13 Conditional Maximum Entropy Models

This chapter presents algorithms for estimating the conditional probability of a class given an example, rather than only predicting the class label for that example. This is motivated by several applications where confidence values are sought, in addition to the class prediction. The algorithms discussed, *conditional Maxent models*, also known as *multinomial logistic regression* algorithms, are among the most well-known and most widely used multi-class classification algorithms. In the special case of two classes, the algorithm is known as *logistic regression*.

As suggested by their name, these algorithms can be viewed as Maxent models for conditional probabilities. To introduce them, we will extend the ideas discussed in the previous chapter (Chapter 12), starting from an extension of the Maxent principle to the conditional case. Next, we will prove a duality theorem leading to an equivalent dual optimization problem for conditional Maxent. We will specifically discuss different aspects of multi-class classification using conditional Maxent and reserve a special section to the analysis of logistic regression.

13.1 Learning problem

We consider a multi-class classification problem with c classes, $c \geq 1$. Let $\mathcal{Y} = \{1, \ldots, c\}$ denote the output space and $\mathcal{D}$ a distribution over $\mathcal{X} \times \mathcal{Y}$. The learner receives a labeled training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ drawn i.i.d. according to $\mathcal{D}$. As in Chapter 12, we assume that, additionally, the learner has access to a feature mapping $\mathbf{\Phi} \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^N$ with $\mathbb{R}^N$ a normed vector space and with $\|\mathbf{\Phi}\|_\infty \leq r$. We will denote by $\mathcal{H}$ a family of real-valued functions containing the component feature functions Φ_j with $j \in [N]$. Note that in the most general case, we may have $N = +\infty$. The problem consists of using the training sample S to learn an accurate conditional probability $\mathsf{p}[\cdot|x]$, for any $x \in \mathcal{X}$.

13.2 Conditional Maxent principle

As for Maxent models, conditional Maxent or logistic regression models can be derived from a key concentration inequality. By the general Rademacher complexity bound (Theorem 3.3), for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ over the choice of a sample of size m:

$$\left\| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbf{\Phi}(x,y)] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x,y)] \right\|_\infty \leq 2\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}, \tag{13.1}$$

where we denote by $\widehat{\mathcal{D}}$ the empirical distribution defined by the sample S. We will also denote by $\widehat{\mathcal{D}}^1(x)$ the empirical distribution of x in the sample S. For any $x \in \mathcal{X}$, let $\mathsf{p}_0[\cdot|x]$ denote a conditional probability, often chosen to be the uniform distribution. Then, the *conditional Maxent principle* consists of seeking conditional probabilities $\mathsf{p}[\cdot|x]$ that are as agnostic as possible, that is as close as possible to the uniform distribution, or, more generally, to priors $\mathsf{p}_0[\cdot|x]$, while verifying an inequality similar to (13.1):

$$\left\| \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1 \\ y\sim\mathsf{p}[\cdot|x]}}[\mathbf{\Phi}(x,y)] - \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x,y)] \right\|_\infty \leq \lambda, \tag{13.2}$$

where $\lambda \geq 0$ is a parameter. Here, closeness is defined via the *conditional relative entropy* (appendix E) based on the empirical marginal distribution $\widehat{\mathcal{D}}^1$ of input points. Choosing $\lambda = 0$ corresponds to standard *conditional Maxent* or *unregularized conditional Maxent* and to requiring the expectation of the features based on $\mathcal{D}^1$ and the conditional probabilities $\mathsf{p}[\cdot|x]$ 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 the conditional Maxent principle does not require specifying a family of conditional probability distributions $\mathcal{P}$ to choose from.

13.3 Conditional Maxent models

Let Δ denote the simplex of the probability distributions over $\mathcal{Y}$, $\mathcal{X}_1 = \text{supp}(\widehat{\mathcal{D}}^1)$ the support of $\widehat{\mathcal{D}}^1$, and $\bar{\mathsf{p}} \in \Delta^{\mathcal{X}_1}$ the family of conditional probabilities, $\bar{\mathsf{p}} = (\mathsf{p}[\cdot|x])_{x\in\mathcal{X}_1}$. Then, the conditional Maxent principle can be formulated as the following optimization problem:

$$\min_{\bar{\mathsf{p}} \in \Delta^{\mathcal{X}_1}} \sum_{x \in \mathcal{X}_1} \widehat{\mathcal{D}}^1(x)\, \mathsf{D}\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) \tag{13.3}$$

$$\text{s.t.} \left\| \mathop{\mathbb{E}}_{\substack{x \sim \widehat{\mathcal{D}}^1 \\ y \sim \mathsf{p}[\cdot|x]}} [\mathbf{\Phi}(x,y)] - \mathop{\mathbb{E}}_{(x,y) \sim \widehat{\mathcal{D}}} [\mathbf{\Phi}(x,y)] \right\|_\infty \leq \lambda.$$

This defines a convex optimization problem since the objective is a positive sum of relative entropies and since the relative entropy D is convex with respect to its arguments (appendix E), since the constraints are affine functions of $\bar{\mathsf{p}}$, and since $\Delta^{\mathcal{X}_1}$ is a convex set. The solution is in fact unique, since the objective is strictly convex as a positive sum of relative entropies, each strictly convex. The empirical conditional probabilities $\widehat{\mathcal{D}}^1(\cdot|x)$, $x \in \mathcal{X}_1$, clearly form a feasible solution, thus problem (12.7) is feasible.

