf{q}}) \geq 0$ and $D(\mathsf{p} \parallel \bar{\mathsf{q}}) = 0$ for $\mathsf{p} = \bar{\mathsf{q}}$, this shows that $\sup_{\mathsf{p} \in \Delta} \big(\mathsf{p} \cdot \mathsf{q} - D(\mathsf{p} \parallel \mathsf{p}_0)\big) = \log\big(\mathrm{E}_{\mathsf{p}_0}[e^{\mathsf{q}}]\big)$ and concludes the proof. □

Theorem 1. *Problem (2) and (3) are equivalent to the dual optimization problem $\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})$:*

$$\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p}). \tag{18}$$

Furthermore, let $\mathsf{p}^* = \mathrm{argmin}_{\mathsf{p}} F(\mathsf{p})$, then, for any $\epsilon > 0$ and any $\mathbf{w}$ such that $|G(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})| < \epsilon$, the following inequality holds: $D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq \epsilon$.

Proof. The proof follows by application of the Fenchel duality theorem (Theorem 5, Appendix A) to the optimization problem (3) with the functions f and g defined for all p and $\mathbf{u}$ by $f(\mathsf{p}) = D(\mathsf{p} \parallel \mathsf{p}_0) + I_\Delta(\mathsf{p})$ and $g(\mathbf{u}) = I_C(\mathbf{u})$ and with A the linear map defined by $A\mathsf{p} = \mathrm{E}_{\mathsf{p}}[\mathbf{\Phi}]$.

A is a bounded linear map since $\|A\| \leq \|\mathbf{\Phi}\|_\infty \leq \Lambda$ and $A^*\mathbf{w} = \mathbf{w} \cdot \mathbf{\Phi}$. Furthermore, define $\mathbf{u} \in \mathbb{F}$ by $\mathbf{u}_k = \mathrm{E}_{\widehat{\mathsf{p}}}[\mathbf{\Phi}_k]$. Then, $\mathbf{u}$ is in $A\,\mathrm{dom} f$ and is in C. Since $\beta_k > 0$ for all k, $\mathbf{u}$ is contained in $\mathrm{int}(C)$. $g = I_C$ equals zero over $\mathrm{int}(C)$ and is therefore continuous over $\mathrm{int}(C)$, thus g is continuous at $\mathbf{u} \in A\,\mathrm{dom} f$.

By Lemma 6, the conjugate of f is the function $f^*\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ defined by $f^*(\mathsf{q}) = \log\big(\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathsf{q}[x]}\big)$ for all $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$. The conjugate function of $g = I_C$ is the function g^* defined for all $\mathbf{w} \in \mathbb{R}^N$ by

$$\begin{aligned}
g^*(\mathbf{w}) &= \sup_{\mathbf{u} \in C} \big(\mathbf{w} \cdot \mathbf{u} - I_C(\mathbf{u})\big) \\
&= \sup_{\mathbf{u} \in C} (\mathbf{w} \cdot \mathbf{u}) \\
&= \sup_{\mathbf{u} \in C} \Big(\sum_{k=1}^p \mathbf{w}_k \cdot \mathbf{u}_k\Big) \\
&= \sum_{k=1}^p \sup_{\|\mathbf{u}_k - \mathrm{E}_S[\mathbf{\Phi}_k]\|_\infty \leq \beta_k} (\mathbf{w}_k \cdot \mathbf{u}_k) \\
&= \sum_{k=1}^p \mathbf{w}_k \cdot \mathop{\mathrm{E}}_S[\mathbf{\Phi}_k] + \sum_{k=1}^p \sup_{\|\mathbf{u}_k\|_\infty \leq \beta_k} (\mathbf{w}_k \cdot \mathbf{u}_k) \\
&= \mathop{\mathrm{E}}_S[\mathbf{w} \cdot \mathbf{\Phi}] + \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1,
\end{aligned}$$

where the penultimate equality holds by definition of the dual norm. In view of these identities, we can write

$$\begin{aligned}
&- f^*(A^*\mathbf{w}) - g^*(-\mathbf{w}) \\
&= -\log\Big(\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w} \cdot \mathbf{\Phi}(x)}\Big) + \mathop{\mathrm{E}}_S[\mathbf{w} \cdot \mathbf{\Phi}] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\
&= -\log Z_{\mathbf{w}} + \frac{1}{m} \sum_{i=1}^m \mathbf{w} \cdot \mathbf{\Phi}(x_i) - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\
&= \frac{1}{m} \sum_{i=1}^m \log \frac{e^{\mathbf{w} \cdot \mathbf{\Phi}(x_i)}}{Z_{\mathbf{w}}} - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\
&= \frac{1}{m} \sum_{i=1}^m \log\left[\frac{\mathsf{p}_{\mathbf{w}}[x_i]}{\mathsf{p}_0[x_i]}\right] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 = G(\mathbf{w}),
\end{aligned}$$

which proves that $\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p})$. For any

