12 Maximum Entropy Models

In this chapter, we introduce and discuss maximum entropy models, also known as *Maxent models*, a widely used family of algorithms for density estimation that can exploit rich feature sets. We first introduce the standard density estimation problem and briefly describe the Maximum Likelihood and Maximum a Posteriori solutions. Next, we describe a richer density estimation problem where the learner additionally has access to features. This is the problem addressed by Maxent models.

We introduce the key principle behind Maxent models and formulate their primal optimization problem. Next, we prove a duality theorem showing that Maxent models coincide with Gibbs distribution solutions of a regularized Maximum Likelihood problem. We present generalization guarantees for these models and also give an algorithm for solving their dual optimization problem using a coordinate descent technique. We further extend these models to the case where an arbitrary Bregman divergence is used with other norms, and prove a general duality theorem leading to an equivalent optimization problem with alternative regularizations. We also give a specific theoretical analysis of Maxent models with L_2-regularization, which are commonly used in applications.

12.1 Density estimation problem

Let $S = (x_1, \ldots, x_m)$ be a sample of size m drawn i.i.d. from an unknown distribution $\mathcal{D}$. Then, the density estimation problem consists of using that sample to select out of a family of possible distributions $\mathcal{P}$ a distribution p that is close to $\mathcal{D}$.

The choice of the family $\mathcal{P}$ is critical. A relatively small family may not contain $\mathcal{D}$ or even any distribution close to $\mathcal{D}$. On the other hand, a very rich family defined by a large set of parameters may make the task of selecting p very difficult if only a sample of a relatively modest size m is available.

12.1.1 Maximum Likelihood (ML) solution

One common solution adopted for selecting a distribution p is based on the *maximum likelihood principle*. This consists of choosing a distribution out of the family $\mathcal{P}$ that assigns the largest probability to the sample S observed. Thus, using the fact that the sample is drawn i.i.d., the solution p_{ML} selected by maximum likelihood is defined by

$$\mathsf{p}_{\mathrm{ML}} = \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \prod_{i=1}^{m} \mathsf{p}(x_i) = \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \sum_{i=1}^{m} \log \mathsf{p}(x_i). \tag{12.1}$$

The maximum likelihood principle can be equivalently formulated in terms of the relative entropy. Let $\widehat{\mathcal{D}}$ denote the empirical distribution corresponding to the sample S. Then, p_{ML} coincides with the distribution p with respect to which the empirical distribution $\widehat{\mathcal{D}}$ admits the smallest relative entropy:

$$\mathsf{p}_{\mathrm{ML}} = \operatorname*{argmin}_{\mathsf{p} \in \mathcal{P}} \mathsf{D}(\widehat{\mathcal{D}} \parallel \mathsf{p}). \tag{12.2}$$

This can be seen straightforwardly from the following:

$$\begin{aligned}
\mathsf{D}(\widehat{\mathcal{D}} \parallel \mathsf{p}) &= \sum_{x} \widehat{\mathcal{D}}(x) \log \widehat{\mathcal{D}}(x) - \sum_{x} \widehat{\mathcal{D}}(x) \log \mathsf{p}(x) \\
&= -\mathsf{H}(\widehat{\mathcal{D}}) - \sum_{x} \frac{\sum_{i=1}^{m} 1_{x=x_i}}{m} \log \mathsf{p}(x) \\
&= -\mathsf{H}(\widehat{\mathcal{D}}) - \sum_{i=1}^{m} \sum_{x} \frac{1_{x=x_i}}{m} \log \mathsf{p}(x) \\
&= -\mathsf{H}(\widehat{\mathcal{D}}) - \sum_{i=1}^{m} \frac{\log \mathsf{p}(x_i)}{m},
\end{aligned}$$

since the first term of the last expression, the negative entropy of the empirical distribution, does not vary with p.

As an example of application of the maximum likelihood principle, suppose we wish to estimate the bias p_0 of a coin from an i.i.d. sample $S = (x_1, \ldots, x_m)$ where $x_i \in \{\mathsf{h}, \mathsf{t}\}$ with h denoting heads and t tails. $p_0 \in [0, 1]$ is the probability of h according to the unknown distribution $\mathcal{D}$. Let $\mathcal{P}$ be the family of all distributions $\mathsf{p} = (p, 1-p)$ where $p \in [0, 1]$ is an arbitrary possible bias value. Let n_{h} denote the number of occurrences of h in S. Then, choosing $\mathsf{p} = (\widehat{p}_S, 1 - \widehat{p}_S) = \widehat{\mathcal{D}}$ where $\widehat{p}_S = \frac{n_{\mathsf{h}}}{m}$ leads to $\mathsf{D}(\widehat{\mathcal{D}} \parallel \mathsf{p}) = 0$, which, by (12.2), shows that $\mathsf{p}_{\mathrm{ML}} = \widehat{\mathcal{D}}$. Thus, the maximum likelihood estimate p_{ML} of the bias is the empirical value

$$p_{\mathrm{ML}} = \frac{n_{\mathsf{h}}}{m}. \tag{12.3}$$

12.1.2 Maximum a Posteriori (MAP) solution

An alternative solution based on the so-called *Maximum a Posteriori* solution consists of selecting a distribution $\mathsf{p} \in \mathcal{P}$ that is the most likely, given the observed sample S and a prior $\mathbb{P}[\mathsf{p}]$ over the distributions $\mathsf{p} \in \mathcal{P}$. By the Bayes rule, the problem can be formulated as follows:

$$\mathsf{p}_{\text{MAP}} = \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \mathbb{P}[\mathsf{p}|S] = \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \frac{\mathbb{P}[S|\mathsf{p}]\,\mathbb{P}[\mathsf{p}]}{\mathbb{P}[S]} = \operatorname*{argmax}_{\mathsf{p} \in \mathcal{P}} \mathbb{P}[S|\mathsf{p}]\,\mathbb{P}[\mathsf{p}]. \tag{12.4}$$

Notice that, for a uniform prior, $\mathbb{P}[\mathsf{p}]$ is a constant and the Maximum a Posteriori solution then coincides with the Maximum Likelihood solution. The following is a standard example illustrating the MAP solution and its difference with the ML solution.

Example 12.1 (Application of the MAP solution) Suppose we need to determine if a patient has a rare disease, given a laboratory test of that patient. We consider a set of two simple distributions: d (disease with probability one) and $\bar{\mathsf{d}}$ (no disease with probability one), thus $\mathcal{P} = \{\mathsf{d}, \bar{\mathsf{d}}\}$. The laboratory test is either pos (positive) or neg (negative), thus $S \in \{\text{pos}, \text{neg}\}$.

Suppose that the disease is rare, say $\mathbb{P}[\mathsf{d}] = .005$ and that the laboratory is relatively accurate: $\mathbb{P}[\text{pos}|\mathsf{d}] = .98$, and $\mathbb{P}[\text{neg}|\bar{\mathsf{d}}] = .95$. Then, if the test is positive, what should be the diagnosis? We can compute the right-hand side of (12.4) for both outcomes, given the positive test result, to determine the MAP estimate:

$$\begin{aligned}
\mathbb{P}[\text{pos}|\mathsf{d}]\,\mathbb{P}[\mathsf{d}] &= .98 \times .005 = .0049\\
\mathbb{P}[\text{pos}|\bar{\mathsf{d}}]\,\mathbb{P}[\bar{\mathsf{d}}] &= (1 - .95) \times (1 - .005) = .04975 > .0049.
\end{aligned}$$

Thus, in this case, the MAP prediction is no disease: according to the MAP solution, with the values indicated, a patient with a positive test result is nonetheless more likely not to have the disease!