For uniform priors $\mathsf{p}_0[\cdot|x]$, problem (13.3) can be equivalently formulated as a conditional entropy maximization, which explains the name given to these models. Let $\bar{\mathsf{H}}(\bar{\mathsf{p}}) = -\mathbb{E}_{x \sim \widehat{\mathcal{D}}^1}\big[\sum_{y \in \mathcal{Y}} \mathsf{p}[y|x] \log \mathsf{p}[y|x]\big]$ denote the conditional entropy of p with respect to the marginal $\widehat{\mathcal{D}}^1$. Then, the objective function of (12.7) can be rewritten as follows:

$$\begin{aligned}
\mathsf{D}\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) &= \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}^1} \left[\sum_{y \in \mathcal{Y}} \mathsf{p}[y|x] \log \frac{\mathsf{p}[y|x]}{\mathsf{p}_0[y|x]} \right] \\
&= \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}^1} \left[-\sum_{y \in \mathcal{Y}} \mathsf{p}[y|x] \log(1/c) + \sum_{y \in \mathcal{Y}} \mathsf{p}[y|x] \log \mathsf{p}[y|x] \right] \\
&= \log(c) - \bar{\mathsf{H}}(\bar{\mathsf{p}}).
\end{aligned}$$

Thus, since $\log(c)$ is a constant, minimizing the objective is then equivalent to maximizing $\bar{\mathsf{H}}(\bar{\mathsf{p}})$.

Conditional Maxent models are the solutions of the optimization problem just described. As in the non-conditional case, 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 conditional Maxent models.

13.4 Dual problem

Here, we derive an equivalent dual problem for (13.3) which, as we will show, can be formulated as a regularized conditional maximum likelihood problem over the family of *Gibbs distributions*.

The Maxent optimization problem (13.3) can be equivalently expressed as the unconstrained optimization problem $\min_{\bar{\mathsf{p}}} F(\bar{\mathsf{p}})$ with, for all $\bar{\mathsf{p}} = (\mathsf{p}[\cdot|x] \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}$,

$$F(\bar{\mathsf{p}}) = \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}^1} \Big[\widetilde{\mathsf{D}}\big(\mathsf{p}[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]\big) \Big] + I_{\mathcal{C}} \Bigg(\mathop{\mathbb{E}}_{\substack{x \sim \widehat{\mathcal{D}}^1 \\ y \sim \mathsf{p}[\cdot|x]}} [\mathbf{\Phi}(x, y)] \Bigg), \tag{13.4}$$

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

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

$$G(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m \log \left[\frac{\mathsf{p}_{\mathbf{w}}[y_i|x_i]}{\mathsf{p}_0[y_i|x_i]} \right] - \lambda \|\mathbf{w}\|_1 \,, \tag{13.5}$$

with, for all $x \in \mathcal{X}_1$ and $y \in \mathcal{Y}$,

$$\mathsf{p}_{\mathbf{w}}[y|x] = \frac{\mathsf{p}_0[y|x] e^{\mathbf{w} \cdot \mathbf{\Phi}(x,y)}}{Z(\mathbf{w}, x)} \quad \text{and} \quad Z(\mathbf{w}, x) = \sum_{y \in \mathcal{Y}} \mathsf{p}_0[y|x] e^{\mathbf{w} \cdot \mathbf{\Phi}(x,y)}. \tag{13.6}$$

Then, the following theorem gives a result similar to the duality theorem presented in the non-conditional case (Theorem 12.2, Section 12.5).

Theorem 13.1 *Problem* (13.3) *is equivalent to the dual optimization problem* $\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})$*:*

$$\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\bar{\mathsf{p}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}} F(\bar{\mathsf{p}}). \tag{13.7}$$

Furthermore, let $\bar{\mathsf{p}}^* = \operatorname{argmin}_{\bar{\mathsf{p}}} F(\bar{\mathsf{p}})$*. Then, for any* $\epsilon > 0$ *and any* $\mathbf{w}$ *such that* $|G(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})| < \epsilon$*, we have* $\mathbb{E}_{x \sim \widehat{\mathcal{D}}^1} \big[\mathsf{D}\big(\bar{\mathsf{p}}^*[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]\big)\big] \leq \epsilon$*.*

The proof is similar to that of Theorem 12.2 and is given at the end of this chapter since it is somewhat longer (Section 13.9).

In view of the theorem, if $\mathbf{w}$ is an ϵ-solution of the dual optimization problem, then $\mathbb{E}_{x \sim \widehat{\mathcal{D}}^1} \big[\mathsf{D}\big(\mathsf{p}^*[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]\big)\big] \leq \epsilon$, which, by Jensen's inequality and Pinsker's inequality (Proposition E.7) implies that

$$\mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}^1} \Big[\big\| \mathsf{p}^*[\cdot|x] - \mathsf{p}_{\mathbf{w}}[\cdot|x] \big\|_1 \Big] \leq \sqrt{ \mathop{\mathbb{E}}_{x \sim \widehat{\mathcal{D}}^1} \Big[\big\| \mathsf{p}^*[\cdot|x] - \mathsf{p}_{\mathbf{w}}[\cdot|x] \big\|_1^2 \Big] } \leq \sqrt{2\epsilon}.$$

Thus, $\mathsf{p}_{\mathbf{w}}[\cdot|x]$ is then $\sqrt{2\epsilon}$-close in $\widehat{\mathcal{D}}^1$-averaged L_1-norm to the optimal solution of the primal and the theorem suggests that the solution of the conditional Maxent problem can be determined by solving the dual problem, which can be written equivalently as follows for a uniform prior:

$$\inf_{\mathbf{w}} \lambda \|\mathbf{w}\|_1 - \frac{1}{m} \sum_{i=1}^m \log \big[\mathsf{p}_{\mathbf{w}}[y_i|x_i]\big]. \tag{13.8}$$

Similar remarks to those made for non-conditional Maxent models apply here. In particular, the solution may not be achieved for any finite $\mathbf{w}$ for $\lambda = 0$, which is why the infimum is needed. Also, this result may seem surprising since it shows that conditional Maxent coincides with conditional Maximum Likelihood ($\lambda = 0$) or regularized conditional Maximum Likelihood ($\lambda > 0$) using for the family $\mathcal{P}$ of conditional probabilities to choose from that of Gibbs distributions, while the conditional Maxent principle does not explicitly specify any family of conditional probabilities $\mathcal{P}$. The reason is the specific choice of the conditional relative entropy as the measure of closeness of $\mathsf{p}[\cdot|x]$ to the prior conditional distributions $\mathsf{p}_0[\cdot|x]$. 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 conditional distributions in maximum likelihood. Also, as already mentioned in the standard Maxent case, Gibbs distributions form a very rich family.