$\mathbf{w} \in \mathbb{R}^N$, we can write

$$
\begin{aligned}
&G(\mathbf{w}) - D(\mathsf{p}^* \parallel \mathsf{p}_0) \\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}^*[x]}{\mathsf{p}_0[x]}\right] \\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 - \\
&\qquad \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}^*[x]}{\mathsf{p}_{\mathbf{w}}[x]} \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] \\
&= -D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\
&\qquad + \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}_{\mathbf{w}}(x)}{\mathsf{p}_0(x)}\right].
\end{aligned}
$$

The difference of the last two terms can be expressed as follows

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] \\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{w}\cdot\mathbf{\Phi}(x) - \log Z_{\mathbf{w}}] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{w}\cdot\mathbf{\Phi}(x) - \log Z_{\mathbf{w}}] \\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{w}\cdot\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{w}\cdot\mathbf{\Phi}(x)].
\end{aligned}
$$

Plugging back this equality and rearranging yields

$$
\begin{aligned}
D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) &= D(\mathsf{p}^* \parallel \mathsf{p}_0) - G(\mathbf{w}) \\
&\quad - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 + \mathbf{w}\cdot\left(\mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}(x)]\right).
\end{aligned}
$$

The solution of the primal optimization, p^*, verifies the constraint $I_C(\mathrm{E}_{\mathsf{p}^*}[\mathbf{\Phi}]) = 0$, that is $\|\mathrm{E}_{x\sim\widehat{p}}[\mathbf{\Phi}_k(x)] - \mathrm{E}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}_k(x)]\|_\infty \leq \beta_k$ for all $k \in [1, p]$. By Hölder's inequality, this implies that

$$
\begin{aligned}
&- \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 + \mathbf{w}\cdot\left(\mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}(x)]\right) \\
&= - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 + \sum_{k=1}^p \mathbf{w}_k\cdot\left(\mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{\Phi}_k(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}_k(x)]\right) \\
&\leq - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 + \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 = 0.
\end{aligned}
$$

Thus, we can write, for any $\mathbf{w} \in \mathbb{R}^N$,

$$
D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq D(\mathsf{p}^* \parallel \mathsf{p}_0) - G(\mathbf{w}).
$$

Now, assume that $\mathbf{w}$ verifies $|G(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w})| \leq \epsilon$ for some $\epsilon > 0$. Then, $D(\mathsf{p}^* \parallel \mathsf{p}_0) - G(\mathbf{w}) = \sup_{\mathbf{w}} G(\mathbf{w}) - G(\mathbf{w}) \leq \epsilon$ implies $D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq \epsilon$. This concludes the proof of the theorem. □

Theorem 4. *Problem (10) is equivalent to dual optimization problem* $\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})$*:*

$$
\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w}) = \min_{\mathsf{p}} \widetilde{F}(\mathsf{p}). \tag{19}
$$

Furthermore, let $\mathsf{p}^* = \operatorname{argmin}_{\mathsf{p}} \widetilde{F}(\mathsf{p})$*. Then, for any* $\epsilon > 0$ *and any* $\mathbf{w}$ *such that* $|\widetilde{G}(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})| < \epsilon$*, we have* $\mathrm{E}_{x\sim\widehat{p}}\left[D\big(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big)\right] \leq \epsilon$.

Proof. The proof follows by application of the Fenchel duality theorem (Theorem 5, Appendix A) to the optimization problem (11) with the functions $\widetilde{f}$ and $\widetilde{g}$ defined for all p and $\mathbf{u}$ by $\widetilde{f}(\mathsf{p}) = \mathrm{E}_{x\sim\widehat{p}}\left[D\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) + I_\Delta\big(\mathsf{p}[\cdot|x]\big)\right]$ and $\widetilde{g}(\mathbf{u}) = I_C(\mathbf{u})$ and with A the linear map defined by $A\mathsf{p} = \mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\mathbf{\Phi}(x,y)]$.

