MAXIMUM ENTROPY DENSITY ESTIMATION AND MODELING GEOGRAPHIC DISTRIBUTIONS OF SPECIES

Miroslav Dudík

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
COMPUTER SCIENCE

September 2007

Abstract

Maximum entropy (maxent) approach, formally equivalent to maximum likelihood, is a widely used density-estimation method. When input datasets are small, maxent is likely to overfit. Overfitting can be eliminated by various smoothing techniques, such as regularization and constraint relaxation, but theory explaining their properties is often missing or needs to be derived for each case separately. In this dissertation, we propose a unified treatment for a large and general class of smoothing techniques. We provide fully general guarantees on their statistical performance and propose optimization algorithms with complete convergence proofs. As special cases, we can easily derive performance guarantees for many known regularization types including L1 and L2-squared regularization. Furthermore, our general approach enables us to derive entirely new regularization functions with superior statistical guarantees. The new regularization functions use information about the structure of the feature space, incorporate information about sample selection bias, and combine information across several related density-estimation tasks. We propose algorithms solving a large and general subclass of generalized maxent problems, including all discussed in the dissertation, and prove their convergence. Our convergence proofs generalize techniques based on information geometry and Bregman divergences as well as those based more directly on compactness.

As an application of maxent, we discuss an important problem in ecology and conservation: the problem of modeling geographic distributions of species. Here, small sample sizes hinder accurate modeling of rare and endangered species. Generalized maxent offers several advantages over previous techniques. In particular, generalized maxent addresses the problem in a statistically sound manner and allows principled extensions to situations when data collection is biased or when we have access to data on many related species. The utility of our unified approach is demonstrated in comprehensive experiments on large real-world datasets. We find that generalized maxent is among the best-performing species-distribution modeling techniques. Our experiments also show that the contributions of this dissertation, i.e., regularization strategies, bias-removal approaches, and multiple-estimation techniques, all significantly improve the predictive performance of maxent.

Acknowledgements

First, I would like to thank my advisor Rob Schapire who gently nudged me along my PhD candidacy and encouraged me to work on problems that were not easy (a blessing in disguise). Discussions with him have been a source of tremendous insights, and observing him in action a source of inspiration. His talent to combine theory with applications and his ability to balance a large picture with attention to detail will continue to inspire me in future.

Next, I would like to thank my readers and collaborators Steven Phillips and Dave Blei, who actively participated in the research of this dissertation. Steven Phillips is the main creator of the *Maxent* software for modeling distributions of species, of which Rob and I furnished small parts and pieces. Without his endless devotion to *Maxent* and *Maxent* users (who now number in thousands), the results presented in this dissertation might not have seen their application (or certainly not as fast). I would like to thank Steven for an extraordinary experience of witnessing the rise of *Maxent* in the course of my PhD candidacy. Dave Blei encouraged me to explore the problem of multiple-density estimation. The maxent solution to this problem has turned into the second longest chapter in this dissertation, and possibly the most advanced application of the theory presented at the beginning. Dave's and Steven's comments on the manuscript of this dissertation and various preceding manuscripts have improved the scientific value and legibility of this text.

I cannot forget to thank my biggest fans, which is my family: my parents Miroslav and Anna, and my sister Anna. Their words of support made it across the Atlantic and worked their charms.

There are many friends that made my Princeton experience wonderful and thanks to whom I was able to keep my sanity during the times of stress. Two of them deserve special mention: Åsa and Martin. They have created places that I have considered my homes and provided me with peace, support, understanding, and, at times, with needed distraction.

My research was supported by NSF grant CCR-0325463.

Contents

Chapter 1

Introduction

Density estimation is a central task in classical statistics as well as statistical learning theory. This dissertation focuses on three problems in density estimation: estimation from few samples, estimation from biased samples, and simultaneous estimation of several related densities.

These three problems share one underlying goal: the efficient use of the training data. For example, given a small number of samples in a many-dimensional space, an efficient approach should use the information from all the dimensions. Many classical techniques fail in this scenario, suffering from the "curse of dimensionality." We are interested in techniques that overcome the curse of dimensionality, enabling a large number of dimensions for small datasets. When data is collected in a biased manner, any knowledge about the bias should improve the predictive performance. We look for techniques that efficiently incorporate information about the bias. Finally, when estimating several densities whose datasets are organized into overlapping groups, such as a hierarchy, the signal shared by the densities in a group should improve the accuracy of individual estimates. Successful multiple-estimation techniques seek balance between the information from the groups and the information from the individual datasets.

Our work is motivated by a new application of density estimation to modeling distributions of plant or animal species, a critical problem in conservation biology and ecology. We are given a set of locations, features describing them, and samples of where different species were observed. Our goal is to estimate the distribution of locations favored by each species based on the features of the kinds of places in which they are found. For example, we will consider species sampled from North America, such as the yellow-throated vireo, blue-headed vireo, and loggerhead shrike. All locations in North America are described by environmental variables such as elevation, annual precipitation, and average daily temperature. This dataset is described in more detail in Chapter 5.

Species distribution modeling exemplifies our three problems. We are often interested in modeling rare species, so the number of available samples may be quite small, calling for a technique that performs well for small sample sizes. The available data is often biased toward locations that are easier to access such as areas near roads, towns, airports, and waterways. The sample selection bias, reflecting the collectors' effort, can be estimated from the set of visited sites. The estimate of the bias provides additional information which can assist in removing the bias. Finally, even though the number of samples per species is quite small, we frequently have access to datasets containing many related species, such as the North American dataset. In such cases, it is desirable to share the strength of prediction across multiple species, using multiple-estimation techniques.

To address the problems of small-sample estimation, biased-sample estimation, and multiple estimation, we apply the *principle of maximum entropy*, and develop a single unified approach. The resulting framework combines key benefits of solutions to all three problems. Estimates from a small number of samples take advantage of a large numbers of features, the information about the sample selection bias is used for bias removal, and the group structure is easily incorporated to further improve the quality of predictions.

1.1 Overview of the Maximum-Entropy Principle

The maximum-entropy principle (maxent) originated in statistical mechanics, in the work of Boltzmann (1871c,b,a) and Gibbs (1902). As an approach to density estimation, it was first proposed by Jaynes (1957), and has since been used in many areas outside statistical mechanics (Kapur and Kesavan, 1992). In computer science, it has been particularly popular in natural language processing (Berger et al., 1996; Della Pietra et al., 1997).

In maxent, one is given a set of known constraints on the target distribution. The target distribution is then estimated by a distribution of maximum entropy satisfying the given constraints. The constraints are frequently based on a set of samples from the target distribution and represented using a set of *features* (real-valued functions) defined on the sample space. Typically, the constraints require the expectation of every feature to match its empirical average.

In species-distribution modeling, the goal is to estimate the density of a species over the pixels in a map. Features are simple functions derived from the environmental variables, and constraints are based on the observed occurrences of the species. For example, when modeling the distribution of the yellow-throated vireo across the pixels of North America, we may use the constraint derived from the feature "an-

nual precipitation," saying that the mean annual precipitation favored by the yellow-throated vireo should equal the average observed precipitation.

To determine the maxent distribution, it is possible to apply the method of Lagrange multipliers. By the Karush-Kuhn-Tucker optimality conditions, the maxent distribution is the maximum-likelihood distribution from an exponential family with features acting as sufficient statistics. In statistical mechanics, such exponential-family distributions are called *Gibbs distributions*. A detailed derivation of the equivalence of maximum entropy and maximum likelihood is in Chapter 2.

The first question that we should answer before delving into the details of the principle of maximum entropy is, "Why maximize the entropy?" or, equivalently, "Why use Gibbs distributions?" We will see below that this question has been quite satisfactorily answered by many different authors. The second question, important for a concrete implementation, is, "What are the right constraints?" This is studied and will be the subject of the theory we develop in Chapter 3. To answer the first question and understand the significance of the second question, we take a brief historical detour.

1.1.1 Maximum Entropy in Statistical Mechanics

We begin with the work of Boltzmann (1871c,b,a), who studied properties of gas bodies, viewed as systems composed of a large number of molecules. One of his central concerns was how the macroscopic state of the system is influenced by the microscopic properties of the system. The macroscopic state (macrostate) includes properties such as total volume, total number of molecules, and total energy. The microscopic state (microstate) is described by the properties of individual molecules such as their velocities and positions.

To simplify the discussion, assume that the molecules of the gas body occupy discrete states. These can be obtained, for example, by the discretization of positions and velocities of the molecules. A crucial quantity on both the macroscopic and the microscopic scale is the energy. The energy of each molecule is the sum of the kinetic energy, which depends only on the velocity of the molecule, and the potential energy, which depends only on the position of the molecule within a force field. We assume that the division of the state space into discrete cells is fine enough so that the energy of molecules within the same cell is almost constant, but coarse enough to allow a large number of molecules per cell. The microstate of the system can be viewed as a vector, listing for each molecule the cell it occupies. The macrostate is determined by the histogram of molecule counts across cells. Therefore, to describe the macrostate, it suffices to calculate the most likely histogram.