We will not analyze the properties of the Maximum Likelihood and Maximum a Posteriori solutions here, which depend on the size of the sample and the choice of the family $\mathcal{P}$. Instead, we will consider a richer density estimation problem where the learner has access to features, which is the learning problem addressed by Maximum Entropy (Maxent) models.

12.2 Density estimation problem augmented with features

As with the standard density estimation problem, we consider a scenario where the learner receives a sample $S = (x_1, \ldots, x_m) \subseteq \mathcal{X}$ of size m drawn i.i.d. according to some distribution $\mathcal{D}$. But, here, additionally, we assume that the learner has access to a feature mapping $\mathbf{\Phi}$ from $\mathcal{X}$ to $\mathbb{R}^N$ with $\|\mathbf{\Phi}\|_\infty \leq r$. In the most general case, we may have $N = +\infty$. We will denote by $\mathcal{H}$ a family of real-valued functions containing the component feature functions Φ_j with $j \in [N]$. Different feature functions can be considered in practice. $\mathcal{H}$ may be the family of threshold functions $\mathbf{x} \mapsto 1_{x_i \leq \theta}$, $\mathbf{x} \in \mathbb{R}^n$, $\theta \in \mathbb{R}$, defined over n variables as for boosting stumps, or it may be a family of functions defined by more complex decision trees or regression trees. Other features often used in practice are monomials of degree k based on the input variables. To simplify the presentation, in what follows, we will assume that the input set $\mathcal{X}$ is finite.

12.3 Maxent principle

Maxent models are derived from a principle based on the key property that, with high probability, the empirical average of any feature is close to its true average. By the Rademacher complexity bound, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ over the choice of a sample S of size m:

$$\Big\| \mathop{\mathbb{E}}_{x \sim \mathcal{D}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] \Big\|_\infty \leq 2\mathfrak{R}_m(\mathcal{H}) + r\sqrt{\frac{\log \frac{2}{\delta}}{2m}}, \tag{12.5}$$

where we denote by $\widehat{\mathcal{D}}$ the empirical distribution defined by the sample S. This is the theoretical guarantee that guides the definition of the Maxent principle.

Let p_0 be a distribution over $\mathcal{X}$ with $\mathsf{p}_0(x) > 0$ for all $x \in \mathcal{X}$, which is often chosen to be the uniform distribution. Then, the *Maxent principle* consists of seeking a distribution p that is as agnostic as possible, that is as close as possible to the uniform distribution or, more generally, to a prior p_0, while verifying an inequality similar to (12.5):

$$\Big\| \mathop{\mathbb{E}}_{x \sim \mathsf{p}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] \Big\|_\infty \leq \lambda, \tag{12.6}$$

where $\lambda \geq 0$ is a parameter. Here, closeness is measured using the relative entropy. Choosing $\lambda = 0$ corresponds to standard *Maxent* or *unregularized Maxent* and to requiring the expectation of the features with respect to p to precisely match the empirical averages. As we will see later, its relaxation, that is the inequality case ($\lambda \neq 0$), translates into a regularization. Notice that, unlike Maximum likelihood, the Maxent principle does not require specifying a family of probability distributions $\mathcal{P}$ to choose from.

12.4 Maxent models

Let Δ denote the simplex of all distributions over $\mathcal{X}$, then, the Maxent principle can be formulated as the following optimization problem:

$$\begin{aligned} &\min_{\mathsf{p}\in\Delta} \mathsf{D}(\mathsf{p} \,\|\, \mathsf{p}_0) \\ \text{subject to: } &\left\| \mathop{\mathbb{E}}_{x\sim\mathsf{p}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] \right\|_\infty \leq \lambda. \end{aligned} \tag{12.7}$$

This defines a convex optimization problem since the relative entropy D is convex with respect to its arguments (appendix E), since the constraints are affine, and since Δ is a convex set. The solution is in fact unique since the relative entropy is strictly convex. The empirical distribution is clearly a feasible point, thus problem (12.7) is feasible.

For a uniform prior p_0, problem (12.7) can be equivalently formulated as an entropy maximization, which explains the name given to these models. Let $\mathsf{H}(\mathsf{p}) = -\sum_{x\in\mathcal{X}} \mathsf{p}(x) \log \mathsf{p}(x)$ denote the entropy of p. Then, the objective function of (12.7) can be rewritten as follows:

$$\begin{aligned} \mathsf{D}(\mathsf{p} \,\|\, \mathsf{p}_0) &= \sum_{x\in\mathcal{X}} \mathsf{p}(x) \log \frac{\mathsf{p}(x)}{\mathsf{p}_0(x)} \\ &= -\sum_{x\in\mathcal{X}} \mathsf{p}(x) \log \mathsf{p}_0(x) + \sum_{x\in\mathcal{X}} \mathsf{p}(x) \log \mathsf{p}(x) \\ &= \log|\mathcal{X}| - \mathsf{H}(\mathsf{p}). \end{aligned}$$

Thus, since $\log|\mathcal{X}|$ is a constant, minimizing the relative entropy $\mathsf{D}(\mathsf{p} \,\|\, \mathsf{p}_0)$ is then equivalent to maximizing $\mathsf{H}(\mathsf{p})$.

Maxent models are the solutions of the optimization problem just described. As already discussed, they admit two important benefits: they are based on a fundamental theoretical guarantee of closeness of empirical and true feature averages, and they do not require specifying a particular family of distributions $\mathcal{P}$. In the next sections, we will further analyze the properties of Maxent models.

12.5 Dual problem

Here, we derive an equivalent dual problem for (12.7) which, as we will show, can be formulated as a regularized maximum likelihood problem over the family of *Gibbs distributions*.

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 Maxent optimization problem (12.7) can be equivalently expressed as the unconstrained optimization problem $\min_{\mathsf{p}} F(\mathsf{p})$ with, for all $\mathsf{p} \in \mathbb{R}^{\mathcal{X}}$,

$$F(\mathsf{p}) = \widetilde{\mathsf{D}}(\mathsf{p} \,\|\, \mathsf{p}_0) + I_{\mathcal{C}}(\mathop{\mathbb{E}}_{\mathsf{p}}[\mathbf{\Phi}]), \tag{12.8}$$

with $\widetilde{\mathsf{D}}(\mathsf{p} \,\|\, \mathsf{p}_0) = \mathsf{D}(\mathsf{p} \,\|\, \mathsf{p}_0)$ if p is in the simplex Δ, $\widetilde{\mathsf{D}}(\mathsf{p} \,\|\, \mathsf{p}_0) = +\infty$ otherwise, and with $\mathcal{C} \subseteq \mathbb{R}^N$ the convex set defined by $\mathcal{C} = \{\mathbf{u} \colon \|\mathbf{u} - \mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x,y)]\,\|_\infty \leq \lambda\}$.