A is a bounded linear map since $\|A\| \leq \|\mathbf{\Phi}\|_\infty \leq \Lambda$. Note that

$$
\begin{aligned}
A\mathsf{p} = \mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\mathbf{\Phi}(x,y)] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{\Phi}(x,y)\widehat{\mathsf{p}}[x]\mathsf{p}[y|x] \\
&= \sum_{x\in\mathrm{supp}(\widehat{p})} (\widehat{\mathsf{p}}[x]\mathbf{\Phi}(x,\cdot))\cdot(\mathsf{p}[\cdot|x]).
\end{aligned}
$$

Thus, the conjugate of A is defined for all $\mathbf{w} \in \mathbb{R}^N$ by $A^*\mathbf{w} = \mathbf{w}\cdot\big(\widehat{p}(x)\mathbf{\Phi}(x,y)\big)$. Furthermore, define $\mathbf{u} \in \mathbb{F}$ by $\mathbf{u}_k = \mathrm{E}_{(x,y)\sim S}[\mathbf{\Phi}_k(x,y)]$. Then, $\mathbf{u}$ is in $A\,\mathrm{dom} f$ and is in C. Since $\beta_k > 0$ for all k, $\mathbf{u}$ is contained in $\mathrm{int}(C)$. $g = I_C$ equals zero over $\mathrm{int}(C)$ and is therefore continuous over $\mathrm{int}(C)$, thus g is continuous at $\mathbf{u} \in A\,\mathrm{dom} f$.

The conjugate function of $\widetilde{f}$ is defined for all $\mathsf{q} = (\mathsf{q}[\cdot|x_i])_{i\in[1,m]}$ by

$$
\begin{aligned}
\widetilde{f}^*(\mathsf{q}) &= \sup_{\mathsf{p}[\cdot|x]\in\Delta} \{\langle \mathsf{p}, \mathsf{q}\rangle - \sum_{x\in\mathcal{X}} \widehat{\mathsf{p}}[x]\, D(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\} \\
&= \sup_{\mathsf{p}[\cdot|x]\in\Delta} \Big\{ \sum_{x\in\mathrm{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] \sum_{y\in\mathcal{Y}} \mathsf{p}[y|x]\mathsf{q}[y|x](\widehat{\mathsf{p}}[x])^{-1} \\
&\qquad\qquad - \sum_{x\in\mathrm{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] D(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\Big\} \\
&= \sum_{x\in\mathrm{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] \sup_{\mathsf{p}[\cdot|x]} \Big\{ \sum_{y\in\mathcal{Y}} \mathsf{p}[y|x] \left(\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}\right) \\
&\qquad\qquad - D(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\Big\} \\
&= \sum_{x\in\mathrm{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] f_x^*\left(\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}\right)
\end{aligned}
$$

where f_x is defined for all $x \in \mathcal{X}$ and $\mathsf{p}' \in \mathbb{R}^{\mathcal{Y}}$ by $f(\mathsf{p}') = D(\mathsf{p}' \parallel \mathsf{p}_0[\cdot|x])$ if $\mathsf{p}' \in \Delta$, $f(\mathsf{p}') = +\infty$ otherwise. By

Lemma 6, $f_x^*\left(\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}\right) = \log\left(\sum_{y\in\mathcal{Y}} \mathsf{p}_0[y|x] e^{\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}}\right)$, thus, $\widetilde{f}^*$ is given by

$$\widetilde{f}^*(\mathsf{q}) = \mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}\left[\log\left(\sum_{y\in\mathcal{Y}} \mathsf{p}_0[y|x] e^{\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}}\right)\right].$$

In view of these identities, we can write

$$\begin{aligned}
&-\widetilde{f}^*(A^*\mathbf{w}) - \widetilde{g}^*(-\mathbf{w})\\
&= -\mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}\left[\log\left(\sum_{y\in\mathcal{Y}} \mathsf{p}_0[y|x] e^{\mathbf{w}\cdot\mathbf{\Phi}(x,y)}\right)\right]\\
&\quad + \mathop{\mathrm{E}}_{S}[\mathbf{w}\cdot\mathbf{\Phi}] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&= -\mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}[\log Z_{\mathbf{w}}(x)] + \frac{1}{m}\sum_{i=1}^m \mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i) - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&= \frac{1}{m}\sum_{i=1}^m \log\frac{e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)}}{Z_{\mathbf{w}}(x_i)} - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&= \frac{1}{m}\sum_{i=1}^m \log\left[\frac{\mathsf{p}_{\mathbf{w}}[y_i|x_i]}{\mathsf{p}_0[y_i|x_i]}\right] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 = \widetilde{G}(\mathbf{w}),
\end{aligned}$$