Here, Boltzmann appealed to the "principle of indifference," and posited that all

the microstates are equally likely. Thus, the most likely histogram is the one that can be realized by the largest number of microstates. To be more concrete, label the discrete cells as $1,2,\ldots,K$, denote the number of molecules in the k-th cell by N_k, and the total number of molecules by N. The total number of ways to realize a concrete allocation into cells is described by the multinomial coefficient

$$\frac{N!}{N_1!N_2!\cdots N_K!} \ . \tag{1.1}$$

Boltzmann looked for the set of occupancies N_k for which the number of possible realizations (1.1) is the maximum, while respecting the law of conservation of energy

$$\sum_{k=1}^{K} N_k E_k = E \ . \tag{1.2}$$

Here, E_k is the energy associated with the state k and E is the total energy.

Instead of maximizing Eq. (1.1) directly, it is computationally simpler to maximize its logarithm. The logarithm of Eq. (1.1) plays a central role in thermodynamics. When multiplied by Boltzmann's constant, it defines the thermodynamic entropy:

$$\text{thermodynamic entropy} \propto \ln\left(\frac{N!}{N_1!N_2!\cdots N_K!}\right) \approx \left(\sum_{k=1}^{K} N_k \ln\frac{N}{N_k}\right) \ .$$

In the last step, we used Stirling's approximation. Boltzmann's problem can be rephrased in terms of frequencies $p_k = N_k/N$ as

$$\text{maximize} \sum_{k=1}^{K} N p_k \ln\frac{1}{p_k} \tag{1.3}$$

$$\text{subject to the constraint} \sum_{k=1}^{K} p_k E_k = E/N \ . \tag{1.4}$$

Now, using the method of Lagrange multipliers, we arrive at the solution to Boltzmann's problem: the Boltzmann distribution

$$p_k \propto e^{\lambda E_k} \ .$$

Here, λ is the Lagrange multiplier ensuring that the average-energy constraint (1.4) is satisfied. Using the expression for the Boltzmann distribution, it is now possible to study various properties of gas bodies, for example, the distribution of gas density in a gravitational field.

1.1.2 Jaynes-Kullback Principle of Maximum Entropy

Boltzmann's reasoning can be re-interpreted using information theory and generalized to problems outside statistical mechanics. This was first noticed by Jaynes (1957), who even suggested that statistical mechanics "may become merely an example of statistical inference."

Jaynes, influenced by the information-theoretical work of Shannon (1948), argued that the thermodynamic entropy in Boltzmann's problem should be replaced by the information-theoretic entropy, quantifying how uncertain we are about the system. Our only knowledge about the system is summarized by the average-energy constraint (1.4). Among all distributions satisfying this constraint, we should choose the one that is "maximally non-committal with regard to the missing information," i.e., the one with the largest information-theoretic entropy

$$\mathrm{H}(p) = -\sum_{k=1}^{K} p_k \ln p_k \quad .$$

Since the information-theoretic entropy is a multiple of the thermodynamic entropy, its maximization yields the result that is identical to Boltzmann's solution.

Moreover, the principle of maximum entropy can be viewed as a generalization of the principle of indifference applied by Boltzmann. In statistical inference, the principle of maximum entropy tells us to represent an unknown distribution by the maximum-entropy distribution satisfying a given set of constraints. In Boltzmann's problem, the only feature is the energy, and the Boltzmann distribution is a one-feature instance of a Gibbs distribution.

In general, let the states be denoted by x and the state space by $\mathcal{X}$. If every state is assigned a vector of feature values $\boldsymbol{f}(x)$ then the resulting Gibbs distribution takes the form

$$p(x) \propto e^{\boldsymbol{\lambda}\cdot\boldsymbol{f}(x)}$$

for the appropriate vector of Lagrange multipliers $\boldsymbol{\lambda}$.

The information-theoretic justification of Jaynes was generalized by Kullback (1959) who assumed that in addition to a set of constraints we are also given a distribution q_0, serving as a default guess—the distribution we would choose if we had no data. He suggested choosing the distribution that is the closest to q_0 among all the distributions satisfying the constraints. The measure of closeness is the relative entropy,

$$\mathrm{D}(p \parallel q_0) = \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q_0(x)} \quad ,$$

also known as the Kullback-Leibler divergence, measuring how much information

about the outcome could be gained by knowing p instead of approximating it by q_0. If q_0 is uniform then the minimum relative entropy criterion is the same as the maximum entropy criterion. The resulting Gibbs distributions take the form

$$p(x) \propto q_0(x) e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x)} \ .$$

The information-theoretic motivation of Jaynes and Kullback was described by Beneš (1965) as "a reasonable and systematic way of throwing up our hands." However, it is not clear why the information-theoretic quantities such as entropy and relative entropy should be appropriate for the density estimation task. Even though a large body of research seems satisfied with the purely information-theoretic motivation (see, for example, references in Shore and Johnson, 1980), the apparent mismatch between the task at hand and the maximum-entropy principle prompted a large body of theoretical research, resulting in a variety of theoretical justifications. We mention three approaches, addressing maxent from three different perspectives: large-deviation theory, axiomatic derivation, and game theory.

1.1.3 Large-deviation Theory

The first perspective, large-deviation theory, is related to the original application of maximum entropy in statistical mechanics. Informally, large-deviation theory studies probabilities of unlikely events. In statistical mechanics, the unlikely events correspond to macrostates (or histograms) with submaximal entropy.

For example, we saw in Boltzmann's problem that the number of realizations of an empirical distribution p by N particles is

$$\frac{N!}{N_1! N_2! \cdots N_K!} = e^{N(\mathrm{H}(p) + o(1))} \ , \tag{1.5}$$

where $o(1)$ denotes an arbitrary function with limit zero as N tends to infinity. Thus, the maximum-entropy distribution $\hat{p}$ can be realized by

$$e^{N(\mathrm{H}(\hat{p}) + o(1))} \tag{1.6}$$

microstates. We would like to compare this number with the total number of realizations whose entropy is submaximal, i.e., whose entropy is less than $\mathrm{H}(\hat{p}) - \varepsilon$ for some fixed ε.

For any particular histogram p, the number of realizations is given by Eq. (1.5). Thus, if $\mathrm{H}(p) \leq \mathrm{H}(\hat{p}) - \varepsilon$ then the number of realizations of the histogram p will be at most $e^{N(\mathrm{H}(\hat{p}) - \varepsilon + o(1))}$. The number of histograms with entropy at most $\mathrm{H}(\hat{p}) - \varepsilon$ is trivially bounded by the total number of histograms of N particles, each in one of K

states. This is at most $(N+1)^K$. Thus the total number of realizations with entropy at most $\mathrm{H}(\hat{p}) - \varepsilon$ is at most

$$(N+1)^K e^{N(\mathrm{H}(\hat{p})-\varepsilon+o(1))} = e^{N(\mathrm{H}(\hat{p})-\varepsilon+o(1))} \quad .$$

If N is sufficiently large, this number of realizations will be exponentially smaller than the number of realizations of $\hat{p}$. Thus, among all empirical distributions satisfying the constraints, all but an exponentially small fraction lie inside an arbitrarily small neighborhood of the maximum-entropy distribution.

The previous reasoning assumes that all realizations are *a priori* equally likely. When each molecule is assigned to its cell according to an *a priori* distribution q_0, the entropy $\mathrm{H}(p)$ in the previous arguments needs to be replaced by the negative relative entropy $-\mathrm{D}(p \parallel q_0)$.

The foregoing arguments are special cases of Sanov's theorem (Sanov, 1957), which states that the empirical distribution under a set of constraints approaches the maximum-entropy distribution. Sanov's theorem can be further generalized to a set of results known as conditional limit theorems (Van Campenhout and Cover, 1981; Csiszár, 1984; Grünwald, 2001). For example, assume that the constraints define a convex set of probability distributions and samples are drawn independently from q_0, which itself does not satisfy the constraints. Then it can be argued by Csiszár's conditional limit theorem that the conditional distribution of the first sample, on the condition that the empirical distribution p satisfies the constraints, converges to the maximum entropy distribution $\hat{p}$. The statement of Csiszár's conditional limit theorem is significantly stronger than that of Sanov's theorem. Rather than a statement about the empirical distribution of all N particles, it is a statement about the marginal distribution of a single particle.

Sanov's theorem and conditional limit theorems characterize the properties of the empirical state of a system under known conditions, assuming that the empirical state is generated by the distribution q_0. They provide a strong justification for maxent in statistical mechanics (Csiszár, 1995), but it is not clear how useful they are in statistical inference, when q_0 is a mere default estimate, and the goal is to infer the *unknown* sample-generating distribution.