Notice that both the primal and the dual optimization problems for conditional Maxent involve only conditional probabilities $\mathsf{p}[\cdot|x]$ for x in $\mathcal{X}_1$, that is for x in the training sample. Thus, they do not provide us with any information about other conditional probabilities. However, the dual shows that, for x in $\mathcal{X}_1$, the solution admits the same general form $\mathsf{p}_{\mathbf{w}}[\cdot|x]$, which only depends on the weight vector $\mathbf{w}$. In view of that, we extend the definition of Maxent conditional probabilities to all $x \in \mathcal{X}$ by using the same general form $\mathsf{p}_{\mathbf{w}}[\cdot|x]$ and the same vector $\mathbf{w}$ for all x.

Observe also that in the definition of the primal or dual problems we could have used some other distribution $\mathcal{Q}$ over $\mathcal{X}$ in lieu of $\widehat{\mathcal{D}}^1$. It is straightforward to verify that the duality theorem would continue to hold in that case using the same proof. In fact, ideally, we would have chosen $\mathcal{Q}$ to be $\mathcal{D}^1$. However, that optimization problem would require knowledge of the feature vectors for all $x \in \operatorname{supp}(\mathcal{D}^1)$, which of course is not accessible to us given a finite sample. The weighted vector $\mathbf{w}$ found when using $\widehat{\mathcal{D}}^1$ can be viewed as an approximation of the one obtained if using $\mathcal{D}^1$.

13.5 Properties

In this section, we discuss several aspects of conditional Maxent models, including the form of the dual optimization problems, the feature vectors used, and prediction with these models.

13.5.1 Optimization problem

L_1-regularized conditional Maxent models are therefore conditional probability models solutions of the primal problem (13.3) or, equivalently, models defined by

$$\mathsf{p}_{\mathbf{w}}[y|x] = \frac{e^{\mathbf{w}\cdot\mathbf{\Phi}(x,y)}}{Z(x)} \quad \text{and} \quad Z(x) = \sum_{y\in\mathcal{Y}} e^{\mathbf{w}\cdot\mathbf{\Phi}(x,y)}, \tag{13.9}$$

where $\mathbf{w}$ is solution of the dual problem

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

with $\lambda \geq 0$ is a parameter. Using the expression of the conditional probabilities, this optimization problem can be written more explicitly as

$$\min_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|_1 + \frac{1}{m}\sum_{i=1}^{m} \log\left[\sum_{y\in\mathcal{Y}} e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,y)-\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)}\right]. \tag{13.10}$$

or, equivalently, as

$$\min_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|_1 - \mathbf{w}\cdot\frac{1}{m}\sum_{i=1}^{m}\mathbf{\Phi}(x_i,y_i) + \frac{1}{m}\sum_{i=1}^{m}\log\left[\sum_{y\in\mathcal{Y}} e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,y)}\right]. \tag{13.11}$$

By definition of the dual problem, this is an unconstrained convex optimization problem in $\mathbf{w}$. This can be also seen from the fact that the log-sum function $\mathbf{w} \mapsto \log\left[\sum_{y\in\mathcal{Y}} e^{\mathbf{w}\cdot\mathbf{\Phi}(x,y)}\right]$ is convex for any $x \in \mathcal{X}$.

There are many optimization solutions available for this problem, including several special-purpose algorithms, general first-order and second-order solutions, and special-purpose distributed solutions. One common method is simply to use stochastic gradient descent (SGD), which has been reported to be more efficient than most special-purpose methods in applications. When the dimension of the feature vectors $\mathbf{\Phi}$ (or the cardinality of the family of feature functions $\mathcal{H}$) is very large, these methods are typically inefficient. An alternative method then consists of applying coordinate descent to solve this problem. In that case, the resulting algorithm coincides with the version of L_1-regularized boosting where, instead of the exponential function, the logistic function is used.

13.5.2 Feature vectors

Using feature vectors $\mathbf{\Phi}(x,y)$ depending on both the input x and the output y is often important in applications. For example, in machine translation, it is convenient to use features whose values may depend on the presence of some words in the input sentence and some others in the output sequence. A common choice of the feature vector is however one where the column vectors $\mathbf{\Phi}(x,y)$ and $\mathbf{w}$ admit c

blocks of equal size and where only the block in $\mathbf{\Phi}(x, y)$ corresponding to the class y is non-zero and equal to a feature vector $\mathbf{\Gamma}(x)$ independent of the class labels:

$$\mathbf{\Phi}(x, y) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{\Gamma}(x) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_{y-1} \\ \mathbf{w}_y \\ \mathbf{w}_{y+1} \\ \vdots \\ \mathbf{w}_c \end{bmatrix}.$$

In view of that, the inner product of $\mathbf{w}$ and $\mathbf{\Phi}(x, y)$ can be expressed in terms of the feature vector $\mathbf{\Gamma}(x)$, which only depends on x, but with a distinct parameter vector $\mathbf{w}_y$:

$$\mathbf{w} \cdot \mathbf{\Phi}(x, y) = \mathbf{w}_y \cdot \mathbf{\Gamma}(x).$$

The optimization problem for L_1-regularized conditional Maxent can then be written in terms of the vectors $\mathbf{w}_y$ as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^N} \lambda \sum_{y \in \mathcal{Y}} \|\mathbf{w}_y\|_1 + \frac{1}{m} \sum_{i=1}^{m} \log \left[\sum_{y \in \mathcal{Y}} e^{\mathbf{w}_y \cdot \mathbf{\Gamma}(x_i) - \mathbf{w}_{y_i} \cdot \mathbf{\Gamma}(x_i)} \right]. \tag{13.12}$$

Notice that, if the vectors $\mathbf{w}_y$ were not correlated via the second term of the objective function (for example if, instead of the log of the sum, this term were replaced by the sum of the logs), then the problem would be reduced to c separate optimization functions learning a distinct weight vector for each class, as in the one-vs-all setup of multi-class classification.

13.5.3 Prediction

Finally, note that the class $\widehat{y}(x)$ predicted by a conditional Maxent model with parameter $\mathbf{w}$ is given by

$$\widehat{y}(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \ \mathsf{p}_{\mathbf{w}}[y|x] = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \ \mathbf{w} \cdot \mathbf{\Phi}(x, y). \tag{13.13}$$

Thus, conditional Maxent models define linear classifiers. Conditional Maxent models are also sometimes referred to as log-*linear models*.