The general form of a Gibbs distribution $\mathsf{p}_{\mathbf{w}}$ with prior p_0, parameter $\mathbf{w}$, and feature vector $\mathbf{\Phi}$ is

$$\mathsf{p}_{\mathbf{w}}[x] = \frac{\mathsf{p}_0[x] e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}}{Z(\mathbf{w})} \tag{12.9}$$

where $Z(\mathbf{w}) = \sum_{x\in\mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}$ is a normalization factor also known as the *partition function*. Let G be the function defined for all $\mathbf{w} \in \mathbb{R}^N$ by

$$G(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \log\left[\frac{\mathsf{p}_{\mathbf{w}}[x_i]}{\mathsf{p}_0[x_i]}\right] - \lambda\|\mathbf{w}\|_1. \tag{12.10}$$

Then, the following theorem shows the equivalence of the primal problem (12.7) or (12.8) and a dual problem based on G.

Theorem 12.2 (Maxent duality) *Problems* (12.7) *or* (12.8) *are equivalent to the optimization problem* $\sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w})$*:*

$$\sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p}). \tag{12.11}$$

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

Proof: The first part of the proof follows by application of the Fenchel duality theorem (theorem B.39) to the optimization problem (12.8) with the functions f, g, and A defined for all $\mathsf{p} \in \mathbb{R}^{\mathcal{X}}$ and $\mathbf{u} \in \mathbb{R}^N$ by $f(\mathsf{p}) = \widetilde{\mathsf{D}}(\mathsf{p} \,\|\, \mathsf{p}_0)$, $g(\mathbf{u}) = I_{\mathcal{C}}(\mathbf{u})$ and $A\mathsf{p} = \sum_{x\in\mathcal{X}} \mathsf{p}(x)\mathbf{\Phi}(x)$. A is a bounded linear map since for any p, we have $\|A\mathsf{p}\| \leq \|\mathsf{p}\|_1 \sup_x \|\mathbf{\Phi}(x)\|_\infty \leq r\|\mathsf{p}\|_1$. Also, notice that for all $\mathbf{w} \in \mathbb{R}^N$, $A^*\mathbf{w} = \mathbf{w}\cdot\mathbf{\Phi}$.

Consider $\mathbf{u}_0 \in \mathbb{R}^N$ defined by $\mathbf{u}_0 = \mathbb{E}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] = A\widehat{\mathcal{D}}$. Since $\widehat{\mathcal{D}}$ is in $\Delta = \operatorname{dom}(f)$, $\mathbf{u}_0$ is in $A(\operatorname{dom}(f))$. Furthermore, since $\lambda > 0$, $\mathbf{u}_0$ is in $\operatorname{int}(\mathcal{C})$. $g = I_{\mathcal{C}}$ equals zero over $\operatorname{int}(\mathcal{C})$ and is therefore continuous over $\operatorname{int}(\mathcal{C})$, thus g is continuous at $\mathbf{u}_0$ and we have $\mathbf{u}_0 \in A(\operatorname{dom}(f)) \cap \operatorname{cont}(g)$. Thus, the assumptions of Theorem B.39 hold.

By Lemma B.37, 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 $g = I_{\mathcal{C}}$ is the function g^* defined for all $\mathbf{w} \in \mathbb{R}^N$ by

$$\begin{aligned}
g^*(\mathbf{w}) &= \sup_{\mathbf{u}} \big(\mathbf{w} \cdot \mathbf{u} - I_{\mathcal{C}}(\mathbf{u})\big) = \sup_{\mathbf{u} \in \mathcal{C}} (\mathbf{w} \cdot \mathbf{u}) \\
&= \sup_{\|\mathbf{u} - \mathbb{E}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}]\|_\infty \leq \lambda} (\mathbf{w} \cdot \mathbf{u}) \\
&= \mathbf{w} \cdot \mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] + \sup_{\|\mathbf{u}\|_\infty \leq \lambda} (\mathbf{w} \cdot \mathbf{u}) \\
&= \mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{w} \cdot \mathbf{\Phi}] + \lambda \|\mathbf{w}\|_1,
\end{aligned}$$

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

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

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

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

$$\begin{aligned}
& G(\mathbf{w}) - \mathsf{D}(\mathsf{p}^* \,\|\, \mathsf{p}_0) + \mathsf{D}(\mathsf{p}^* \,\|\, \mathsf{p}_{\mathbf{w}}) \\
&= \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}} \left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \lambda \|\mathbf{w}\|_1 - \mathop{\mathbb{E}}_{x \sim \mathsf{p}^*} \left[\log \frac{\mathsf{p}^*[x]}{\mathsf{p}_0[x]}\right] + \mathop{\mathbb{E}}_{x \sim \mathsf{p}^*} \left[\log \frac{\mathsf{p}^*[x]}{\mathsf{p}_{\mathbf{w}}[x]}\right] \\
&= -\lambda \|\mathbf{w}\|_1 + \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}} \left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \mathop{\mathbb{E}}_{x \sim \mathsf{p}^*} \left[\log \frac{\mathsf{p}_{\mathbf{w}}(x)}{\mathsf{p}_0(x)}\right] \\
&= -\lambda \|\mathbf{w}\|_1 + \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}} [\mathbf{w} \cdot \mathbf{\Phi}(x) - \log Z(\mathbf{w})] - \mathop{\mathbb{E}}_{x \sim \mathsf{p}^*} [\mathbf{w} \cdot \mathbf{\Phi}(x) - \log Z(\mathbf{w})] \\
&= -\lambda \|\mathbf{w}\|_1 + \mathbf{w} \cdot \Big[\mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}} [\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x \sim \mathsf{p}^*} [\mathbf{\Phi}(x)]\Big].
\end{aligned}$$

The solution of the primal optimization, p^*, verifies the constraint $I_{\mathcal{C}}(\mathbb{E}_{\mathsf{p}^*}[\mathbf{\Phi}]) = 0$, that is $\| \mathbb{E}_{x \sim \widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] - \mathbb{E}_{x \sim \mathsf{p}^*}[\mathbf{\Phi}(x)]\|_\infty \leq \lambda$. By Hölder's inequality, this implies the following inequality:

$$-\lambda \|\mathbf{w}\|_1 + \mathbf{w} \cdot \Big[\mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}} [\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x \sim \mathsf{p}^*} [\mathbf{\Phi}(x)]\Big] \leq -\lambda \|\mathbf{w}\|_1 + \lambda \|\mathbf{w}\|_1 = 0.$$

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

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

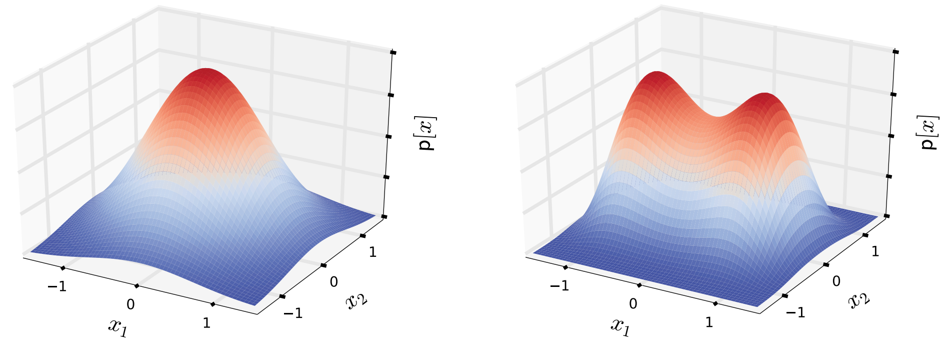
Now, assume that $\mathbf{w}$ verifies $|G(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})| \leq \epsilon$ for some $\epsilon > 0$. Then, $\mathsf{D}(\mathsf{p}^* \,\|\, \mathsf{p}_0) - G(\mathbf{w}) = (\sup_{\mathbf{w}} G(\mathbf{w})) - G(\mathbf{w}) \leq \epsilon$ implies $\mathsf{D}(\mathsf{p}^*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