1.1.4 Axiomatic Approaches

The problem of statistical inference is addressed more directly by axiomatic approaches (Shore and Johnson, 1980; Skilling, 1988; Csiszár, 1991). These approaches begin by formulating a set of properties desirable for consistent statistical inference, such as invariance under changes of coordinates and consistency under decompositions

into disjoint subsystems. From these properties it is then derived that the only method of statistical inference which satisfies all the conditions simultaneously is the principle of maximum entropy.

Unfortunately, the desirable properties postulated in axiomatic approaches are not entirely self-evident. In this respect, axiomatic approaches resemble the original justifications of Jaynes and Kullback. The desirability conditions—be it the maximum ignorance of Jaynes, the minimum relative entropy of Kullback, or a set of consistency properties of Shore and Johnson—are somewhat arbitrary and seem external to the problem of density estimation (at least from the perspective of a machine learning practitioner, such as the author).

In this dissertation, the solution quality is often evaluated on a test set consisting of samples which are withheld during training. Frequencies observed in the test set are seen as approximations of some "true" probabilities, which would be obtained in the limit of infinitely many samples. This view of probabilities is called *frequency interpretation* (Hájek, 2007).

In machine learning and statistics, the most common alternative to frequency interpretation is *Bayesian interpretation*. Bayesian interpretation postulates prior probabilities over all densities in a given family. After seeing the evidence, an orthodox Bayesian does not produce a single density estimate, but instead derives a posterior distribution over all possible densities based on the prior and the evidence. A less orthodox Bayesian may choose a single density that maximizes the posterior.

Frequency and Bayesian interpretations relate density estimation problems to observed samples either directly, through frequencies, or indirectly, through the posterior. Such a link is, however, missing in the justifications of maximum entropy introduced above. Information-theoretic arguments are based on the principle of maximum indifference, and therefore yield *classical interpretation* of probabilities (Hájek, 2007). This interpretation is what we use when analyzing a shuffled deck of cards. We assume that each permutation is equally likely, similar to Boltzmann's assumption that each microstate is equally likely. Axiomatic approaches are similar to the principle of indifference in that they try to deduce probabilities "from the basic principles," rather than from a model of sample generation. The principle of indifference is replaced by a set of consistency requirements. Both lines of justification shed little light on how to derive constraints from the observations.

1.1.5 Game-theoretic Perspective

The final justification of maximum entropy, introduced by Topsøe (1979), adopts a view that is more accessible to frequency interpretation. We will exploit this view to connect sample generation with density estimation.

Topsøe frames maxent within game theory similar to the decision-theoretic setup of classical statistics. Specifically, he considers the density-estimation game with two players: nature and the decision maker. Nature is allowed to choose any distribution π that satisfies a given set of constraints. The decision maker only knows the set of constraints, but not the distribution π, and would like to choose a distribution q that achieves the largest expected log likelihood relative to the log likelihood achieved by a default estimate q_0. Thus the decision maker tries to maximize

$$\sum_{x\in\mathcal{X}} \pi(x)\ln q(x) - \sum_{x\in\mathcal{X}} \pi(x)\ln q_0(x) = \sum_{x\in\mathcal{X}} \pi(x)\ln\frac{q(x)}{q_0(x)} \ .$$

The best strategy of the decision maker is to choose the distribution $\hat{q}$ that maximizes the worst-case log likelihood:

$$\hat{q} = \operatorname*{argmax}_{q\in\Delta} \min_{\pi\in\mathcal{P}} \left(\sum_{x\in\mathcal{X}} \pi(x)\ln\frac{q(x)}{q_0(x)} \right) \ ,$$

where Δ denotes the set of all densities on a given sample space and $\mathcal{P}$ denotes the set of densities satisfying given constraints. Topsøe shows that the max-min likelihood density $\hat{q}$ is identical to the minimum relative entropy (or maximum entropy) density

$$\hat{p} = \operatorname*{argmin}_{p\in\mathcal{P}} \mathrm{D}(p \parallel q_0) \ .$$

One might ask, “Why maximize the likelihood?" and there are several justifications, including optimal gambling and optimal coding (Cover and Thomas, 1991). Are these justifications less arbitrary than the justifications of maxent by information theory and axiomatic derivations?

We believe that there is a difference. Instead of imposing desirability conditions and seeking the rational guess of the distribution, max-min likelihood directly identifies the performance measure and optimizes it relative to the “true” distribution. The assumption about the existence of a single true distribution (obtained possibly in the limit of infinitely many samples) is inherently frequentist. Thus, max-min likelihood can be viewed as a frequentist interpretation of maximum entropy. In this work, we adopt the frequentist view, but for historical reasons we continue to refer to the problem as maxent. We use the name maxent for both the maximum entropy formalism of Jaynes and the minimum relative entropy formalism of Kullback.

1.1.6 Maximum Entropy versus Maximum Likelihood

As mentioned before, maxent with equality constraints based on empirical averages is equivalent to maximum likelihood in an exponential family. The dual interpretation of maxent as maximum likelihood has been suggested as an alternative justification of maxent (Jaynes, 1978). However, the maximum-likelihood setting in classical statistics (see, for example, Lehmann and Casella, 1998, Chapter 6) differs from the maxent setting in several aspects. First, in maximum likelihood, the true distribution is assumed to be from the same family as the distributions over which the likelihood is maximized. Maxent poses no parametric assumptions on the truth. Second, the goal of maximum likelihood is parameter estimation rather than density estimation, so the analysis typically focuses on the asymptotic properties of parameter estimates. In maxent, it is more natural to compare actual distributions instead of parameters: the maxent distribution with the true distribution, or possibly the maxent distribution with the best approximation of the truth by a Gibbs distribution. Thus, likelihood serves as a measure of maxent performance rather than a device for estimating parameters.

1.1.7 Constraints and Overfitting in Maximum Entropy

A crucial question arising in any concrete implementation of maxent is how to choose the set of constraints. The most common are equality constraints on feature expectations, introduced in the original papers of Jaynes and Kullback. Although other types of constraints appear in the literature (Csiszár, 1975; Jaynes, 1978; Shore and Johnson, 1980; Khudanpur, 1995), they have received little attention in practical applications until the 2000s. Yet, according to the max-min likelihood interpretation, the choice of correct constraints may be essential. For if the true distribution does not lie in the set $\mathcal{P}$, then the decision maker is foolishly optimizing against a wrong enemy. Conversely, if the set $\mathcal{P}$ is too large, then the decision maker is too cautious, optimizing against the enemies that should be perhaps ruled out.

When equality constraints are based on empirical averages, it should not be surprising that maxent can severely overfit the training data. For instance, in our application, we sometimes consider threshold features for each environmental variable. These are binary features equal to one if an environmental variable is larger than a fixed threshold and zero otherwise. Thus, there is a continuum of features for each variable, and together they force the maxent distribution to be non-zero only at values achieved by the samples. The problem is that in general, the empirical averages of features will almost never be equal to their true expectations, so the target distribution does not satisfy the constraints imposed on the output distribution. This

problem is exacerbated by small sample sizes. From the dual perspective, the exponential family is too expressive and the maximum likelihood distribution overfits. Common approaches to counter overfitting are parameter regularization (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001; Zhang, 2005), introduction of a prior (Williams, 1995; Goodman, 2004), feature selection (Berger et al., 1996; Della Pietra et al., 1997), discounting (Lau, 1994; Rosenfeld, 1996; Chen and Rosenfeld, 2000), and constraint relaxation (Khudanpur, 1995; Kazama and Tsujii, 2003; Jedynak and Khudanpur, 2005).

Regularization techniques control overfitting by introducing a penalty term in the dual objective (i.e., log likelihood). The penalty is usually a monotone function of a norm of the parameter vector, such as the ℓ_1 norm (the sum of absolute values) or the ℓ_2 norm (the Euclidean norm). Optimization of the regularized objective seeks balance between goodness of fit and complexity of the solution. A Bayesian approach to prevent overfitting is introduction of a prior and, instead of maximization of likelihood, maximization of the posterior. The resulting models thus balance prior knowledge with the observed data. Other approaches, including feature selection, discounting, and constraint relaxation, maximize entropy subject to adjusted constraints which encourage simpler solutions. Thus, there are many ways of modifying maxent to control overfitting calling for a general treatment.

1.2 Outline and Contributions

In this work, we study a generalized form of maxent. Although mentioned by other authors as *fuzzy maxent* (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001), we give the first complete theoretical treatment of this very general framework, including fully general and unified performance guarantees, algorithms, and convergence proofs. Our unified treatment leads to a principled approach to problems of small-sample estimation, biased estimation, and multiple estimation.