13.6 Generalization bounds

In this section, we will present learning guarantees for conditional Maxent models in two different settings: one where the dimension of the feature vectors $\mathbf{\Phi}$ (or the cardinality of the family of feature functions $\mathcal{H}$) is infinite or extremely large and where a coordinate-descent or boosting-type algorithm is more suitable, and another one where the dimension of the feature vectors $\mathbf{\Phi}$ is finite and not too large.

We start with the case where the dimension of the feature vectors $\mathbf{\Phi}$ is very large. The following margin-based guarantee holds in that case.

Theorem 13.2 *For any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, the following holds for all $\rho > 0$ and $f \in \mathcal{F} = \{(x, y) \mapsto \mathbf{w} \cdot \mathbf{\Phi}(x, y) \colon \|\mathbf{w}\|_1 \leq 1\}$:*

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} \log_{u_0}\Big(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}}\Big) + \frac{8c}{\rho} \mathfrak{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\log \log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}},$$

where $u_0 = \log(1 + 1/e)$ and $\Pi_1(\mathcal{H}) = \{x \mapsto \phi(x, y) \colon \phi \in \mathcal{H}, y \in \mathcal{Y}\}$.

Proof: For any $f \colon (x, y) \mapsto \mathbf{w} \cdot \mathbf{\Phi}(x, y)$ and $i \in [m]$, let $\rho_f(x_i, y_i)$ denote the margin of f at (x_i, y_i):

$$\rho_f(x_i, y_i) = \min_{y \neq y_i} f(x_i, y_i) - f(x_i, y) = \min_{y \neq y_i} \mathbf{w} \cdot (\mathbf{\Phi}(x_i, y_i) - \mathbf{\Phi}(x_i, y)).$$

Fix $\rho > 0$. Then, by Theorem 9.2, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{H}$ and $\rho \in (0, 2r]$:

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{\rho_f(x_i, y_i) \leq \rho} + \frac{4c}{\rho} \mathfrak{R}_m(\Pi_1(\mathcal{F})) + \sqrt{\frac{\log \log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}},$$

where $\Pi_1(\mathcal{F}) = \{x \mapsto f(x, y) \colon y \in \mathcal{Y}, f \in \mathcal{H}\}$. The inequality trivially holds for all $\rho > 0$ since for $\rho \geq 2r$, by Hölder's inequality, we have $|\mathbf{w} \cdot \mathbf{\Phi}(x, y)| \leq \|\mathbf{w}\|_1 \, \|\mathbf{\Phi}(x, y)\|_\infty \leq r$ for $\|\mathbf{w}\|_1 \leq 1$, and thus $\min_{y \neq y_i} f(x_i, y_i) - f(x_i, y) \leq 2r \leq \rho$ for all $i \in [m]$ and $y \in \mathcal{Y}$. Now, for any $\rho > 0$, the ρ-margin loss can be upper bounded by the ρ-logistic loss:

$$\forall u \in \mathbb{R}, 1_{u \leq \rho} = 1_{\frac{u}{\rho} - 1 \leq 0} \leq \log_{u_0}(1 + e^{-\frac{u}{\rho}}).$$

Thus, the ρ-margin loss of f at (x_i, y_i) can be upper bounded as follows:

$$\begin{aligned} 1_{\rho_f(x_i, y_i) \leq \rho} &\leq \log_{u_0}(1 + e^{-\frac{\rho(f, x_i, y_i)}{\rho}}) \\ &= \log_{u_0}(1 + \max_{y \neq y_i} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}}) \\ &\leq \log_{u_0}\left(1 + \sum_{y \neq y_i} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}}\right) = \log_{u_0}\left(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}}\right). \end{aligned}$$

Thus, with probability at least $1 - \delta$, the following inequality holds for all $f \in \mathcal{H}$ and $\rho > 0$:

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} \log_{u_0}\Big(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}}\Big) + \frac{4c}{\rho} \mathfrak{R}_m(\Pi_1(\mathcal{F})) + \sqrt{\frac{\log \log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

For any sample $S = (x_1, \ldots, x_m)$ of size m, the empirical Rademacher complexity of $\Pi_1(\mathcal{F})$ can be bounded as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\Pi_1(\mathcal{F})) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{\|\mathbf{w}\|_1 \leq 1 \\ y \in \mathcal{Y}}} \sum_{i=1}^m \sigma_i \sum_{j=1}^N w_j \Phi_j(x_i, y) \Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{\|\mathbf{w}\|_1 \leq 1 \\ y \in \mathcal{Y}}} \sum_{j=1}^N w_j \sum_{i=1}^m \sigma_i \Phi_j(x_i, y) \Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{j \in [N] \\ y \in \mathcal{Y}}} \Bigg| \sum_{i=1}^m \sigma_i \Phi_j(x_i, y) \Bigg| \Bigg] \\
&\leq \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{\Phi \in \mathcal{H} \\ y \in \mathcal{Y}}} \Bigg| \sum_{i=1}^m \sigma_i \Phi(x_i, y) \Bigg| \Bigg] \leq 2\, \widehat{\mathfrak{R}}_S(\Pi_1(\mathcal{H})),
\end{aligned}$$

which completes the proof. □

The learning guarantee of the theorem is remarkable since it does not depend on the dimension N and since it only depends on the complexity of the family $\mathcal{H}$ of feature functions (or base hypotheses). Since for any $\rho > 0$, f/ρ admits the same generalization error as f, the theorem implies that with probability at least $1 - \delta$, the following inequality holds for all $f \in \{(x, y) \mapsto \mathbf{w} \cdot \boldsymbol{\Phi}(x, y) \colon \|\mathbf{w}\|_1 \leq \frac{1}{\rho}\}$ and $\rho > 0$:

$$R(f) \leq \frac{1}{m} \sum_{i=1}^m \log_{u_0}\Bigg(\sum_{y \in \mathcal{Y}} e^{f(x_i, y) - f(x_i, y_i)} \Bigg) + \frac{8c}{\rho} \mathfrak{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\log \log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

This inequality can be used to derive an algorithm that selects $\mathbf{w}$ and $\rho > 0$ to minimize the right-hand side. The minimization with respect to ρ does not lead to a convex optimization and depends on theoretical constant factors affecting the second and third term. Thus, instead, ρ is left as a free parameter of the algorithm, typically determined via cross-validation.