which proves that $\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w}) = \min_{\mathsf{p}} \widetilde{F}(\mathsf{p})$. The second part of the proof is similar to that of Theorem 1. For any $\mathbf{w}\in\mathbb{R}^N$, we can write

$$\begin{aligned}
&\widetilde{G}(\mathbf{w}) - \mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])]\\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&\quad - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}^*[y|x]}{\mathsf{p}_0[y|x]}\right]\\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 -\\
&\quad \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}^*[y|x]}{\mathsf{p}_{\mathbf{w}}[y|x]}\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right]\\
&= -\mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&\quad + \mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right].
\end{aligned}$$

The difference of the last two terms can be expressed as follows

$$\begin{aligned}
&\mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right]\\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y) - \log Z_{\mathbf{w}}(x)]\\
&\quad - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y) - \log Z_{\mathbf{w}}(x)]\\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)].
\end{aligned}$$

Plugging back this equality and rearranging yields

$$\begin{aligned}
&\mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])]\\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - \widetilde{G}(\mathbf{w}) - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&\quad + \mathbf{w}\cdot\left[\mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)]\right].
\end{aligned}$$

The solution of the primal optimization, p^*, verifies the constraint $I_C(\mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}(x,y)]) = 0$, that is $\|\mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}_k(x,y)] - \mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{\Phi}_k(x,y)]\|_\infty \leq \beta_k$ for all $k\in[1,p]$. By Hölder's inequality, this implies that

$$\begin{aligned}
&-\sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1\\
&\quad + \mathbf{w}\cdot\left[\mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)]\right]\\
&\leq -\sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 + \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 = 0.
\end{aligned}$$

Thus, we can write, for any $\mathbf{w}\in\mathbb{R}^N$,

$$\begin{aligned}
&\mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])]\\
&\qquad\qquad \leq \mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - \widetilde{G}(\mathbf{w}).
\end{aligned}$$

Now, assume that $\mathbf{w}$ verifies $|\widetilde{G}(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})| \leq \epsilon$ for some $\epsilon > 0$. Then, $\mathrm{E}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - \widetilde{G}(\mathbf{w}) = \sup_{\mathbf{w}} \widetilde{G}(\mathbf{w}) - \widetilde{G}(\mathbf{w}) \leq \epsilon$ implies $\mathrm{E}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] \leq \epsilon$. This concludes the proof of the theorem. □

B. Pseudocode of StructMaxent1

Figure 2 shows the pseudocode of StructMaxent1.

StructMaxent1$(S = (x_1, \dots, x_m))$

for $t \leftarrow 1$ **to** T **do**
  for $k \leftarrow 1$ **to** p **and** $j \leftarrow 1$ **to** N_k **do**
    if $(w_{t-1,k,j} \neq 0)$ **then**
      $d_{k,j} \leftarrow \beta_k \operatorname{sgn}(w_{t-1,k,j}) + \epsilon_{t-1,k,j}$
    elseif $|\epsilon_{t-1,k,j}| \leq \beta_k$ **then**
      $d_{k,j} \leftarrow 0$
    else $d_{k,j} \leftarrow -\beta_k \operatorname{sgn}(\epsilon_{t-1,k,j}) + \epsilon_{t-1,k,j}$
  $(k, j) \leftarrow \operatorname{argmax}_{(k,j) \in [1,p] \times [1,N_k]} |d_{k,j}|$
  $\beta \leftarrow \frac{\overline{\Phi}^+_{t-1,k,j}\overline{\Phi}^-_{k,j}e^{-2w_{k,j}\Lambda} - \overline{\Phi}^+_{k,j}\overline{\Phi}^-_{t-1,k,j}}{\overline{\Phi}^+_{t-1,k,j}e^{-2w_{k,j}\Lambda} - \overline{\Phi}^-_{t-1,k,j}}$
  if $(|\beta| \leq \beta_k)$ **then**
    $\eta \leftarrow -w_{t-1,k,j}$
  elseif $(\beta > \beta_k)$ **then**
    $\eta \leftarrow \frac{1}{2\Lambda} \log \left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k - \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k - \overline{\Phi}^-_{k,j})} \right]$
  else $\eta \leftarrow \frac{1}{2\Lambda} \log \left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k + \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k + \overline{\Phi}^-_{k,j})} \right]$
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \eta \mathbf{e}_{k,j}$
  $\mathsf{p}_{\mathbf{w}_t} \leftarrow \frac{\mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}{\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}$
return $\mathsf{p}_{\mathbf{w}_t}$