In the problem of estimating a single density, from small as well as large samples, our general results allow us to easily derive performance guarantees for many known regularized formulations, including ℓ_1, ℓ_2, ℓ_2^2, and ℓ_1+ℓ_2^2 regularizations. More specifically, we derive guarantees on the performance of maxent solutions compared to the "best" Gibbs distribution $q^\star$. Our guarantees are derived by bounding deviations of empirical feature averages from their expectations, a setting in which we can take advantage of a wide array of uniform convergence results. Our bounds depend very favorably on the number or complexity of features. For example, we prove that ℓ_1-regularized maxent yields accurate models as long as the features are bounded and their number is smaller than the exponential of the sample size.

A crucial insight of our general analysis is that maxent relaxations corresponding to tighter constraints on the feature expectations yield better performance guarantees. We use this insight throughout the dissertation to derive novel regularization functions and a corresponding analysis for our three problems. In our first problem, small-sample estimation, the performance of ℓ_1 regularization can be improved when some information about the structure of the feature space is available, for example, when some features are known to be squares or products of other "base" features, corresponding to constraints on variances or covariances of the base features. We apply our general framework to derive improved generalization bounds using an entirely new form of regularization. These results improve on bounds for previous forms of regularization by up to a factor of eight—an improvement that would otherwise require a 64-fold increase in the number of training examples.

In Chapter 4, we propose algorithms solving a large and general subclass of generalized maxent problems. We show convergence of our algorithms using a technique that unifies previous approaches and extends them to a more general setting. Specifically, our unified approach generalizes techniques based on information geometry and Bregman divergences (Della Pietra et al., 1997, 2001; Collins et al., 2002) as well as those based more directly on compactness. The main novel ingredient is a modified definition of an auxiliary function, a customary measure of progress, which we view as a surrogate for the difference between the primal and dual objective rather than a bound on the change in the dual objective.

Standard maxent algorithms such as iterative scaling (Darroch and Ratcliff, 1972; Della Pietra et al., 1997), gradient descent, Newton and quasi-Newton methods (Cesa-Bianchi et al., 1994; Malouf, 2002; Salakhutdinov et al., 2003), and their regularized versions (Lau, 1994; Williams, 1995; Chen and Rosenfeld, 2000; Kazama and Tsujii, 2003; Goodman, 2004; Krishnapuram et al., 2005) perform a sequence of feature-weight updates until convergence. In each step, they update all feature weights. This is impractical when the number of features is very large. Instead, we propose a sequential-update algorithm that updates only one feature weight in each iteration, along the lines of algorithms studied by Collins, Schapire, and Singer (2002), and Lebanon and Lafferty (2001). This leads to a boosting-like approach permitting the selection of the best feature from a very large class. For instance, for ℓ_1-regularized maxent, the best threshold feature associated with a single variable can be found in a single linear pass through the (pre-sorted) data, even though conceptually we are selecting from an infinite class of features. Other boosting-like approaches to density estimation have been proposed by Welling, Zemel, and Hinton (2003), and Rosset and Segal (2003).

Sequential updates are especially desirable when the number of features is very

large or when they are produced by a weak learner. When the number of features is relatively small, yet we want to use benefits of regularization to prevent overfitting on small sample sets, it might be more efficient to solve generalized maxent by parallel updates, similar to standard algorithms such as iterative scaling and quasi-Newton approaches. To address this problem, we derive a parallel-update version of our algorithm, generalizing the iterative-scaling approaches mentioned above, and prove its convergence.

In Chapter 5, we return to the application of maxent to species-distribution modeling. We explore the performance of ℓ_1-regularized maxent in an application of estimating distributions of four bird species in North America. We analyze how the choice of the feature set influences the predictive accuracy of maxent, depending on the number of occurrence records. We explore effects of regularization on predictive accuracy and interpretability of the maxent models.

We also evaluate maxent on a comprehensive dataset of 226 species from 6 regions. This dataset was developed by a working group at the National Center for Ecological Analysis and Synthesis (NCEAS) as part of a large-scale comparison of species distribution modeling methods (Elith, Graham et al., 2006). We refer to the data as "the NCEAS dataset," and the comparison of methods as "the NCEAS comparison." Both the NCEAS dataset and methods participating in the NCEAS comparison are described in more detail in Chapter 5. Here, we simply note that the NCEAS dataset has two portions: the training portion, with data of low quality as typical in many applications, and the evaluation portion, obtained by independent, rigorously planned surveys.

In preliminary experiments on the North American bird dataset, we observed that the performance of maxent depends on the choice of feature classes and the amount of regularization. We optimize the performance of maxent on the NCEAS dataset by tuning the regularization parameters on a small portion of the training data. The models are then constructed from all of the training data and evaluated on the evaluation data.

Among the twelve methods in the NCEAS comparison, ℓ_1-regularized maxent is among the best methods alongside boosted regression trees (Leathwick et al., 2006), generalized dissimilarity models (Ferrier et al., 2002) and multivariate adaptive regression splines with the community-level selection of basis functions (Leathwick et al., 2005). Among these, however, maxent is the only method designed for presence-only data. The remaining methods are based on regression and require data on species absence. Since the data on species absence is expensive to collect and thus missing in most datasets, the absences are usually replaced by pseudo-absences. This complicates the analysis as well as interpretation of the resulting models.

The careful tuning of maxent regularization parameters on the NCEAS dataset has one additional goal: to determine well-performing "default settings." Default settings are desirable, because the parameter tuning may be prohibitively time-consuming to do separately for each species, or unreliable for small or biased datasets. Additionally, even with the abundance of good quality data, users interested in the application of species models need not have the statistical knowledge required for detailed tuning. To assess the quality of the settings determined from the NCEAS training data, we compare their performance with the optimal performance of the settings tuned on the evaluation data itself. We find that the potential improvements in performance are very small, and conclude that the settings determined on the NCEAS training data can be used as default settings.

In Chapter 6, we explore maxent solutions to the problem of biased estimation. We propose two bias correction approaches. The first approach is based on our generalization analysis, according to which the maxent constraints should determine tight confidence regions for the unbiased feature means. We use the biased sample and the information about sample-selection bias to determine unbiased confidence regions. The second approach first estimates the biased distribution, using the biased sample, and then factors the bias out. We evaluate our bias correction approaches on a synthetic dataset as well as the NCEAS dataset. We view the training portion of the NCEAS dataset as the biased training data and the evaluation portion as the unbiased test data. The information about sample selection bias is provided by the total of training records across all species, demonstrating how the collectors' effort varies across the regions of interest.

We find that bias correction approaches improve maxent performance in both synthetic and real-data experiments. The improvement in performance on the NCEAS dataset is especially dramatic, comparable with the gap between the four best performing methods and the remaining eight methods of the NCEAS comparison. Inclusion of bias information leads to similar improvements in regression-based techniques.

In Chapter 7, we turn to the problem of multiple estimation. We use insights from our generalization analysis to develop *hierarchical maximum entropy density estimation*, a procedure that allows sharing of information among single-density problems. The datasets are grouped, and the individual estimates are adjusted to reflect that grouping. With this approach, estimates from small sample sizes are influenced by the estimates for which we have more confidence; estimates from large sample sizes are less influenced by others. In statistics, this is known as *hierarchical/multilevel modeling* (Gelman and Hill, 2007) or *shrinkage*, introduced by Stein (1956) and James and Stein (1961). In machine learning, hierarchical models have been used,

for example, by McCallum et al. (1998) and Teh et al. (2005). These methods are also related to *multitask* or *transfer learning* (Caruana, 1993; Baxter, 2000; Raina et al., 2006)

In hierarchical maximum entropy, we assume that we are given a fixed class hierarchy. We fit the joint distribution of all classes, placing constraints on individual class distributions as well as on groups of classes defined by the hierarchy. We show that our approach is closely related to maximum *a posteriori* estimation with a hierarchical prior, or maximum likelihood estimation with hierarchical regularization (shrinkage). We apply our generalization theory and demonstrate how to choose hyperparameters in this setting. We prove strong generalization guarantees. We report the utility of hierarchical maximum entropy on a small synthetic dataset and on two regions of the NCEAS dataset.

Work in this dissertation overlaps with the previously published work of the author (Phillips, Dudík, and Schapire, 2004; Dudík, Phillips, and Schapire, 2004; Dudík, Schapire, and Phillips, 2006; Elith, Graham et al., 2006; Dudík and Schapire, 2006; Dudík, Phillips, and Schapire, 2007; Dudík, Blei, and Schapire, 2007; Phillips and Dudík, 2007; Phillips, Dudík et al., 2007).

Chapter 2

Maximum Entropy and Convex Duality

In this chapter, we introduce a generalization of maxent and derive its convex dual—maximum regularized log likelihood. This generalization has been mentioned by other authors as *fuzzy maxent* (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001), but the duals were derived only for a few specific cases. We derive the general dual using a convex-analysis result known as Fenchel's duality theorem. Before describing the generalized maxent setting, we review the basic maxent setting of Jaynes and Kullback.