In view of the theorem, if $\mathbf{w}$ is an ϵ-solution of the dual optimization problem, then $\mathsf{D}(\mathsf{p}^* \,\|\, \mathsf{p}_{\mathbf{w}}) \leq \epsilon$, which, by Pinsker's inequality (Proposition E.7) 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 Maxent problem can be determined by solving the dual problem, which can be written equivalently as follows:

$$\inf_{\mathbf{w} \in \mathbb{R}^N} \lambda \|\mathbf{w}\|_1 - \frac{1}{m} \sum_{i=1}^{m} \log \mathsf{p}_{\mathbf{w}}[x_i]. \tag{12.12}$$

Notice that the solution may not be achieved for any finite $\mathbf{w}$ for $\lambda = 0$, which is why the infimum is needed. This result may seem surprising since it shows that Maxent coincides with Maximum Likelihood ($\lambda = 0$) or regularized Maximum Likelihood ($\lambda > 0$) over a specific family $\mathcal{P}$ of distributions, that of Gibbs distributions, while, as pointed out earlier, the Maxent principle does not explicitly specify any family $\mathcal{P}$. What can then explain that the solution of Maxent belongs to the specific family of Gibbs distributions? The reason is the specific choice of the relative entropy as the measure of closeness of p to the prior distribution p_0. Other measures of closeness between distributions lead to different forms for the solution. Thus, in some sense, the choice of the measure of closeness is the (dual) counterpart of that of the family of distributions $\mathcal{P}$ in maximum likelihood.

Gibbs distributions form a very rich family. In particular, when $\mathcal{X}$ is a subset of a vector space and the features $\Phi_j(\mathbf{x})$ associated to $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$ are monomials of degree at most 2 based on the input variables x_j, that is $x_j x_k$, x_j, or the constant $a \in \mathbb{R}$, then $\mathbf{w} \cdot \mathbf{\Phi}(x)$ is a quadratic form as a function of the x_js. Thus, Gibbs distributions include the family of distributions defined by the normalized exponential of a quadratic form, which includes as a special case Gaussian distributions but also bi-modal distributions and normalized exponentials of non-positive definite quadratic forms. More complex multi-modal distributions can be further defined using higher-order monomials or more complex functions of the input variables. Figure 12.1 shows two examples of Gibbs distributions illustrating the richness of this family.



(a) (b)

Figure 12.1
Examples of Gibbs distributions in $\mathbb{R}^2$. (a) Unimodal Gaussian distribution $\mathsf{p}[(x_1, x_2)] = \frac{e^{-(x_1^2+x_2^2)}}{Z}$; (b) Bimodal distribution $\mathsf{p}[(x_1, x_2)] = \frac{e^{-(x_1^4+x_2^4)+x_1^2-x_2^2}}{Z}$. In each case, Z is a normalization factor.

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

Theorem 12.3 *Fix $\delta > 0$. Let $\widehat{\mathbf{w}}$ be a solution of the optimization* (12.12) *for $\lambda = 2\mathfrak{R}_m(\mathcal{H}) + r\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$. Then, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{D}$, the following inequality holds:*

$$\mathcal{L}_\mathcal{D}(\widehat{\mathbf{w}}) \le \inf_\mathbf{w} \mathcal{L}_\mathcal{D}(\mathbf{w}) + 2\|\mathbf{w}\|_1 \left[2\mathfrak{R}_m(\mathcal{H}) + r\sqrt{\frac{\log\frac{2}{\delta}}{2m}}\right].$$

Proof: Using the definition of $\mathcal{L}_\mathcal{D}(\mathbf{w})$ and $\mathcal{L}_S(\mathbf{w})$, Hölder's inequality, and inequality (12.5), with probability at least $1-\delta$, the following holds:

$$\mathcal{L}_\mathcal{D}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) = \widehat{\mathbf{w}} \cdot \big[\mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] - \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbf{\Phi}]\big] \le \|\widehat{\mathbf{w}}\|_1 \, \|\mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] - \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbf{\Phi}]\|_\infty \le \lambda\|\widehat{\mathbf{w}}\|_1.$$

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

$$\begin{aligned}\mathcal{L}_\mathcal{D}(\widehat{\mathbf{w}}) - \mathcal{L}_\mathcal{D}(\mathbf{w}) &= \mathcal{L}_\mathcal{D}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_\mathcal{D}(\mathbf{w})\\ &\le \lambda\|\widehat{\mathbf{w}}\|_1 + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_\mathcal{D}(\mathbf{w})\\ &\le \lambda\|\mathbf{w}\|_1 + \mathcal{L}_S(\mathbf{w}) - \mathcal{L}_\mathcal{D}(\mathbf{w}) \le 2\lambda\|\mathbf{w}\|_1,\end{aligned}$$

where we used for the last inequality the left inequality counterpart of inequality (12.5). This concludes the proof. □

Assume that $\mathbf{w}^*$ achieves the infimum of the loss, that is $\mathcal{L}_{\mathcal{D}}(\mathbf{w}^*) = \inf_{\mathbf{w}} \mathcal{L}_{\mathcal{D}}(\mathbf{w})$ and that $\mathfrak{R}_m(\mathcal{H}) = (1/\sqrt{m})$. Then, the theorem shows that, with high probability, the following inequality holds:

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \leq \inf_{\mathbf{w}} \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + O\bigg(\frac{\|\mathbf{w}^*\|_1}{\sqrt{m}}\bigg).$$

12.7 Coordinate descent algorithm

The dual objective function in the optimization (12.12) is convex since the Lagrange dual is always concave (appendix B). Ignoring the constant term $-\frac{1}{m}\sum_{i=1}^m \log \mathsf{p}_0[x_i]$, the optimization problem (12.12) can be rewritten as $\inf_{\mathbf{w}} J(\mathbf{w})$ with

$$J(\mathbf{w}) = \lambda\|\mathbf{w}\|_1 - \mathbf{w} \cdot \mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] + \log\left[\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}\right].$$

Note in particular that the function $\mathbf{w} \mapsto \log\left[\sum_{x\in\mathcal{X}} \mathsf{p}_0[x]e^{\mathbf{w}\cdot\mathbf{\Phi}(x)}\right]$ is convex as the conjugate function f^* of the function f defined in the proof of Theorem 12.2.

Different optimization techniques can be used to solve this convex optimization problem, including standard stochastic gradient descent and several special-purpose techniques. In this section, we will describe a solution based on coordinate descent which is particularly advantageous in presence of a very large number of features.

Function J is not differentiable but since it is convex, it admits a subdifferential at any point. The Maxent algorithm we describe consists of applying coordinate descent to the objective function (12.12).