Figure 2. Pseudocode of the StructMaxent1 algorithm. For all $(k, j) \in [1, p] \times [1, N_k]$, $\beta_k = 2\mathfrak{R}_m(H_k) + \beta$, $\epsilon_{t-1,k,j} = \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] - \mathrm{E}_S[\Phi_{k,j}]$ and, for any $s \in \{-1, +1\}$, $\overline{\Phi}^s_{t-1,k,j} = \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + s\Lambda$ and $\overline{\Phi}^s_{k,j} = \mathrm{E}_S[\Phi_{k,j}] + s\Lambda$. The closed-form solutions for the step size given here assume that the conditions (8) hold.

C. Algorithm

In this section we derive the step size for the StructMaxent1 and StructMaxent2 algorithms presented in Section 2.4 and Appendix B.

Observe that

$$F(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1}) \tag{20}$$
$$= \beta_k(|w_{k,j} + \eta| - |w_{k,j}|) - \eta \mathop{\mathrm{E}}_S[\Phi_{k,j}] + \log \left[\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta \Phi_{k,j}}] \right].$$

Since $\Phi_{k,j} \in [-\Lambda, +\Lambda]$, by the convexity of $x \mapsto e^{\eta x}$, we can write

$$e^{\eta \Phi_{k,j}} \leq \frac{\Lambda - \Phi_{k,j}}{2\Lambda} e^{-\eta\Lambda} + \frac{\Phi_{k,j} + \Lambda}{2\Lambda} e^{\eta\Lambda}.$$

Taking the expectation and the log yields

$$\log \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta \Phi_{k,j}}] \leq \log \left[\frac{\overline{\Phi}^+_{t-1,k,j} e^{\eta\Lambda} - \overline{\Phi}^-_{t-1,k,j} e^{-\eta\Lambda}}{2\Lambda} \right]$$
$$= -\eta\Lambda + \log \left[\frac{\overline{\Phi}^+_{t-1,k,j} e^{2\eta\Lambda} - \overline{\Phi}^-_{t-1,k,j}}{2\Lambda} \right],$$

where we used the following notation:

$$\overline{\Phi}^s_{t-1,k,j} = \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + s\Lambda \qquad \overline{\Phi}^s_{k,j} = \mathop{\mathrm{E}}_S[\Phi_{k,j}] + s\Lambda,$$

for all $(k, j) \in [1, p] \times [1, N_k]$ and $s \in \{-1, +1\}$.

Plugging back this inequality in (20) and ignoring constant terms, minimizing the resulting upper bound on $F(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1})$ becomes equivalent to minimizing $\psi(\eta)$ defined for all $\eta \in \mathbb{R}$ by

$$\psi(\eta) = \beta_k |w_{k,j} + \eta| - \eta \overline{\Phi}^+_{k,j} + \log \left[\overline{\Phi}^+_{t-1,k,j} e^{2\eta\Lambda} - \overline{\Phi}^-_{t-1,k,j} \right].$$

Let η^* denote the minimizer of $\psi(\eta)$. If $w_{t-1,k,j} + \eta^* = 0$, then the subdifferential of $|w_{t-1,k,j} + \eta|$ at η^* is the set $\{\nu \colon \nu \in [-1, +1]\}$. Thus, in that case, the subdifferential $\partial\psi(\eta^*)$, contains 0 iff there exists $\nu \in [-1, +1]$ such that

$$\beta_k \nu - \overline{\Phi}^+_{k,j} + \frac{2\Lambda \overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda} - \Phi^-_{t-1,k,j}} = 0$$
$$\Leftrightarrow \overline{\Phi}^+_{k,j} - \frac{2\Lambda \overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda} - \Phi^-_{t-1,k,j}} = \beta_k \nu.$$

Thus, the condition is equivalent to

$$\left| \overline{\Phi}^+_{k,j} - \frac{2\Lambda \overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda} - \Phi^-_{t-1,k,j}} \right| \leq \beta_k,$$

which can be rewritten as

$$\left| \frac{\overline{\Phi}^+_{t-1,k,j}\overline{\Phi}^-_{k,j} e^{-2w_{k,j}\Lambda} - \overline{\Phi}^+_{k,j}\overline{\Phi}^-_{t-1,k,j}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{k,j}\Lambda} - \overline{\Phi}^-_{t-1,k,j}} \right| \leq \beta_k.$$