2.1 Basic Maximum Entropy

The goal of density estimation is to estimate an unknown density π over a *sample space* $\mathcal{X}$. Throughout this dissertation we assume that $\mathcal{X}$ is discrete; hence any density π can be identified with a probability mass function on $\mathcal{X}$. As empirical information, we are given a set of *samples* $x_1, \ldots, x_m$ drawn independently at random from π. The corresponding empirical distribution is denoted by $\tilde{\pi}$:

$$\tilde{\pi}(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(x_i = x)$$

where $\mathbb{1}(P(x))$ denotes the *binary indicator*, which is a function of x, equal to one when the predicate $P(x)$ is true, and equal to zero when the predicate $P(x)$ is false. The available information about the space $\mathcal{X}$ is expressed by *features* f_j where $f_j : \mathcal{X} \to \mathbb{R}$ and j comes from an index set $\mathcal{J}$. The set of features is denoted by $\mathcal{F}$, the vector of all features is denoted by $\boldsymbol{f}$.

One naive approach to estimating π is an approximation by the empirical distri-

bution $\tilde{\pi}$. However, when the size of the sample space $\mathcal{X}$ is much larger than the number of samples m, the empirical distribution $\tilde{\pi}$ will be quite distant, under any reasonable measure, from π.

On the other hand, for a given function f, we do expect $\mathbf{E}_{X\sim\tilde{\pi}}[f(X)]$, the empirical average of f, to be rather close to its true expectation $\mathbf{E}_{X\sim\pi}[f(X)]$. It is quite natural, therefore, to seek an approximation p under which f_j's expectation is equal to $\mathbf{E}_{X\sim\tilde{\pi}}[f_j(X)]$ for every f_j. There will typically be many distributions satisfying these constraints. The *maximum entropy principle* suggests that, from among all distributions satisfying these constraints, we choose the one of maximum entropy, i.e., the one that is closest to uniform. However, the *default estimate* of π, i.e., the distribution we would choose if we had no sample data, may be in some cases non-uniform. In a more general setup, we therefore seek a distribution that minimizes entropy relative to the default estimate q_0.

Instead of maximizing entropy or minimizing the relative entropy, we could posit a family of distributions $q_{\boldsymbol{\lambda}}$ parameterized by $\boldsymbol{\lambda}$, and approximate π by the maximum likelihood distribution from the family, i.e., by the distribution which maximizes the likelihood of the data

$$\prod_{i=1}^{m} q_{\boldsymbol{\lambda}}(x_i) \ .$$

It can be proved (Della Pietra et al., 1997) that the maximum entropy distribution and the maximum likelihood distribution are equal when the family of distributions is the *exponential family* with features as *sufficient statistics* and the distribution q_0 as the *base measure*; in other words, when the family is defined by

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

where $Z_{\boldsymbol{\lambda}} = \sum_{x\in\mathcal{X}} q_0(x)e^{\boldsymbol{\lambda}\cdot\boldsymbol{f}(x)}$ is the normalizing constant, and $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the vector of parameters. Distributions $q_{\boldsymbol{\lambda}}$ will also be referred to as *Gibbs distributions*.

Instead of maximizing the likelihood, we could equivalently minimize the empirical log loss (negative normalized log likelihood relative to the default q_0)

$$\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) = -\frac{1}{m}\sum_{i=1}^{m} \ln\frac{q_{\boldsymbol{\lambda}}(x_i)}{q_0(x_i)} \ .$$

Summarizing the previous, we obtain the following two optimization problems

$$\min_{p\in\Delta} \mathrm{D}(p \parallel q_0) \text{ subject to } \mathbf{E}_p[\boldsymbol{f}] = \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] \tag{2.1}$$

$$\inf_{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{J}}} \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) \tag{2.2}$$

where Δ is the simplex of probability distributions over $\mathcal{X}$. We use the shorthand $\mathbf{E}_p[\boldsymbol{f}]$ for $\mathbf{E}_{X\sim p}[\boldsymbol{f}(X)]$.

In general, we write

$$\mathrm{L}_r(\boldsymbol{\lambda}) = -\mathbf{E}_r\left[\ln \frac{q_{\boldsymbol{\lambda}}}{q_0}\right] = \ln Z_{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\cdot\mathbf{E}_r[\boldsymbol{f}] \tag{2.3}$$

to denote the log loss of $q_{\boldsymbol{\lambda}}$ on the distribution r relative to the default q_0. It differs from the relative entropy $\mathrm{D}(r \parallel q_{\boldsymbol{\lambda}})$ only by the additive constant $\mathrm{D}(r \parallel q_0)$:

$$\mathrm{L}_r(\boldsymbol{\lambda}) = \mathrm{D}(r \parallel q_{\boldsymbol{\lambda}}) - \mathrm{D}(r \parallel q_0) \ . \tag{2.4}$$

We will use the two interchangeably as objective functions. In particular, note that Eq. (2.2) minimizes the relative entropy between the empirical distribution $\tilde{\pi}$ and distributions q in the closure of the exponential family.

2.2 Feature Types and Exponential Families

The choice of feature types determines the exponential family used in the maxent dual to approximate π. We stress that π need not belong to this exponential family. Here we discuss various feature types and the resulting exponential families. To be concrete, we consider the species modeling setup where features are derived from environmental variables $v : \mathcal{X} \to \mathbb{R}$, $v \in \mathcal{V}$. Environmental variables might be continuous, such as altitude, annual precipitation, and average temperature, or categorical, such as soil type or vegetation type.

2.2.1 Linear, Quadratic, and Product Features

For a continuous variable v, its corresponding linear and quadratic features are defined by

$$f_v(x) = v(x) \ , \qquad f_{v^2}(x) = v^2(x) \ .$$

For a pair of distinct continuous variables v, w, the corresponding product feature is

$$f_{vw}(x) = v(x)w(x) \ .$$

Linear features in basic maxent require that the expectations of individual variables match their empirical means. For example, an average elevation where a yellow-throated vireo was observed should match the expected elevation according to the maxent model. Linear and quadratic variables jointly constrain both means and variances of the environmental variables. Product features with the respective linear

features constrain means and covariances.

For an example of an exponential family resulting from these features, consider the feature set including linear and quadratic features. In this case, Gibbs distributions take the form

$$q_{\boldsymbol{\lambda}}(x) = q_0(x) \exp\left\{ \sum_{v \in \mathcal{V}} \left[\lambda_{v^2} v^2(x) + \lambda_v v(x) \right] \right\} \Big/ Z_{\boldsymbol{\lambda}} \ . \tag{2.5}$$

Formally, this resembles a Gaussian with uncorrelated components. Specifically, consider a Gaussian random variable taking values $\boldsymbol{\xi} \in \mathbb{R}^K$, with mean $\boldsymbol{\mu}$, and a diagonal covariance matrix with variances $\sigma_1^2, \ldots, \sigma_K^2$. The density of $\boldsymbol{\xi}$ is then

$$p(\boldsymbol{\xi}) \propto \exp\left\{ \sum_{k=1}^{K} -\frac{(\xi_k - \mu_k)^2}{2\sigma_k^2} \right\} \propto \exp\left\{ \sum_{k=1}^{K} \left[-\frac{1}{2\sigma_k^2} \xi_k^2 + \frac{\mu_k}{\sigma_k^2} \xi_k \right] \right\} \ . \tag{2.6}$$

Now, identifying components ξ_k with variables v, and setting $\lambda_v = \mu_k / \sigma_k^2$ and $\lambda_{v^2} = -1/(2\sigma_k^2)$, we find that the exponents of Eqs. (2.5) and (2.6) formally agree. The only difference is that the base measure of the Gaussian density is the Lebesgue measure whereas the base measure of $q_{\boldsymbol{\lambda}}$ is q_0. However, as a result, the two exponential families have qualitatively different properties. For example, λ_{v^2} can be both positive and negative, whereas the variance σ_k^2 is always nonnegative. Thus, the exponential family defined by Eq. (2.5) includes multimodal distributions, whereas diagonal Gaussian distributions defined by Eq. (2.6) are always unimodal.

2.2.2 Categorical Indicator Features

For a categorical variable $v : \mathcal{X} \to \mathcal{C}$, where $\mathcal{C}$ is a discrete subset of $\mathbb{R}$, we define a categorical indicator for each category $c \in \mathcal{C}$ as $f_{v=c}(x) = \mathbb{1}(v(x) = c)$.