Now, since only the first term of the right-hand side depends on $\mathbf{w}$, for any $\rho > 0$, the bound suggests selecting $\mathbf{w}$ as the solution of the following optimization problem:

$$\min_{\|\mathbf{w}\|_1 \leq \frac{1}{\rho}} \frac{1}{m} \sum_{i=1}^m \log \Bigg(\sum_{y \in \mathcal{Y}} e^{\mathbf{w} \cdot \boldsymbol{\Phi}(x_i, y) - \mathbf{w} \cdot \boldsymbol{\Phi}(x_i, y_i)} \Bigg). \tag{13.14}$$

Introducing a Lagrange variable $\lambda \geq 0$, the optimization problem can be written equivalently as

$$\min_{\mathbf{w}} \lambda \|\mathbf{w}\|_1 + \frac{1}{m} \sum_{i=1}^{m} \log \bigg(\sum_{y \in \mathcal{Y}} e^{\mathbf{w} \cdot \mathbf{\Phi}(x_i, y) - \mathbf{w} \cdot \mathbf{\Phi}(x_i, y_i)} \bigg). \tag{13.15}$$

Since for any choice of ρ in the constraint of (13.14), there exists an equivalent dual variable λ in the formulation of (13.15) that achieves the same optimal $\mathbf{w}$, λ can be freely selected via cross-validation. The resulting algorithm precisely coincides with conditional Maxent.

When the dimension N of the feature vectors $\mathbf{\Phi}$ is finite, the following margin-based guarantee holds.

Theorem 13.3 *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, the following holds for all $\rho > 0$ and $f \in \mathcal{F} = \{(x, y) \mapsto \mathbf{w} \cdot \mathbf{\Phi}(x, y) \colon \|\mathbf{w}\|_1 \leq 1\}$:*

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} \log_{u_0} \bigg(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}} \bigg) + \frac{4cr\sqrt{2 \log(2cN)}}{\rho} + \sqrt{\frac{\log \log_2 \frac{4r}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}},$$

where $u_0 = \log(1 + 1/e)$.

Proof: The proof coincides with that of Theorem 13.2, modulo the upper bound on $\mathfrak{R}_m(\Pi_1(\mathcal{F}))$. For any sample $S = (x_1, \ldots, x_m)$ of size m, the empirical Rademacher complexity of $\Pi_1(\mathcal{F})$ can be bounded as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\Pi_1(\mathcal{F})) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{\|\mathbf{w}\|_1 \leq 1 \\ y \in \mathcal{Y}}} \sum_{i=1}^{m} \sigma_i \mathbf{w} \cdot \mathbf{\Phi}(x_i, y) \Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{\|\mathbf{w}\|_1 \leq 1 \\ y \in \mathcal{Y}}} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{\Phi}(x_i, y) \Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{y \in \mathcal{Y}} \bigg\| \sum_{i=1}^{m} \sigma_i \mathbf{\Phi}(x_i, y) \bigg\|_\infty \Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \Bigg[\sup_{\substack{j \in [N] \\ y \in \mathcal{Y}, s \in \{-1, +1\}}} s \sum_{i=1}^{m} \sigma_i \Phi_j(x_i, y) \Bigg] \\
&\leq r\sqrt{2 \log(2cN)},
\end{aligned}$$

where the third equality holds by definition of the dual norm, and the last inequality by the maximal inequality (Corollary D.11), since the supremum is taken over $2cN$ choices. □

This learning guarantee of the theorem is very favorable even for relatively high-dimensional problems since its dependency on the dimension N is only logarithmic.

13.7 Logistic regression

The binary case of conditional Maxent models ($c = 2$) is known as *logistic regression* and is one of the most well-known algorithms for binary classification.

13.7.1 Optimization problem

In the binary case, the sum appearing in the optimization problem of conditional Maxent models can be simplified as follows:

$$\begin{aligned}\sum_{y \in \mathcal{Y}} e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,y)-\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)} &= e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,+1)-\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)} + e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,-1)-\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)}\\ &= 1 + e^{-y_i\mathbf{w}\cdot[\mathbf{\Phi}(x_i,+1)-\mathbf{\Phi}(x_i,-1)]}\\ &= 1 + e^{-y_i\mathbf{w}\cdot\mathbf{\Psi}(x_i)},\end{aligned}$$

where for all $x \in \mathcal{X}$, $\mathbf{\Psi}(x) = \mathbf{\Phi}(x,+1) - \mathbf{\Phi}(x,-1)$. This leads to the following optimization problem, which defines *L_1-regularized logistic regression*:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \lambda\|\mathbf{w}\|_1 + \frac{1}{m}\sum_{i=1}^{m}\log\Big[1 + e^{-y_i\mathbf{w}\cdot\mathbf{\Psi}(x_i)}\Big]. \tag{13.16}$$

As discussed in the general case, this is a convex optimization problem which admits a variety of different solutions. A common solution is SGD, another one is coordinate descent. When coordinate descent is used, then the algorithm coincides with the alternative to AdaBoost where the logistic loss is used instead of the exponential loss ($\phi(-u) = \log_2(1 + e^{-u}) \geq 1_{u\leq 0}$).