If $w_{t-1,k,j} + \eta^* > 0$, then ψ is differentiable at η^* and $\psi'(\eta^*) = 0$, that is

$$\beta - \overline{\Phi}^+_{k,j} + \frac{2\Lambda \overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda} - \overline{\Phi}^-_{t-1,k,j}} = 0$$
$$\Leftrightarrow e^{2\eta^*\Lambda} = \frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k - \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k - \overline{\Phi}^-_{k,j})}$$
$$\Leftrightarrow \eta^* = \frac{1}{2\Lambda} \log \left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k - \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k - \overline{\Phi}^-_{k,j})} \right].$$

For the step size η^* to be in $\mathbb{R}$, the following conditions must be met:

$$(\overline{\Phi}^-_{t-1,k,j} \neq 0) \wedge (\overline{\Phi}^+_{t-1,k,j} \neq 0) \wedge$$
$$((\beta_k - \overline{\Phi}^+_{k,j}) < 0) \wedge ((\beta_k - \overline{\Phi}^-_{k,j}) \neq 0),$$

that is

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\}) \wedge (\mathop{\mathrm{E}}_S[\Phi_{k,j}] > -\Lambda + \beta_k). \tag{21}$$

The condition $w_{t-1,k,j} + \eta^* > 0$ is equivalent to $e^{2\eta^*\Lambda} > e^{-2w_{t-1,k,j}\Lambda}$, which, in view of the expression of $e^{2\eta^*\Lambda}$ given above can be written as

$$\frac{\overline{\Phi}^+_{t-1,k,j}\overline{\Phi}^-_{k,j}e^{-2w_{k,j}\Lambda} - \overline{\Phi}^+_{k,j}\overline{\Phi}^-_{t-1,k,j}}{\overline{\Phi}^+_{t-1,k,j}e^{-2w_{k,j}\Lambda} - \overline{\Phi}^-_{t-1,k,j}} > \beta_k.$$

Similarly, if $w_{t-1,k,j} + \eta^* < 0$, ψ is differentiable at η^* and $\psi'(\eta^*) = 0$, which gives

$$\eta^* = \frac{1}{2\Lambda}\log\left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k + \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k + \overline{\Phi}^-_{k,j})}\right].$$

Again for the step size η^* to be in $\mathbb{R}$, the following conditions must be met:

$$(\overline{\Phi}^-_{t-1,k,j} \neq 0) \wedge (\overline{\Phi}^+_{t-1,k,j} \neq 0) \wedge ((\beta_k + \overline{\Phi}^+_{k,j}) \neq 0) \wedge ((\beta_k + \overline{\Phi}^-_{k,j}) < 0),$$

that is

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\}) \wedge (\mathop{\mathrm{E}}_{S}[\Phi_{k,j}] < \Lambda - \beta_k).$$

Combining with condition 21, the following condition on Φ, Λ and β_k must be satisfied:

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\}) \wedge (-\Lambda + \beta_k < \mathop{\mathrm{E}}_{S}[\Phi_{k,j}] < \Lambda - \beta_k).$$

Figure 2 shows the pseudocode of our algorithm using the closed-form solution for the step size just presented.

An alternative method consists of using a somewhat looser upper bound for $\log \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}]$ using Hoeffding's lemma and $\Phi_{k,j} \in [-\Lambda, +\Lambda]$:

$$\log \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}] \leq \eta \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + \frac{\eta^2\Lambda^2}{2}.$$

Combining this inequality with (20) and disregarding constant terms, minimizing the resulting upper bound on $F(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1})$ becomes equivalent to minimizing $\varphi(\eta)$ defined for all $\eta \in \mathbb{R}$ by

$$\varphi(\eta) = \beta_k|w_{k,j} + \eta| + \eta\epsilon_{t-1,k,j} + \frac{\eta^2\Lambda^2}{2}.$$