2.2.3 Threshold Features

For a continuous variable v and a threshold θ, there are two threshold features

$$f_{v \geq \theta}(x) = \mathbb{1}(v(x) \geq \theta) \ , \qquad f_{v < \theta}(x) = \mathbb{1}(v(x) < \theta) \ .$$

Formally, we consider a continuum of threshold features for each variable. In practice, it suffices to consider a single threshold between each pair of consecutive values appearing in the sample space $\mathcal{X}$. Thus, in the worst case, there will be $|\mathcal{X}| - 1$ distinct threshold features for each variable. Note that for each variable, the sum of its threshold features weighted by the corresponding λ's can express an arbitrary piecewise constant function of the variable. Linear combinations of threshold fea-

tures across all variables can represent arbitrary additive "responses" in the exponent. The *response* refers to the exponent of the Gibbs distribution when viewed as a function of environmental variables rather than the point x. For example, Eq. (2.5) models arbitrary quadratic responses (without interactions, which can be introduced by adding product features).

2.2.4 Hinge Features and Splines

Threshold features can model arbitrary piecewise constant functions. The resulting responses to environmental variables will be, however, discontinuous. To obtain continuous responses, we introduce hinge features, which model continuous piecewise linear functions of variables. There are two types of hinge features for each continuous variable v and threshold θ

$$f_{\text{hinge};v\geq\theta}(x) = \begin{cases} \dfrac{v(x)-\theta}{v_{\max}-\theta} & \text{if } v(x)\geq\theta \\ \\ 0 & \text{if } v(x)<\theta, \end{cases} \qquad f_{\text{hinge};v<\theta}(x) = \begin{cases} 0 & \text{if } v(x)\geq\theta \\ \\ \dfrac{\theta-v(x)}{\theta-v_{\min}} & \text{if } v(x)<\theta, \end{cases}$$

where $v_{\min}$ and $v_{\max}$ are the minimum and maximum values of v on $\mathcal{X}$. The scaling by $v_{\max}-\theta$ and $\theta-v_{\min}$ ensures that hinge features have values in $[0,1]$.

Formally, hinge features can be defined in terms of *clamped linear functions*

$$\mathsf{lh}(t;a,b) = \begin{cases} 0 & \text{if } \dfrac{t-a}{b-a}<0 \\ \\ \dfrac{t-a}{b-a} & \text{if } \dfrac{t-a}{b-a}\in[0,1] \\ \\ 1 & \text{if } \dfrac{t-a}{b-a}\geq 1 \ . \end{cases}$$

Specifically,

$$f_{\text{hinge};v\geq\theta}(x) = \mathsf{lh}(v(x);\theta,v_{\max}) \ , \qquad f_{\text{hinge};v<\theta}(x) = \mathsf{lh}(v(x);\theta,v_{\min}) \ .$$

Similar to threshold features, we only consider a single threshold between each pair of consecutive variable values. Unlike threshold features, it is not possible to represent all hinge features in this manner. However, when hinge features are used jointly with threshold features, then one hinge and one threshold feature at the same threshold can represent a hinge feature at an arbitrary threshold between the two given variable values.

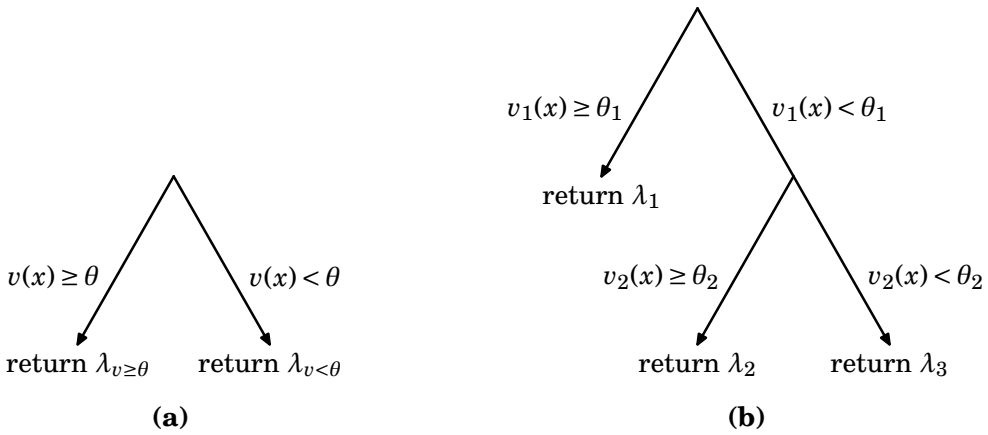


Figure 2.1. *Examples of a regression stump (a), and a regression tree (b).*

Hinge features give rise to continuous response functions, but their first derivatives are discontinuous. When a higher-order continuity is needed then hinge features can be replaced by a suitable spline basis.

2.2.5 Regression Trees and Multivariate Splines

Threshold and hinge features model arbitrary additive responses. Here we consider their generalizations that model higher-order interactions. Consider a pair of threshold features $f_{v \geq \theta}$ and $f_{v < \theta}$ for a fixed variable v and threshold θ. The linear combination

$$\lambda_{v \geq \theta} f_{v \geq \theta}(x) + \lambda_{v < \theta} f_{v < \theta}(x)$$

can be viewed as a regression stump which assigns the value $\lambda_{v \geq \theta}$ to the points x with $v(x) \geq \theta$, and value $\lambda_{v < \theta}$ to the points x with $v(x) < \theta$, as depicted in Fig. 2.1(a). Threshold features can be generalized to implement regression trees of arbitrary depth by introducing more complicated features corresponding to paths from the root to leaves. For example, the regression tree in Fig. 2.1(b) can be represented by the linear combination

$$\lambda_1 f_1(x) + \lambda_2 f_2(x) + \lambda_3 f_3(x)$$

where f_1, f_2, f_3 are *decision-path features*, or simply *decision paths*,

$$\begin{aligned} f_1(x) &= \mathbb{1}(v_1(x) \geq \theta_1) \\ f_2(x) &= \mathbb{1}(\{v_1(x) < \theta_1\} \cap \{v_2(x) \geq \theta_2\}) \\ f_3(x) &= \mathbb{1}(\{v_1(x) < \theta_1\} \cap \{v_2(x) < \theta_2\}) \ . \end{aligned}$$

Note that decision paths can be written as products of threshold features, for example,

$$f_2(x) = \mathbb{1}(\{v_1(x) < \theta_1\} \cap \{v_2(x) \geq \theta_2\}) = \mathbb{1}(v_1(x) < \theta_1)\mathbb{1}(v_2(x) \geq \theta_2) \ .$$

In classification, decision-path features are building blocks of alternating decision trees (Freund and Mason, 1999).

Similar to decision paths, it is possible to introduce products of hinge features, which we call *path hinge features*, yielding continuous versions of regression trees. For example, f_2 could be replaced by

$$f_2'(x) = \mathrm{lh}(v_1(x); \theta_1, v_{1;\min})\mathrm{lh}(v_2(x); \theta_2, v_{2;\max}) \ .$$

In regression settings, path hinge features are used for example in multivariate adaptive regression splines (Friedman, 1991). Again, if smooth first or second derivatives are desired, it is possible to use products of higher-order splines.

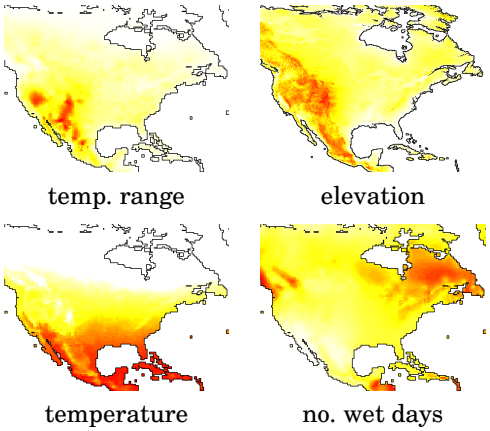
2.3 Overfitting and Smoothing

As mentioned in Chapter 1, maxent can severely overfit training data when the constraints on the output distribution are based on empirical averages, especially for a very large number of features. For instance, constraints derived from threshold or hinge features force the output distribution to be non-zero only at values achieved by the samples (see Fig. 2.2).

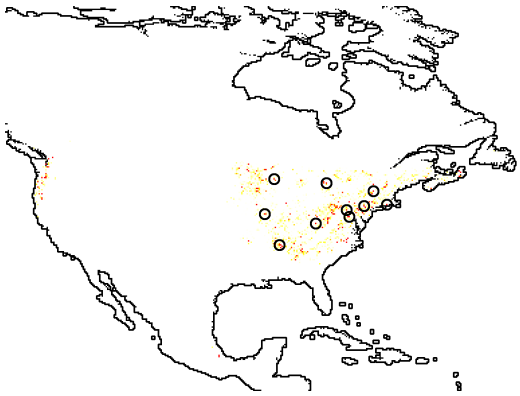
The problem is that in general, the empirical averages of features will almost never be equal to their true expectations, so the target distribution itself does not satisfy the constraints imposed on the output distribution. From the dual perspective, the chosen exponential family is too expressive. As a result, maxent fits an overly complex solution to the training data, while failing to capture the dependencies of the true distribution.