Direction Let $\mathbf{w}_{t-1}$ denote the weight vector defined after $(t-1)$ iterations. At each iteration $t \in [T]$, the direction $\mathbf{e}_j$, $j \in [N]$ considered by coordinate descent is $\delta J(\mathbf{w}_{t-1}, \mathbf{e}_j)$. If $w_{t-1,j} \neq 0$, then J admits a directional derivative along $\mathbf{e}_j$ given by

$$J'(\mathbf{w}_{t-1}, \mathbf{e}_j) = \lambda \operatorname{sgn}(w_{t-1,j}) + \epsilon_{t-1,j}.$$

where $\epsilon_{t-1,j} = \mathbb{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_j] - \mathbb{E}_{\widehat{\mathcal{D}}}[\Phi_j]$. If $w_{t-1,j} = 0$, J admits right and left directional derivatives along $\mathbf{e}_j$:

$$J'_+(\mathbf{w}_{t-1}, \mathbf{e}_j) = \lambda + \epsilon_{t-1,j} \qquad J'_-(\mathbf{w}_{t-1}, \mathbf{e}_j) = -\lambda + \epsilon_{t-1,j}.$$

```
CDMaxent($S = (x_1, \ldots, x_m)$)
for $t \leftarrow 1$ to $T$ do
    for $j \leftarrow 1$ to $N$ do
        if $(w_{t-1,j} \neq 0)$ then
            $d_j \leftarrow \lambda \operatorname{sgn}(w_{t-1,j}) + \epsilon_{t-1,j}$
        elseif $|\epsilon_{t-1,j}| \leq \lambda$ then
            $d_j \leftarrow 0$
        else $d_j \leftarrow -\lambda \operatorname{sgn}(\epsilon_{t-1,j}) + \epsilon_{t-1,j}$
    $j \leftarrow \operatorname*{argmax}_{j \in [N]} |d_j|$
    if $(|w_{t-1,j} r^2 - \epsilon_{t-1,j}| \leq \lambda)$ then
        $\eta \leftarrow -w_{t-1,j}$
    elseif $(w_{t-1,j} r^2 - \epsilon_{t-1,j} > \lambda)$ then
        $\eta \leftarrow \frac{1}{r^2}[-\lambda - \epsilon_{t-1,j}]$
    else $\eta \leftarrow \frac{1}{r^2}[\lambda - \epsilon_{t-1,j}]$
    $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \eta \mathbf{e}_j$
    $\mathsf{p}_{\mathbf{w}_t} \leftarrow \frac{\mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}{\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}$
return $\mathsf{p}_{\mathbf{w}_t}$
```

Figure 12.2
Pseudocode of the Coordinate Descent Maxent algorithm. For all $j \in [N]$, $\epsilon_{t-1,j} = \mathbb{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_j] - \mathbb{E}_{\widehat{\mathcal{D}}}[\Phi_j]$.

Thus, in summary, we can define, for all $j \in [N]$,

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

The coordinate descent algorithm selects the direction $\mathbf{e}_j$ with the largest absolute value of $\delta J(\mathbf{w}_{t-1}, \mathbf{e}_j)$.

Step size Given the direction $\mathbf{e}_j$, the optimal step value η is given by $\operatorname{argmin}_\eta J(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_j)$. η can be found via a line search or other numerical methods. A closed-form expression for the step can also be derived by minimizing an upper bound on $J(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_j)$. Notice that we can write

$$J(\mathbf{w}_{t-1}+\eta\,\mathbf{e}_j)-J(\mathbf{w}_{t-1}) = \lambda(|w_j+\eta|-|w_j|)-\eta\operatorname*{\mathbb{E}}_{S}[\Phi_j]+\log\left[\operatorname*{\mathbb{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_j}]\right]. \quad (12.13)$$

In view of $\Phi_j \in [-r, +r]$, by Hoeffding's lemma, the following inequality holds:

$$\log \operatorname*{\mathbb{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_j}] \leq \eta \operatorname*{\mathbb{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_j] + \frac{\eta^2 r^2}{2}.$$

Combining this inequality with (12.13) and disregarding constant terms, minimizing the resulting upper bound on $J(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_j) - J(\mathbf{w}_{t-1})$ becomes equivalent to minimizing $\varphi(\eta)$ defined for all $\eta \in \mathbb{R}$ by

$$\varphi(\eta) = \lambda|w_j + \eta| + \eta\epsilon_{t-1,j} + \frac{\eta^2 r^2}{2}.$$

Let η^* denote the minimizer of $\varphi(\eta)$. If $w_{t-1,j} + \eta^* = 0$, then the subdifferential of $|w_{t-1,j} + \eta|$ 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

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

The condition is therefore equivalent to $|w_{t-1,j} r^2 - \epsilon_{t-1,j}| \leq \lambda$. If $w_{t-1,j} + \eta^* > 0$, then φ is differentiable at η^* and $\varphi'(\eta^*) = 0$, that is

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

In view of that expression, the condition $w_{t-1,j} + \eta^* > 0$ is equivalent to $w_{t-1,j} r^2 - \epsilon_{t-1,j} > \lambda$. Similarly, if $w_{t-1,j} + \eta^* < 0$, φ is differentiable at η^* and $\varphi'(\eta^*) = 0$, which gives

$$\eta^* = \frac{1}{r^2}[\lambda - \epsilon_{t-1,j}].$$

Figure 12.2 shows the pseudocode of the Coordinate Descent Maxent algorithm using the closed-form solution for the step size just presented. Note that we do not need to update distribution $\mathsf{p}_{\mathbf{w}_t}$ at every iteration of the algorithm (line 15) and we only need to be able to compute $\mathbb{E}_{\mathsf{p}_{\mathbf{w}_t}}[\Phi_j]$ which defines $\epsilon_{t,j}$. Various approximation strategies can be used to do this efficiently, including for instance rejection sampling techniques.

12.8 Extensions

As already pointed out, the Gibbs distribution form of the Maxent models is tightly related to the choice of the divergence (relative entropy) used to measure closeness in the Maxent principle. For distributions, the relative entropy coincides with the unnormalized relative entropy, which is a Bregman divergence. Maxent models can be generalized by using an arbitrary Bregman divergence B_Ψ instead (appendix E), where Ψ is a convex function. Moreover, other norms $\|\cdot\|$ can be used to bound the difference of the empirical and true average feature vectors. This leads to the following general primal optimization problem for Maxent models:

$$\begin{aligned} &\min_{\mathsf{p}\in\Delta} \mathsf{B}_\Psi(\mathsf{p}\,\|\,\mathsf{p}_0) \\ &\text{subject to: } \Big\| \mathop{\mathbb{E}}_{x\sim\mathsf{p}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] \Big\| \leq \lambda, \end{aligned} \tag{12.14}$$

which, as with (12.7), is a convex optimization problem since B_Ψ is convex with respect to its first argument and in fact strictly convex if Ψ is strictly convex. The following general duality theorem gives the form of the dual problem equivalent to (12.14) in terms of the conjugate function Ψ^* of Ψ. Here, $\|\cdot\|$ is an arbitrary norm over $\mathbb{R}^N$ and $\|\cdot\|_*$ its conjugate. We will assume here that $\sup_x \|\mathbf{\Phi}(x)\| \leq r$.