Let η^* denote the minimizer of $\varphi(\eta)$. If $w_{t-1,k,j} + \eta^* = 0$, then the subdifferential of $|w_{t-1,k,j} + \eta|$ at η^* is the set $\{\nu\colon \nu \in [-1, +1]\}$. Thus, in that case, the subdifferential $\partial\varphi(\eta^*)$ contains 0 iff there exists $\nu \in [-1, +1]$ such that

$$\beta_k\nu + \epsilon_{t-1,k,j} + \eta^*\Lambda^2 = 0 \Leftrightarrow w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j} = \beta_k\nu.$$

The condition is therefore equivalent to

$$|w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j}| \leq \beta_k.$$

If $w_{t-1,k,j} + \eta^* > 0$, then φ is differentiable at η^* and $\varphi'(\eta^*) = 0$, that is

$$\beta_k + \epsilon_{t-1,k,j} + \eta^*\Lambda^2 = 0 \Leftrightarrow \eta^* = \frac{1}{\Lambda^2}[-\beta_k - \epsilon_{t-1,k,j}].$$

In view of that expression, the condition $w_{t-1,k,j} + \eta^* > 0$ is equivalent to

$$w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j} > \beta_k.$$

Similarly, if $w_{t-1,k,j} + \eta^* < 0$, φ is differentiable at η^* and $\varphi'(\eta^*) = 0$, which gives

$$\eta^* = \frac{1}{\Lambda^2}[\beta_k - \epsilon_{t-1,k,j}].$$

Figure 1 shows the pseudocode of our algorithm using the closed-form solution for the step size just presented.

D. Convergence analysis

In this section, we give convergence guarantees for both versions of the StructMaxent algorithm.

Theorem 3. *Let $(\mathbf{w}_t)_t$ be the sequence of parameter vectors generated by StructMaxent1 or StructMaxent2. Then, $(\mathbf{w}_t)_t$ converges to the optimal solution $\mathbf{w}^*$ of (6).*

Proof. We begin with the proof for StructMaxent2. Our proof is based on Lemma 19 of (Dudík et al., 2007), which implies that it suffices to show that $F(\mathbf{w}_t)$ admits a finite limit and that there exists a sequence $\mathbf{u}_t$ such that $R(\mathbf{u}_t, \mathbf{w}_t) \to 0$ as $t \to \infty$, where R is some auxiliary function. A function R is said to be *auxiliary* if

$$R(\mathbf{u}, \mathbf{w}) = I_C(\mathbf{u}) + \sum_{k=1}^{p}\beta_k\|\mathbf{w}_k\|_1 + \mathbf{w}\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] + \mathbf{w}\cdot\mathbf{u} + B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]),$$

where B is a Bregman divergences. We will use the Bregman divergence based on the squared difference:

$$B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]) = \frac{\|\mathbf{u} - \mathrm{E}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]\|_2^2}{2\Lambda^2}.$$

Let $g_0(\mathbf{u}) = I_C(\mathbf{u}) + \mathbf{w}\cdot\mathbf{u}$ and observe that using the same arguments as in the proof of Theorem 1, we can write

$$\begin{aligned} g_0^*(\mathbf{r}) &= \sup_{\mathbf{u}\in C}\left((\mathbf{r} - \mathbf{w})\cdot\mathbf{u} - I_C(\mathbf{u})\right) \\ &= (\mathbf{r} - \mathbf{w})\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] + \sum_{k=1}^{p}\beta_k\|\mathbf{r}_k - \mathbf{w}_k\|_1. \end{aligned}$$

Similarly, if $f_0(\mathbf{u}) = B(\mathbf{u} \parallel \mathrm{E}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}])$, then

$$\begin{aligned} f_0^*(\mathbf{r}) &= \sup_{\mathbf{u}}(\mathbf{r}\cdot\mathbf{u} - B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}])) \\ &= \frac{\Lambda^2\|\mathbf{r}\|_2}{2} + \mathbf{r}\cdot\mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]. \end{aligned}$$

Therefore, applying Theorem 5 with $A = I$, we obtain

$$\begin{aligned} \inf_{\mathbf{u}} R(\mathbf{u}, \mathbf{w}_t) = \sup_{\mathbf{r}} \Big(&- \frac{\Lambda^2\|\mathbf{r}\|_2}{2} - \mathbf{r}\cdot\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_t}}[\mathbf{\Phi}] - \mathbf{r}\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] \\ &+ \sum_{k=1}^{p}\beta_k(\|\mathbf{w}_{t,k}\|_1 - \|\mathbf{r}_k + \mathbf{w}_{t,k}\|_1)\Big), \end{aligned}$$