13.7.2 Logistic model

In the binary case, the conditional probability defined by the weight vector $\mathbf{w}$ can be expressed as follows:

$$\mathsf{p}_{\mathbf{w}}[y = +1 \mid x] = \frac{e^{\mathbf{w}\cdot\mathbf{\Phi}(x,+1)}}{Z(x)}, \tag{13.17}$$

with $Z(x) = e^{\mathbf{w}\cdot\mathbf{\Phi}(x,+1)} + e^{\mathbf{w}\cdot\mathbf{\Phi}(x,-1)}$. Thus, prediction is based on a linear decision rule defined by the sign of log-odds ratio:

$$\log\frac{\mathsf{p}_{\mathbf{w}}[y = +1 \mid x]}{\mathsf{p}_{\mathbf{w}}[y = -1 \mid x]} = \mathbf{w}\cdot\big(\mathbf{\Phi}(x,+1) - \mathbf{\Phi}(x,-1)\big) = \mathbf{w}\cdot\mathbf{\Psi}(x).$$

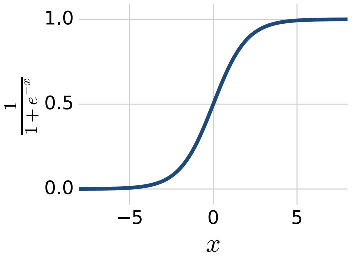

Figure 13.1
Plot of the logistic function f_{logistic}.

This is why logistic regression is also known as a log-*linear model*. Observe also that the conditional probability admits the following *logistic form*:

$$\mathsf{p}_{\mathbf{w}}[y = +1 \mid x] = \frac{1}{1 + e^{-\mathbf{w}\cdot[\mathbf{\Phi}(x,+1) - \mathbf{\Phi}(x,-1)]}} = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{\Psi}(x)}} = f_{\text{logistic}}\big(\mathbf{w} \cdot \mathbf{\Psi}(x)\big),$$

where f_{logistic} is the function defined over $\mathbb{R}$ by $f_{\text{logistic}} \colon x \mapsto \frac{1}{1+e^{-x}}$. Figure 13.1 shows the plot of this function. The logistic function maps the images of the linear function $x \mapsto \mathbf{\Psi}(x)$ to the interval $[0, 1]$, which makes them interpretable as probabilities.

L_1-regularized logistic regression benefits from the strong learning guarantees already presented for conditional maxent models, in the special case of two classes ($c = 2$). The learning guarantees for L_2-regularized logistic regression will be similarly special cases of those presented in the next section.

13.8 L_2-regularization

A common variant of conditional Maxent models is one where the dimension N is finite and where the regularization is based on the norm-2 squared of the weight vector $\mathbf{w}$. The optimization problem is thus given by

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

where for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathsf{p}_{\mathbf{w}}[y|x] = \frac{\exp(\mathbf{w} \cdot \mathbf{\Phi}(x, y))}{Z(x)} \quad \text{and} \quad Z(x) = \sum_{y\in\mathcal{Y}} \exp(\mathbf{w} \cdot \mathbf{\Phi}(x, y)). \tag{13.18}$$

As for the norm-1 regularization, there are many optimization solutions available for this problem, including special-purpose algorithms, general first-order and second-order solutions, and special-purpose distributed solutions. Here, the objective is additionally differentiable. A common optimization method is simply stochastic gradient descent (SGD).

In contrast to norm-1-regularized conditional Maxent models, which lead to sparser weight vectors, norm-2 conditional Maxent models lead to non-sparse solutions, which may be preferable and lead to more accurate solutions in some applications such as natural language processing. The following margin-based guarantee holds for norm-2 regularized conditional Maxent, assuming that the norm-2 of the feature vector is bounded.

Theorem 13.4 *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, the following holds for all $\rho > 0$ and $f \in \mathcal{F} = \{(x, y) \mapsto \mathbf{w} \cdot \mathbf{\Phi}(x, y) \colon \|\mathbf{w}\|_2 \leq 1\}$:*

$$R(f) \leq \frac{1}{m} \sum_{i=1}^{m} \log_{u_0}\Big(\sum_{y \in \mathcal{Y}} e^{\frac{f(x_i, y) - f(x_i, y_i)}{\rho}}\Big) + \frac{4 r_2 c^2}{\rho \sqrt{m}} + \sqrt{\frac{\log \log_2 \frac{4 r_2}{\rho}}{m}} + \sqrt{\frac{\log \frac{2}{\delta}}{2m}},$$

where $u_0 = \log(1 + 1/e)$ and $r_2 = \sup_{(x,y)} \|\mathbf{\Phi}(x, y)\|_2$.

Proof: The proof is similar to that of Theorem 13.3, modulo the observation that here $|\mathbf{w} \cdot \mathbf{\Phi}(x, y)| \leq \|\mathbf{w}\|_2 \|\mathbf{\Phi}(x, y)\|_2 \leq r_2$ and modulo the upper bound on $\mathfrak{R}_m(\Pi_1(\mathcal{F}))$. For any sample $S = (x_1, \ldots, x_m)$ of size m, the empirical Rademacher complexity of $\Pi_1(\mathcal{F})$ can be bounded as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\Pi_1(\mathcal{F})) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Bigg[\sup_{\substack{\|\mathbf{w}\|_2 \leq 1 \\ y \in \mathcal{Y}}} \sum_{i=1}^{m} \sigma_i \mathbf{w} \cdot \mathbf{\Phi}(x_i, y)\Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Bigg[\sup_{\substack{\|\mathbf{w}\|_2 \leq 1 \\ y \in \mathcal{Y}}} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{\Phi}(x_i, y)\Bigg] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Bigg[\sup_{y \in \mathcal{Y}} \Bigg\|\sum_{i=1}^{m} \sigma_i \mathbf{\Phi}(x_i, y)\Bigg\|_2\Bigg] \\
&\leq \frac{1}{m} \sum_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Bigg[\Bigg\|\sum_{i=1}^{m} \sigma_i \mathbf{\Phi}(x_i, y)\Bigg\|_2\Bigg] \\
&\leq \frac{1}{m} \sum_{y \in \mathcal{Y}} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\Bigg[\Bigg\|\sum_{i=1}^{m} \sigma_i \mathbf{\Phi}(x_i, y)\Bigg\|_2^2\Bigg]} \\
&= \frac{1}{m} \sum_{y \in \mathcal{Y}} \sqrt{\sum_{i=1}^{m} \|\mathbf{\Phi}(x_i, y)\|_2^2\Bigg]} \leq \frac{r_2 c}{\sqrt{m}},
\end{aligned}$$

where the third equality holds by definition of the dual norm, and the second inequality by Jensen's inequality. □

The learning guarantee of the theorem for L_2-regularized conditional maxent models admits the advantage that the bound does not depend on the dimension. It can be very favorable for r_2 relatively small. The algorithm can then be very effective, provided that a small error can be achieved by a non-sparse weight vector.