Common approaches to counter overfitting are parameter regularization (Lau, 1994; Chen and Rosenfeld, 2000; Lebanon and Lafferty, 2001; Zhang, 2005), introduction of a prior (Williams, 1995; Goodman, 2004), feature selection (Berger et al., 1996; Della Pietra et al., 1997), discounting (Lau, 1994; Rosenfeld, 1996; Chen and Rosenfeld, 2000) and constraint relaxation (Khudanpur, 1995; Kazama and Tsujii, 2003). Here we briefly discuss each of them.

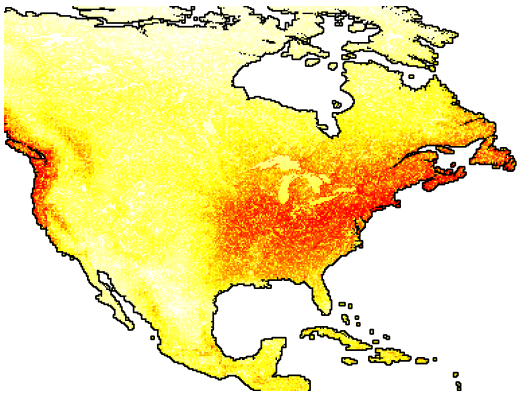
(a) Occurrence records.



(b) Example environmental variables.



(c) Maxent model: overfitting.



(d) Maxent model with ℓ_1 regularization.

Figure 2.2. *Overfitting and regularization.* We have used hinge features derived from seven environmental variables (four of them shown) to model the bird species "yellow-throated vireo" (more details in Section 5.3). (a,b) Maxent inputs: ten occurrence records and examples of environmental variables (larger values shown darker or red). (c) Maxent without regularization overfits, zooming on a small number of pixels whose environmental-variable values are represented exactly in the training set. Note the high predicted probabilities (shown as darker or red) located in the exact centers of the circles which correspond to the locations of training samples. The remaining dark pixels exactly match at least one of the environmental-variable values observed in the training set. (d) The overfitting disappears with the use of regularization.

2.3.1 Feature Selection and Constraint Exclusion

Perhaps the simplest way to obtain smoother distributions is to delete some of the constraints since maximizing the entropy subject to fewer constraints yields distributions closer to uniform. For categorical variables, we may, for example, omit constraints on indicators of the categories with too few observations. Alternatively, we could employ a more sophisticated feature selection scheme, weeding out "unreliable" features, such as features with large sample variances.

A technique complementary to constraint exclusion is feature induction. In feature induction, one begins with a relatively simple set of features, for example, threshold features, and then introduces new features to improve the training accuracy. New features are derived from the existing feature set. For example, beginning with threshold features as trivial decision paths, one may consider all decision paths obtained by appending a single node to all existing decision paths.

2.3.2 Discounting

Discounting techniques are applied predominantly to categorical features. Discounting is based on the observation that categories with low counts in the training sample typically overestimate true probabilities of occurrences, whereas categories with zero counts typically underestimate these probabilities. Instead of removing the corresponding constraints altogether, discounting decreases the target mean values for indicators of low-count categories and adds the missing mass to zero-count categories.

2.3.3 Regularization

Regularization is a common approach to smoothing in optimization and approximation, originally introduced by Tikhonov (1963b,a), Ivanov (1962), and Phillips (1962) as a method of finding solutions to *ill-posed problems*. In statistics, regularization was first introduced implicitly as *shrinkage* (Stein, 1956; James and Stein, 1961), and later explicitly as part of *ridge regression* (Hoerl and Kennard, 1970).

The main idea is to include in the objective a penalty for the ruggedness of the solution. The goal is to remove some of the noise present in finite sampling and to make the optimum unique. The two most commonly used penalty functions are the ℓ_1 norm

$$\|\boldsymbol{\lambda}\|_1 = \sum_{j \in \mathcal{J}} |\lambda_j|$$

and the ℓ_2 norm squared

$$\|\boldsymbol{\lambda}\|_2^2 = \sum_{j \in \mathcal{J}} \lambda_j^2 \ .$$

In the context of least squares regression, they are called the *lasso* penalty and the *ridge* penalty. They can be applied to the maximum likelihood problem as follows:

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{J}}}\left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\beta\|\boldsymbol{\lambda}\|_1\right]\ , \qquad \min_{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{J}}}\left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\frac{\alpha}{2}\|\boldsymbol{\lambda}\|_2^2\right]\ .$$

Here, β and α are tuning parameters specifying the tradeoff between the tightness of fit, as expressed by the log loss, and the complexity of the solution, as expressed by the regularization. To understand how the norms characterize the complexity of the solution, consider the example of threshold features. Here larger values of $|\lambda_j|$ correspond to larger jumps in response curves of the corresponding variables.[1] Hinge features behave similarly, with larger values of $|\lambda_j|$ corresponding to larger changes in the slope. As a result, the responses characterized by large norms are more rugged and possibly more prone to fitting models of the noise.

2.3.4 Introduction of a Prior

In frequentist settings, maxent is often justified by its dual formulation as the maximum likelihood. In Bayesian settings, the likelihood function should be complemented with a prior over parameters, and instead of maximizing the likelihood, we should determine the posterior distribution, or, less orthodoxly, maximize the posterior. For example, if the prior over $\boldsymbol{\lambda}$ is a Gaussian with mean zero and a diagonal covariance matrix with components σ_j^2 then the posterior is proportional to

$$\left(\prod_{j\in\mathcal{J}} e^{-\lambda_j^2/(2\sigma_j^2)}\right)\left(\prod_{i=1}^{m} q_{\boldsymbol{\lambda}}(x_i)\right)\ .$$

Taking the negative log of the posterior, we find that maximizing the posterior is equivalent to minimizing the ℓ_2^2-regularized log loss:

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{J}}}\left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\sum_{j\in\mathcal{J}}\frac{\lambda_j^2}{2m\sigma_j^2}\right]\ .$$

Similarly, ℓ_1-style regularization is equivalent to a Laplace prior. Other priors give rise to other regularization types, and conversely other regularizations can be viewed as representing various priors.

[1] Responses modeled by threshold features are additive, so they can be decomposed into responses to individual variables.

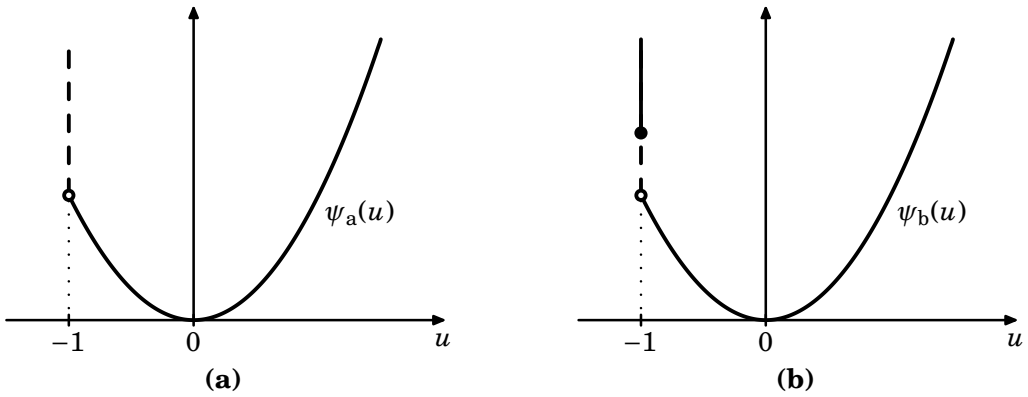


Figure 2.3. *Examples of convex functions which are not closed.* (a) $\psi_a(u) = u^2$ if $u > -1$, and $\psi_a(u) = \infty$ otherwise. (b) $\psi_b(u) = u^2$ if $u > -1$, $\psi_b(-1) = 1.5$, and $\psi_b(u) = \infty$ if $u < -1$.

2.3.5 Constraint Relaxation

A crucial problem with basic maxent is that the true distribution itself does not match the constraints exactly. There are many possibilities how to relax constraints. For example, equality constraints can be relaxed to inequalities

$$\left|\mathbf{E}_p[f_j] - \mathbf{E}_{\tilde{\pi}}[f_j]\right| \leq \beta_j \text{ for all } j \in \mathcal{J}.$$

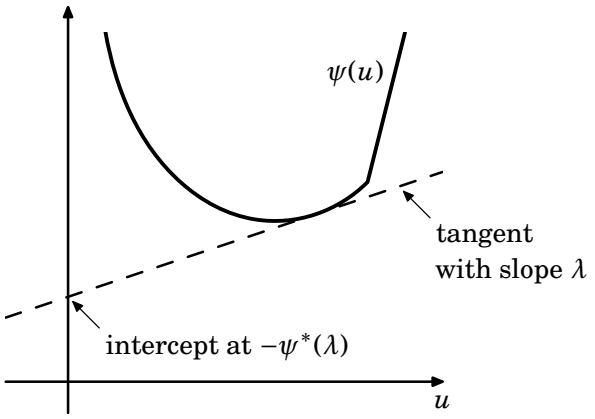
The maxent distribution under the relaxed constraints is closer to the default distribution q_0. Thus, when q_0 is uniform, the solution of relaxed maxent will be smoother than the solution of basic maxent.