and we define $\mathbf{u}_t$ to be the solution of this optimization problem, which, in view of Theorem 5, does exist. We will now argue that $R(\mathbf{u}_t, \mathbf{w}_t) \to 0$ as $t \to \infty$. Note that

$$\begin{aligned} R(\mathbf{u}_t, \mathbf{w}_t) = -\sum_{k=1}^{p}\sum_{j=1}^{N_k}\inf_r \Big(&\frac{\lambda^2 r}{2} + r(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_t}}[\Phi_{k,j}] - \mathop{\mathrm{E}}_{S}[\Phi_{k,j}]) \\ &+ \beta_k|w_{t,k,j}| - \beta_k|r + w_{t,k,j}|\Big). \end{aligned}$$

Recall that, by definition of StructMaxent2, the following holds for all $(k, j) \in [1, p] \times [1, N_k]$:

$$\begin{aligned} &F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \qquad (22) \\ &\quad\geq -\inf_r \Big(\frac{\Lambda^2 r}{2} + r(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_t}}[\Phi_{k,j}] - \mathop{\mathrm{E}}_{S}[\Phi_{k,j}]) \\ &\qquad\qquad + \beta_k|w_{t,k,j}| - \beta_k|r + w_{t,k,j}|\Big) \\ &\quad\geq 0, \end{aligned}$$

where the last inequality follows by taking $r = 0$. Therefore, to complete the proof, it suffices to show that $\lim_{t\to\infty} F(\mathbf{w}_t)$ is finite, since then $F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \to 0$ and $R(\mathbf{u}_t, \mathbf{w}_t) \to 0$. By (22), $F(\mathbf{w}_t)$ is decreasing and it suffices to show that $F(\mathbf{w}_t)$ is bounded below. This is an immediate consequence of the feasibility of the optimization problem $\inf_{\mathbf{w}} F(\mathbf{w})$ which was established in Section 2.2 and the proof for StructMaxent2 is now complete.

The proof for StructMaxent1 requires the use a different Bregman divergence B defined as follows:

$$B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]) = \sum_{k=1}^{p}\sum_{j=1}^{N_k} D_0(\varphi_{kj}(\mathbf{u}) \parallel \varphi_{kj}(\mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}])),$$

where D_0 is unnormalized relative entropy, $\varphi_{kj}(\mathbf{u}) = ((\Lambda - u_{k,j}), (\Lambda + u_{k,j}))$ and $\|\mathbf{u}\|_\infty \leq \Lambda$. The rest of the argument remains the same. □

E. Bounds on Rademacher complexities

In this section, we give the proof of the upper bounds on Rademacher complexities given in (15):

$$\begin{aligned} \mathfrak{R}_m(H_k^{\text{mono}}) &\leq \sqrt{\frac{2k\log d}{m}} \\ \mathfrak{R}_m(H_k^{\text{trees}}) &\leq \sqrt{\frac{(4k+2)\log_2(d+2)\log(m+1)}{m}}. \end{aligned}$$

The first inequality is an immediate consequence of Massart's lemma, which states that

$$\frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\Big] \leq \frac{r\sqrt{2\log|A|}}{m},$$

where $A \subset \mathbb{R}^n$ is a finite set, $r = \max_{\mathbf{x}\in A}\|\mathbf{x}\|_2$ and σ_is are Rademacher random variables. If we take A to be the image of the sample under H_k^{mono} then $|A| \leq |H_k^{\text{mono}}| \leq d^k$. Moreover, if the features in H_1^{mono} are normalized to belong to $[-1, 1]$ then $\Lambda = 1$ and $r = \sqrt{m}$. Combining these results with Massart's lemma leads to the desired bound.

Now we derive the second bound of (15). Since each binary decision tree in H_k^{trees}, can be viewed as a binary classifier, Massart's lemma yields that

$$\mathfrak{R}_m(H_k^{\text{trees}}) \leq \sqrt{\frac{2\log\Pi_{H_k^{\text{trees}}}(m)}{m}},$$

where $\Pi_{H_k^{\text{trees}}}(m)$ is the growth function of H_k^{trees}. We use Sauer's lemma to bound the growth function: $\Pi_{H_k^{\text{trees}}}(m) \leq (em)^{\text{VC-dim}(H_k^{\text{trees}})}$. For the family of binary decision trees in dimension d it is known that $\text{VC-dim}(H_k^{\text{trees}}) \leq (2k+1)\log_2(d+2)$ (Mansour, 1997) and the desired bound follows.