Theorem 12.4 *Let Ψ be a convex function defined over $\mathbb{R}^{\mathcal{X}}$. Then, problem* (12.14) *admits the following equivalent dual:*

$$\begin{aligned} &\min_{\mathsf{p}\in\Delta} \mathsf{B}_\Psi(\mathsf{p}\,\|\,\mathsf{p}_0) \\ &\qquad \textit{subject to: } \Big\| \mathop{\mathbb{E}}_{x\sim\mathsf{p}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] \Big\| \leq \lambda \\ &= \sup_{\mathbf{w}\in\mathbb{R}^N} -\Psi^*\big(\mathbf{w}\cdot\mathbf{\Phi} + \nabla\Psi(\mathsf{p}_0)\big) + \mathbf{w}\cdot \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] - \lambda\|\mathbf{w}\|_* - C(\mathsf{p}_0), \end{aligned}$$

where $C(\mathsf{p}_0) = \Psi(\mathsf{p}_0) - \langle\nabla\Psi(\mathsf{p}_0), \mathsf{p}_0\rangle$.

Proof: The proof is similar to that of Theorem 12.2 and follows by application of the Fenchel duality theorem (Theorem B.39) to the following optimization problem:

$$\min_{\mathsf{p}} f(\mathsf{p}) + g(A\mathsf{p}), \tag{12.15}$$

with the functions f, g, and A defined for all $\mathsf{p}\in\mathbb{R}^{\mathcal{X}}$ and $\mathbf{u}\in\mathbb{R}^N$ by $f(p) = \mathsf{B}_\Psi(\mathsf{p}\,\|\,\mathsf{p}_0) + I_\Delta(p)$, $g(\mathbf{u}) = I_{\mathcal{C}}(\mathbf{u})$, and $A\mathsf{p} = \sum_{x\in\mathcal{X}} \mathsf{p}(x)\mathbf{\Phi}(x)$. Given these definitions, problem (12.15) is equivalent to (12.14). A is a bounded linear map since for any p, we have $\|A\mathsf{p}\| \leq \|\mathsf{p}\|_1 \sup_x \|\mathbf{\Phi}(x)\| \leq r\|\mathsf{p}\|_1$. Also, notice that for all $\mathbf{w}\in\mathbb{R}^N$, $A^*\mathbf{w} = \mathbf{w}\cdot\mathbf{\Phi}$.

Consider $\mathbf{u}_0\in\mathbb{R}^N$ defined by $\mathbf{u}_0 = \mathbb{E}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] = A\widehat{\mathcal{D}}$. Since $\widehat{\mathcal{D}}$ is in $\Delta = \mathrm{dom}(f)$, $\mathbf{u}_0$ is in $A(\mathrm{dom}(f))$. Furthermore, since $\lambda > 0$, $\mathbf{u}_0$ is in $\mathrm{int}(\mathcal{C})$. $g = I_{\mathcal{C}}$ equals zero over $\mathrm{int}(\mathcal{C})$ and is therefore continuous over $\mathrm{int}(\mathcal{C})$, thus g is continuous at $\mathbf{u}_0$ and we have $\mathbf{u}_0 \in A(\mathrm{dom}(f)) \cap \mathrm{cont}(g)$. Thus, the assumptions of Theorem B.39 hold.

The conjugate function of f is defined for all $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$ by

$$\begin{aligned}
f^*(\mathsf{q}) &= \sup_{\mathsf{p}} \langle \mathsf{p}, \mathsf{q} \rangle - \mathsf{B}_{\Psi}(\mathsf{p} \,\|\, \mathsf{p}_0) - I_{\Delta}(\mathsf{p}) \\
&= \sup_{\mathsf{p} \in \Delta} \langle \mathsf{p}, \mathsf{q} \rangle - \mathsf{B}_{\Psi}(\mathsf{p} \,\|\, \mathsf{p}_0) \\
&= \sup_{\mathsf{p} \in \Delta} \langle \mathsf{p}, \mathsf{q} \rangle - \Psi(\mathsf{p}) + \Psi(\mathsf{p}_0) + \langle \nabla \Psi(\mathsf{p}_0), \mathsf{p} - \mathsf{p}_0 \rangle \\
&= \sup_{\mathsf{p} \in \Delta} \langle \mathsf{p}, \mathsf{q} + \nabla \Psi(\mathsf{p}_0) \rangle - \Psi(\mathsf{p}) + \Psi(\mathsf{p}_0) - \langle \nabla \Psi(\mathsf{p}_0), \mathsf{p}_0 \rangle \\
&= \Psi^*(\mathsf{q} + \nabla \Psi(\mathsf{p}_0)) + \Psi(\mathsf{p}_0) - \langle \nabla \Psi(\mathsf{p}_0), \mathsf{p}_0 \rangle.
\end{aligned}$$

The conjugate function of $g = I_{\mathcal{C}}$ is defined for all $\mathbf{w} \in \mathbb{R}^N$ by

$$\begin{aligned}
g^*(\mathbf{w}) &= \sup_{\mathbf{u}} \langle \mathbf{w}, \mathbf{u} \rangle - I_{\mathcal{C}}(\mathbf{u}) \\
&= \sup_{\mathbf{u} \in \mathcal{C}} \langle \mathbf{w}, \mathbf{u} \rangle \\
&= \sup_{\|\mathbf{u} - \mathbb{E}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}]\| \leq \lambda} \langle \mathbf{w}, \mathbf{u} \rangle \\
&= \langle \mathbf{w}, \mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] \rangle + \sup_{\|\mathbf{u}\| \leq \lambda} \langle \mathbf{w}, \mathbf{u} \rangle = \langle \mathbf{w}, \mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] \rangle + \lambda \|\mathbf{w}\|_*,
\end{aligned}$$

where the last equality holds by definition of the dual norm. In view of these identities, by Theorem B.39, we have

$$\begin{aligned}
\min_{\mathsf{p}} f(\mathsf{p}) + g(A\mathsf{p}) &= \sup_{\mathbf{w} \in \mathbb{R}^N} -f^*(A^*\mathbf{w}) - g^*(\mathbf{w}) \\
&= \sup_{\mathbf{w} \in \mathbb{R}^N} -\Psi^*(\mathbf{w} \cdot \mathbf{\Phi} + \nabla \Psi(\mathsf{p}_0)) + \mathbf{w} \cdot \mathop{\mathbb{E}}_{\widehat{\mathcal{D}}}[\mathbf{\Phi}] - \lambda \|\mathbf{w}\|_* \\
&\qquad - \Psi(\mathsf{p}_0) + \langle \nabla \Psi(\mathsf{p}_0), \mathsf{p}_0 \rangle,
\end{aligned}$$

which completes the proof. □

Note, the previous proof and its use of Fenchel duality holds even when considering norms that are not inner product norms and, more generally, Banach spaces are considered (as mentioned in section B.4).

Much of the analysis and theoretical guarantees presented in previous sections in the special case of the unnormalized relative entropy straightforwardly extend to a broad family of Bregman divergences.