13.9 Proof of the duality theorem

In this section, we give the full proof of Theorem 13.1.

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 optimization problem (13.4) with the functions f and g defined for all $\bar{\mathsf{p}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}$ and $\mathbf{u} \in \mathbb{R}^N$ by $f(\bar{\mathsf{p}}) = \mathbb{E}_{x \sim \widehat{\mathcal{D}}^1}\Big[\widetilde{\mathsf{D}}\big(\mathsf{p}[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]\big)\Big]$, $g(\mathbf{u}) = I_{\mathcal{C}}(\mathbf{u})$ and $A\mathsf{p} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \widehat{\mathcal{D}}^1(x)\mathsf{p}[y|x]\mathbf{\Phi}(x, y)$. A is a bounded linear map since we have $\|A\bar{\mathsf{p}}\| \leq \|\bar{\mathsf{p}}\|_1 \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} \|\mathbf{\Phi}(x, y)\|_\infty \leq r\|\bar{\mathsf{p}}\|_1$ for any $\bar{\mathsf{p}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}$. Also, notice that the conjugate of A is given for all $\mathbf{w} \in \mathbb{R}^N$ and $(x, y) \in \mathcal{X}_1 \times \mathcal{Y}$ by $(A^*\mathbf{w})(x, y) = \mathbf{w} \cdot \big(\widehat{\mathcal{D}}^1(x)\mathbf{\Phi}(x, y)\big)$.

Consider $\mathbf{u}_0 \in \mathbb{R}^N$ defined by $\mathbf{u}_0 = \mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x, y)] = A\bar{\mathsf{p}}_0$ with $\bar{\mathsf{p}}_0 = (\mathcal{D}(\cdot|x))_{x \in \mathcal{X}_1}$. Since $\bar{\mathsf{p}}_0$ is in $\mathrm{dom}(f) = \Delta^{\mathcal{X}_1}$, $\mathbf{u}_0$ is in $A(\mathrm{dom}(f))$. Furthermore, since λ is positive, $\mathbf{u}_0$ is contained 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 $\bar{\mathsf{q}} = (\mathsf{q}[\cdot|x])_{x \in \mathcal{X}_1} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}$ by

$$
\begin{aligned}
f^*(\bar{\mathsf{q}}) &= \sup_{\bar{\mathsf{p}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}} \left\{ \langle \mathsf{p}, \mathsf{q} \rangle - \sum_{x \in \mathcal{X}} \widehat{\mathcal{D}}^1(x)\, \widetilde{\mathsf{D}}(\mathsf{p}[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]) \right\} \\
&= \sup_{\bar{\mathsf{p}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}} \left\{ \sum_{x \in \mathcal{X}_1} \widehat{\mathcal{D}}^1[x] \sum_{y \in \mathcal{Y}} \frac{\mathsf{p}[y|x]\mathsf{q}[y|x]}{\widehat{\mathcal{D}}^1[x]} - \sum_{x \in \mathcal{X}_1} \widehat{\mathcal{D}}^1[x]\, \widetilde{\mathsf{D}}(\mathsf{p}[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]) \right\} \\
&= \sum_{x \in \mathcal{X}_1} \widehat{\mathcal{D}}^1(x) \sup_{\bar{\mathsf{p}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}} \left\{ \sum_{y \in \mathcal{Y}} \mathsf{p}[y|x] \left[\frac{\mathsf{q}[y|x]}{\widehat{\mathcal{D}}^1(x)} \right] - \widetilde{\mathsf{D}}(\mathsf{p}[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x]) \right\} \\
&= \sum_{x \in \mathcal{X}_1} \widehat{\mathcal{D}}^1(x) f_x^*\left(\frac{\mathsf{q}[y|x]}{\widehat{\mathcal{D}}^1(x)} \right),
\end{aligned}
$$

where, for all $x \in \mathcal{X}_1$ and $\mathsf{p} \in \mathbb{R}^{\mathcal{X}_1}$, f_x is defined by $f_x(\bar{\mathsf{p}}) = \widetilde{\mathsf{D}}(\mathsf{p}[\cdot|x] \,\|\, \mathsf{p}_0[\cdot|x])$. By Lemma B.37, the conjugate function f_x^* is given for all $\bar{\mathsf{q}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}$ by $f_x^*\left(\frac{\mathsf{q}[y|x]}{\widehat{\mathcal{D}}^1(x)}\right) = \log\left(\sum_{y \in \mathcal{Y}} \mathsf{p}_0[y|x] e^{\frac{\mathsf{q}[y|x]}{\widehat{\mathcal{D}}^1(x)}}\right)$. Thus, f^* is given for all $\bar{\mathsf{q}} \in (\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}$ by

$$
f^*(\mathsf{q}) = \mathbb{E}_{x \sim \widehat{\mathcal{D}}^1}\left[\log\left(\sum_{y \in \mathcal{Y}} \mathsf{p}_0[y|x] e^{\frac{\mathsf{q}[y|x]}{\widehat{\mathcal{D}}^1(x)}} \right) \right].
$$

As in the proof of Theorem 12.2, the conjugate function of $g = I_{\mathcal{C}}$ is given for all $\mathbf{w} \in \mathbb{R}^N$ by $g^*(\mathbf{w}) = \mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{w} \cdot \mathbf{\Phi}(x, y)] + \lambda\|\mathbf{w}\|_1$. In view of these identities, we can write, for all $\mathbf{w} \in \mathbb{R}^N$,

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

which proves that $\sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w}) = \min_{\bar{\mathsf{p}}\in(\mathbb{R}^{\mathcal{Y}})^{\mathcal{X}_1}} F(\bar{\mathsf{p}})$.