12.9 L_2-regularization

In this section, we study a common variant of the Maxent algorithm where a regularization based on the norm-2 squared of the weight vector $\mathbf{w}$ is used. Observe that this is not covered by the general framework discussed in the previous section where the regularization was based on some norm of $\mathbf{w}$. The corresponding (dual) optimization problem is the following:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|_2^2 - \frac{1}{m}\sum_{i=1}^{m} \log \mathsf{p}_{\mathbf{w}}[x_i]. \tag{12.16}$$

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

Theorem 12.5 *Let $\widehat{\mathbf{w}}$ be a solution of the optimization problem* (12.16)*. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m from $\mathcal{D}$, the following inequality holds:*

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \leq \inf_{\mathbf{w}} \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2 + \frac{r^2}{\lambda m}\left(1 + \sqrt{\log\frac{1}{\delta}}\right)^2.$$

Proof: Let $\widehat{D}$ denote the empirical distribution defined by the sample S. Then, the optimization problem (12.16) can be formulated as follows:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|_2^2 - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}\big[\log \mathsf{p}_{\mathbf{w}}[x]\big] = \lambda\|\mathbf{w}\|_2^2 - \mathbf{w}\cdot \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] + \log Z(\mathbf{w}),$$

where $Z(\mathbf{w}) = \big(\sum_x \exp(\mathbf{w}\cdot\mathbf{\Phi}(x))\big)$. Similarly, let $\mathbf{w}_{\mathcal{D}}$ denote the solution of the minimization problem with the distribution $\mathcal{D}$:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|_2^2 - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\big[\log \mathsf{p}_{\mathbf{w}}[x]\big] = \lambda\|\mathbf{w}\|_2^2 - \mathbf{w}\cdot \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)] + \log Z(\mathbf{w}).$$

We first give an upper-bound on $\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}})$ valid for all $\mathbf{w}\in\mathbb{R}^N$, starting with a decomposition of $\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}})$ as a sum of terms, next using the expression of $\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}})$ in terms of average feature values, then the optimality of $\widehat{\mathbf{w}}$, next the expression of $\mathcal{L}_S(\mathbf{w}_{\mathcal{D}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}})$ in terms of average feature values, and finally the Cauchy-Schwarz inequality and the optimality of $\mathbf{w}_{\mathcal{D}}$:

$$\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \\
&= \mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) - \mathcal{L}_S(\widehat{\mathbf{w}}) + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \lambda\|\widehat{\mathbf{w}}\|_2^2 - \lambda\|\widehat{\mathbf{w}}\|_2^2 \\
&= \widehat{\mathbf{w}}\cdot\Big[\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big] + \mathcal{L}_S(\widehat{\mathbf{w}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \lambda\|\widehat{\mathbf{w}}\|_2^2 - \lambda\|\widehat{\mathbf{w}}\|_2^2 \\
&\leq \widehat{\mathbf{w}}\cdot\Big[\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big] + \mathcal{L}_S(\mathbf{w}_{\mathcal{D}}) - \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \lambda\|\mathbf{w}_{\mathcal{D}}\|_2^2 - \lambda\|\widehat{\mathbf{w}}\|_2^2 \\
&\leq [\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}]\cdot\Big[\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big] + \mathcal{L}_{\mathcal{D}}(\mathbf{w}_{\mathcal{D}}) + \lambda\|\mathbf{w}_{\mathcal{D}}\|_2^2 - \lambda\|\widehat{\mathbf{w}}\|_2^2 \\
&\leq \|\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}\|_2 \Big\|\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]\Big\|_2 + \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2.
\end{aligned}$$

Next, we bound $\|\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}\|_2$ using the fact that $\widehat{\mathbf{w}}$ and $\mathbf{w}_{\mathcal{D}}$ are solutions of the minimization of convex and differentiable objectives functions, whose gradients must be zero at the minimizing values:

$$\begin{aligned} 2\lambda\widehat{\mathbf{w}} - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] + \nabla \log Z(\widehat{\mathbf{w}}) = 0 \\ 2\lambda\mathbf{w}_{\mathcal{D}} - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)] + \nabla \log Z(\mathbf{w}_{\mathcal{D}}) = 0, \end{aligned}$$

which implies

$$2\lambda(\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}) = \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)] + \nabla \log Z(\mathbf{w}_{\mathcal{D}}) - \nabla \log Z(\widehat{\mathbf{w}}).$$

Multiplying both sides by $(\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}})$ gives

$$\begin{aligned} & 2\lambda\|\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}\|_2^2 \\ & = \Big[\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big] \cdot [\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}] - [\nabla \log Z(\widehat{\mathbf{w}}) - \nabla \log Z(\mathbf{w}_{\mathcal{D}})] \cdot [\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}] \\ & \leq \Big[\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big] \cdot [\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}], \end{aligned}$$

in view of the convexity of $\mathbf{w} \mapsto \log Z(\mathbf{w})$. Using the Cauchy-Schwarz inequality and simplifying, we obtain

$$\|\widehat{\mathbf{w}} - \mathbf{w}_{\mathcal{D}}\|_2 \leq \frac{\Big\|\mathbb{E}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathbb{E}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big\|_2}{2\lambda}.$$

Plugging this back in the upper bound previously derived for $\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}})$ yields

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \leq \frac{\Big\|\mathbb{E}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathbb{E}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big\|_2^2}{2\lambda} + \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2.$$

We now use McDiarmid's inequality to bound $\Big\|\mathbb{E}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathbb{E}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big\|_2$. Let $\Psi(S)$ denote this quantity for a sample S. Let S' be a sample differing from S by one point, say x_m for S, x'_m for S'. Then, by the triangle inequality,

$$\begin{aligned} |\Psi(S') - \Psi(S)| &= \bigg|\Big\|\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}'}[\boldsymbol{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big\|_2 - \Big\|\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\boldsymbol{\Phi}(x)]\Big\|_2\bigg| \\ &\leq \Big\|\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}'}[\boldsymbol{\Phi}(x)] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}}[\boldsymbol{\Phi}(x)]\Big\|_2 \\ &\leq \Big\|\frac{\boldsymbol{\Phi}(x'_m) - \boldsymbol{\Phi}(x_m)}{m}\Big\|_2 \leq \frac{2r}{m}. \end{aligned}$$

Thus, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds

$$\Psi(S) \leq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\Psi(S)] + 2r\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

For any $i \in [m]$, let $\mathbf{Z}_i$ denote the random variable $\mathbb{E}_{x\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x_i)] - \mathbb{E}_{x\sim\mathcal{D}}[\mathbf{\Phi}(x)]$. Then, by Jensen's inequality, $\mathbb{E}_{S\sim\mathcal{D}^m}[\Psi(S)]$ can be upper-bounded as follows:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\Psi(S)] = \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{Z}_i\right\|_2\right] \le \sqrt{\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{Z}_i\right\|_2^2\right]}.$$

Since the random variables $\mathbf{Z}_i$s are i.i.d. and centered ($\mathbb{E}[\mathbf{Z}_i] = 0$), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{Z}_i\right\|^2\right] &= \frac{1}{m^2}\left[\sum_{i=1}^{m}\mathbb{E}\left[\|\mathbf{Z}_i\|^2\right] + \sum_{i\neq j}\mathbb{E}[\mathbf{Z}_i]\cdot\mathbb{E}[\mathbf{Z}_j]\right]\\ &= \frac{\mathbb{E}\left[\|\mathbf{Z}_1\|^2\right]}{m}\\ &= \frac{\mathbb{E}\left[\|\mathbf{Z}_1\|^2 + \|\mathbf{Z}_2\|^2\right]}{2m}\\ &= \frac{\mathbb{E}\left[\|\mathbf{Z}_1 - \mathbf{Z}_2\|^2\right]}{2m}.\end{aligned}$$

where, for the last equality, we used the fact that $\mathbb{E}[\mathbf{Z}_1\cdot\mathbf{Z}_2] = \mathbb{E}[\mathbf{Z}_1]\cdot\mathbb{E}[\mathbf{Z}_2] = 0$. This shows that $\mathbb{E}[\Psi(S)] \le \frac{2r}{\sqrt{2m}}$ and that, with probability at least $1-\delta$, the following holds

$$\Psi(S) \le \frac{2r}{\sqrt{2m}}\left(1 + \sqrt{\log\frac{1}{\delta}}\right),$$

and therefore also

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) &\le \frac{1}{2\lambda}\frac{2r^2}{m}\left(1 + \sqrt{\log\frac{1}{\delta}}\right)^2 + \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2\\ &\le \frac{r^2}{\lambda m}\left(1 + \sqrt{\log\frac{1}{\delta}}\right)^2 + \mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2,\end{aligned}$$

which ends the proof. □

Assume that $\mathbf{w}^*$ achieves the infimum of the loss, that is $\mathcal{L}_{\mathcal{D}}(\mathbf{w}^*) = \inf_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w})$ and that we are given an upper bound Λ_2 on its norm: $\|\mathbf{w}^*\|_2 \le \Lambda_2$. Then, we can use that upper bound and choose λ to minimize the two terms containing λ: $\lambda\Lambda_2^2 = \frac{r^2}{\lambda m}$, that is $\lambda = \frac{r}{\Lambda_2\sqrt{m}}$ and the theorem would then guarantee the following inequality with probability $1-\delta$ for $\widehat{\mathbf{w}}$:

$$\mathcal{L}_{\mathcal{D}}(\widehat{\mathbf{w}}) \le \inf_{\mathbf{w}}\mathcal{L}_{\mathcal{D}}(\mathbf{w}) + \frac{r\Lambda_2}{\sqrt{m}}\left[1 + \left(1 + \sqrt{\log\frac{1}{\delta}}\right)^2\right].$$

12.10 Chapter notes

The Maxent principle was first explicitly advocated by Jaynes [1957] (see also Jaynes [1983]) who referred to Shannon's notion of entropy (appendix E) to support this principle. As seen in Section 12.5, standard Maxent models coincide with Gibbs distributions, as in the original Boltzmann models in statistical mechanics. In fact, Jaynes [1957] argued that statistical mechanics could be viewed as a form of statistical inference, as opposed to a physical theory, and that the thermodynamic notion of entropy could be replaced by the information-theoretical notion. The justification of the Maxent principle presented in this chapter is instead based upon learning theory arguments.

Maximum entropy models, commonly referred to as Maxent models, are used in a variety of tasks in natural language processing [Berger et al., 1996, Rosenfeld, 1996, Pietra et al., 1997, Malouf, 2002, Manning and 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]. One key benefit of Maxent models is that they allow the use of diverse features that can be selected and augmented by the user. 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 and Tsujii, 2003] or the L_2-norm [Chen and Rosenfeld, 2000, Lebanon and 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 vectors in a Bayesian interpretation [Williams, 1994, Goodman, 2004], thereby making Maxent models coincide with Maximum a Posteriori solutions with specific choices of the prior.

An extensive theoretical study of these regularizations and the introduction of other more general ones were given by Dudík, Phillips, and Schapire [2007] and by Altun and Smola [2006] who studied the extensions to arbitrary Bregman divergences and norms (Section 12.8) using Fenchel duality (see also [Lafferty, Pietra, and Pietra, 1997]). Cortes, Kuznetsov, Mohri, and Syed [2015] give a more general family of density estimation models, *Structural Maxent models*, with feature functions selected from a union of possibly very complex sub-families for which they also give a duality theorem, strong learning guarantees, and algorithms. These models can also be viewed as Maxent with a more general type of regularization.

The Maxent duality theorem is due to Pietra, Pietra, and Lafferty [1997] (see also [Dudík et al., 2007] and [Altun and Smola, 2006]). Theorem 12.2 is a slight extension giving a guarantee for an ϵ-solution of the dual and is a special instance of a more general theorem given for Structural Maxent models [Cortes et al., 2015]. The generalization bounds of Sections 12.6 and 12.9 and their proofs are variants of results due to Dudík et al. [2007]. The stability analysis used in the proof of Theorem 12.5 is equivalent to the one described in Chapter 14 using Bregman divergences.

A variety of different techniques have been suggested to solve the Maxent optimization problem including standard gradient descent and stochastic gradient descent. Some specific algorithms were introduced for this problem, including *generalized iterative scaling* (GIS) [Darroch and Ratcliff, 1972] and *improved iterative scaling* (IIS) [Pietra et al., 1997]. It was shown by Malouf [2002] that these algorithms perform poorly in several natural language processing tasks in comparison with conjugate gradient techniques and limited-memory BFGS methods (see also [Andrew and Gao, 2007]). The coordinate descent solution presented in this chapter is due to Cortes et al. [2015]. It is a simpler version of an algorithm of Dudík et al. [2007] which uses a tighter upper bound on $J(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_j)$ but which is subject to various technical conditions. Both algorithms benefit from a similar asymptotic convergence rate [Cortes et al., 2015] and are particularly adapted to cases where the number of features is very large and where updating all feature weights is impractical. A sequential greedy approximation due to Zhang [2003b] is also advocated by Altun and Smola [2006] as a general algorithm for general forms of the Maxent problem.

12.11 Exercises

12.1 Convexity. Prove directly that the function $\mathbf{w} \mapsto \log Z(\mathbf{w}) = \log(\sum_{x \in \mathcal{X}} e^{\mathbf{w} \cdot \mathbf{\Phi}(x)})$ is convex (*Hint*: compute its Hessian).

12.2 Lagrange duality. Derive the dual problem of the Maxent problem and justify it carefully in the case of the stricter constraint of positivity for the distribution p: $\mathsf{p}(x) > 0$ for all $x \in \mathcal{X}$.

12.3 Dual of norm-2 squared regularized Maxent. Derive the dual formulation of the norm-2 squared regularized Maxent optimization shown in equation (12.16).

12.4 Extension to Bregman divergences. Derive theoretical guarantees for the extensions discussed in Section 12.8. What additional property is needed for the Bregman divergence so that your learning guarantees hold?

12.5 L_2-regularization. Let $\mathbf{w}$ be the solution of Maxent with a norm-2 squared regularization.

(a) Prove the following inequality: $\|\mathbf{w}\|_2 \leq \frac{2r}{\lambda}$ (*Hint*: you could compare the values of the objective function at $\mathbf{w}$ and 0.). Generalize this result to other $\|\cdot\|_p^p$-regularizations with $p > 1$.

(b) Use the previous question to derive an explicit learning guarantee for Maxent with norm-2 squared regularization (*Hint*: you could use the last inequality given in Section 12.9 and derive an explicit expression for Λ_2).