The second part of the proof is similar to that of Theorem 12.2. For any $\mathbf{w} \in \mathbb{R}^N$, we can write

$$\begin{aligned}
&G(\mathbf{w}) - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}^1}[\mathsf{D}(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] + \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}^1}[\mathsf{D}(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \lambda\|\mathbf{w}\|_1 - \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log \frac{\mathsf{p}^*[y|x]}{\mathsf{p}_0[y|x]}\right] + \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log \frac{\mathsf{p}^*[y|x]}{\mathsf{p}_{\mathbf{w}}[y|x]}\right] \\
&= -\lambda\|\mathbf{w}\|_1 + \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] \\
&= -\lambda\|\mathbf{w}\|_1 + \mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{w} \cdot \mathbf{\Phi}(x, y) - \log Z(\mathbf{w}, x)] - \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w} \cdot \mathbf{\Phi}(x, y) - \log Z(\mathbf{w}, x)] \\
&= -\lambda\|\mathbf{w}\|_1 + \mathbf{w} \cdot \Big[\mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x, y)] - \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}(x, y)]\Big].
\end{aligned}$$

As the solution of the primal optimization, $\bar{\mathsf{p}}^*$ verifies $I_{\mathcal{C}}\Big(\mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}(x, y)]\Big) = 0$, that is $\Big\|\mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}(x, y)] - \mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{\Phi}(x, y)]\Big\|_\infty \leq \lambda$. By Hölder's inequality, this implies the following inequality:

$$-\|\mathbf{w}\|_1 + \mathbf{w} \cdot \Big[\mathop{\mathbb{E}}_{(x,y)\sim\widehat{\mathcal{D}}}[\mathbf{w} \cdot \mathbf{\Phi}(x, y)] - \mathop{\mathbb{E}}_{\substack{x\sim\widehat{\mathcal{D}}^1\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w} \cdot \mathbf{\Phi}(x, y)]\Big] \leq -\|\mathbf{w}\|_1 + \|\mathbf{w}\|_1 = 0.$$

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

$$\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}^1}\Big[\mathsf{D}(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])\Big] \leq \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}^1}\Big[\mathsf{D}(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\Big] - G(\mathbf{w}).$$

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, $\mathbb{E}_{x\sim\widehat{\mathcal{D}}^1}[\mathsf{D}(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - G(\mathbf{w}) = (\sup_{\mathbf{w}} G(\mathbf{w})) - G(\mathbf{w}) \leq \epsilon$ implies the inequality $\mathbb{E}_{x\sim\widehat{\mathcal{D}}^1}[\mathsf{D}(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] \leq \epsilon$. This concludes the proof of the theorem. □

13.10 Chapter notes

The logistic regression model is a classical model in statistics. The term *logistic* was introduced by the Belgian mathematician Verhulst [1838, 1845]. An early reference for logistic regression is the publication of Berkson [1944] who advocated the use of the logistic function, instead of the cumulative distribution function of the standard normal distribution (*probit model*).

Conditional maximum entropy models in natural language processing were introduced by Berger et al. [1996] and were widely adopted for a variety of different tasks, including part-of-speech tagging, parsing, machine translation, and text categorization (see tutorial by Manning and Klein [2003]). Our presentation of the conditional Maxent principle, including their regularized variants, the duality theorem for conditional Maxent models (Theorem 13.1) and their theoretical justifications are based on [Cortes, Kuznetsov, Mohri, and Syed, 2015]. This chapter provided two types of justification for these models: one based on the conditional Maxent principle, another based on standard generalization bounds.

As in the case of Maxent models for density estimation, conditional Maxent models can be extended by using other Bregman divergences [Lafferty, Pietra, and Pietra, 1997] and other regularizations. Lafferty [1999] presented a general framework for incremental algorithms based on Bregman divergences that admits logistic regression as as special case, see also [Collins et al., 2002] who showed that boosting and logistic regression were special instances of a common framework based on Bregman divergences. The regularized conditional Maxent models presented in this chapter can be extended similarly using other Bregman divergences. In the binary classification case, when coordinate descent is used to solve the optimization problem of regularized conditional Maxent models, the algorithm coincides with L_1-regularized AdaBoost modulo the use of the logistic loss instead of the exponential loss.

Cortes, Kuznetsov, Mohri, and Syed [2015] presented a more general family of conditional probability models, *conditional structural Maxent models*, for which they also presented a duality theorem and gave strong learning guarantees. These Maxent models are based on feature functions selected from a union of possibly very complex sub-families. The resulting algorithms coincide with the DeepBoost algorithms of Cortes, Mohri, and Syed [2014] in the binary classification case or the multi-class DeepBoost algorithm of Kuznetsov, Mohri, and Syed [2014] in the multi-class classification case, when the logistic function is used as a convex surrogate loss function.

13.11 Exercises

13.1 Extension to Bregman divergences.

(a) Show how conditional Maxent models can be extended by using arbitrary Bregman divergences instead of the (unnormalized) relative entropy.

(b) Prove a duality theorem similar to Theorem 13.1 for theses extensions.

(c) Derive theoretical guarantees for these extensions. What additional property is needed for the Bregman divergence so that your learning guarantees hold?

13.2 Stability analysis for L_2-regularized conditional Maxent.

(a) Give an upper bound on the stability of the L_2-regularized conditional Maxent in terms of the sample size and λ (*Hint*: use the techniques and results of Chapter 14).

(b) Use the previous question to derive a stability-based generalization guarantee for the algorithm.

13.3 Maximum conditional Maxent. An alternative measure of closeness, instead of the conditional relative entropy, is the maximum relative entropy over all $x \in \mathcal{X}_1$.

(a) Write the primal optimization problem for this maximum conditional Maxent formulation. Show that it is a convex optimization problem, and discuss its feasibility and the uniqueness of its solution.

(b) Prove a duality theorem for maximum conditional Maxent and write the equivalent dual problem.

(c) Analyze the properties of maximum conditional Maxent and give a generalization bound for the algorithm.

13.4 Conditional Maxent with other marginal distributions: discuss and analyze conditional Maxent models when using a distribution $\mathcal{Q}$ over $\mathcal{X}$ instead of $\widehat{\mathcal{D}}^1$. Prove that a duality theorem similar to Theorem 13.1 holds.