Note that constraints need not be separable, i.e., they need not decompose into independent constraints on individual feature means. For example, consider the vector of categorical indicators $\boldsymbol{f}_{\mathcal{C}}$ derived from a single categorical variable $v : \mathcal{X} \to \mathcal{C}$. For an arbitrary distribution p, the expectation $\mathbf{E}_p[\boldsymbol{f}_{\mathcal{C}}]$ is itself a distribution over categories. Thus, it may be more natural to use relative entropy to measure the deviation between $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}_{\mathcal{C}}]$ and $\mathbf{E}_p[\boldsymbol{f}_{\mathcal{C}}]$, and hence introduce a non-separable inequality constraint

$$\mathrm{D}\big(\mathbf{E}_p[\boldsymbol{f}_{\mathcal{C}}] \,\big\|\, \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}_{\mathcal{C}}]\big) \leq \beta \ . \tag{2.7}$$

2.4 Convex Analysis Background

We have seen that there are many ways to prevent overfitting in maxent, calling for a unified treatment. Before we introduce such a treatment, we will need a few concepts from convex analysis. These concepts will be used throughout this dissertation. For a more detailed exposition see Rockafellar (1970) or Boyd and Vandenberghe (2004).



Fenchel's inequality:

$$\psi(u) \geq \underbrace{\lambda u - \psi^*(\lambda)}_{\text{equation of tangent with slope } \lambda}$$

Figure 2.4. *Convex conjugates and Fenchel's inequality.* Value $-\psi^*(\lambda)$ is defined as the vertical-axis intercept of the tangent to ψ's epigraph with slope λ.

Consider a function $\psi : \mathbb{R}^n \to (-\infty, \infty]$. The *effective domain* of ψ is the set $\operatorname{dom}\psi = \{\boldsymbol{u} \in \mathbb{R}^n : \psi(\boldsymbol{u}) < \infty\}$. A point $\boldsymbol{u}$ where $\psi(\boldsymbol{u}) < \infty$ is called *feasible*. The *epigraph* of ψ is the set of points above its graph $\{(\boldsymbol{u}, t) \in \mathbb{R}^n \times \mathbb{R} : t \geq \psi(\boldsymbol{u})\}$. We say that ψ is *convex* if its epigraph is a convex set. A convex function is called *proper* if it is not uniformly equal to ∞. It is called *closed* if its epigraph is closed. For a proper convex function, closedness is equivalent to lower semi-continuity (ψ is lower semi-continuous if $\liminf_{\boldsymbol{u}' \to \boldsymbol{u}} \psi(\boldsymbol{u}') \geq \psi(\boldsymbol{u})$ for all $\boldsymbol{u}$). Examples of convex functions which are not closed are given in Fig. 2.3.

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

$$\psi^*(\boldsymbol{\lambda}) = \sup_{\boldsymbol{u} \in \mathbb{R}^n} [\boldsymbol{\lambda} \cdot \boldsymbol{u} - \psi(\boldsymbol{u})] \ . \tag{2.8}$$

The conjugate provides an alternative description of ψ in terms of tangents to ψ's epigraph. Specifically, $-\psi^*(\boldsymbol{\lambda})$ is defined as the vertical-axis intercept of a tangent to ψ's epigraph (see Fig. 2.4). The definition of the conjugate immediately yields *Fenchel's inequality*

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

which simply states that the graph of a convex function lies above its tangent (see Fig. 2.4). It turns out that the conjugate ψ^* is a closed proper convex function and $\psi^{**} = \psi$ (for a proof see Rockafellar, 1970, Corollary 12.2.1).[2]

In this work we use several examples of closed proper convex functions. The first of them is the relative entropy, viewed as a function of its first argument and

[2]Convex conjugates are defined for arbitrary functions (not necessarily closed or convex) and Fenchel's inequality remains valid. However, the identity $\psi = \psi^{**}$ holds only for closed proper convex functions.

extended to $\mathbb{R}^{\mathcal{X}}$ as follows:

$$\psi(p) = \begin{cases} \mathrm{D}(p \parallel q_0) & \text{if } p \in \Delta \\ \infty & \text{otherwise} \end{cases} \tag{2.9}$$

where $q_0 \in \Delta$ is assumed fixed. In the following two propositions we derive the conjugate of the relative entropy.

Proposition 2.1. *If $p, q \in \Delta$ then $\mathrm{D}(p \parallel q) \geq 0$, with equality if and only if $p = q$.*

Proof. By Jensen's inequality

$$\begin{aligned} \mathrm{D}(p \parallel q) &= \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = -\sum_{x \in \mathcal{X}} p(x) \ln \frac{q(x)}{p(x)} \\ &\geq -\ln \left(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)} \right) = -\ln \left(\sum_{x \in \mathcal{X}} q(x) \right) = 0 \ , \end{aligned}$$

with equality if and only if $p(x)/q(x)$ is a constant, i.e., if $p(x) = q(x)$ for all x (since both p and q are probability densities). ■

Proposition 2.2. *The conjugate of the relative entropy is the log partition function*

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

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

Proof. To argue that $-\psi^*(r)$ specifies vertical-axis intercepts of tangents to ψ's epigraph, it suffices to show that Fenchel's inequality is satisfied for all $r \in \mathbb{R}^{\mathcal{X}}$ and $p \in \Delta$, and that for every $r \in \mathbb{R}^{\mathcal{X}}$ there exists $p_r \in \Delta$ for which Fenchel's inequality holds with equality.

Set

$$p_r(x) = \frac{q_0(x) e^{r(x)}}{\sum_{x \in \mathcal{X}} q_0(x) e^{r(x)}} \ .$$

First we show that Fenchel's inequality holds:

$$\begin{aligned} \sum_{x \in \mathcal{X}} r(x) p(x) - \psi^*(r) &= \sum_{x \in \mathcal{X}} r(x) p(x) - \ln \left(\sum_{x \in \mathcal{X}} q_0(x) e^{r(x)} \right) \\ &= \sum_{x \in \mathcal{X}} p(x) \ln \left(\frac{e^r(x)}{\sum_{x \in \mathcal{X}} q_0(x) e^{r(x)}} \right) \\ &= \mathrm{D}(p \parallel q_0) - \mathrm{D}(p \parallel p_r) \\ &\leq \mathrm{D}(p \parallel q_0) = \psi(p) \end{aligned}$$

where the last inequality follows by Proposition 2.1. By Proposition 2.1, we also obtain than equality holds instead of the last inequality if $p = p_r$. ■

The second example of a closed proper convex function is the unnormalized relative entropy

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

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

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

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

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

Similar to the relative entropy, the unnormalized relative entropy $\widetilde{\mathrm{D}}(p \parallel q)$ is always non-negative, and zero if and only if $p = q$. Its conjugate can be derived directly by setting partial derivatives on the right-hand side of Eq. (2.8) equal to zero.

The relative entropy is a measure of the distance between distributions, whereas the unnormalized relative entropy is a measure of the distance between non-negative vectors. Although neither of them is a metric (for example, they are not symmetric), they satisfy the following two properties

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

where B stands for either D or $\widetilde{\mathrm{D}}$. These properties are motivated by the formalism of Bregman divergences (Bregman, 1967; Censor and Lent, 1981; Censor and Zenios, 1997), which generalize some common distance measures such as the squared Euclidean distance.[3]

Next example of a closed proper convex function is a *convex indicator* of a closed convex set $C \subseteq \mathbb{R}^n$, denoted by I_C, which equals 0 when its argument lies in C and infinity otherwise. We will also use the notation $\mathrm{I}(P(\boldsymbol{u}))$ or $\mathrm{I}(\boldsymbol{u}; P(\boldsymbol{u}))$ to denote $\mathrm{I}_C(\boldsymbol{u})$

[3]Property (B1) is satisfied by all Bregman divergences, whereas property (B2) is satisfied by all Bregman divergences under further conditions on $\boldsymbol{a}_t$ and $\boldsymbol{b}_t$ (see Censor and Zenios, 1997, Definition 2.1.1). The unnormalized relative entropy is a Bregman divergence. The relative entropy is not a Bregman divergence because its domain has an empty interior. However, in many applications, the relative entropy inherits properties of Bregman divergences because it is a restriction of the unnormalized relative entropy to the simplex.

where $P(\boldsymbol{u})$ is a predicate defining the set C. The conjugate of a convex indicator is a *support function*, which satisfies (by the definition of conjugacy)

$$\mathrm{I}_C^*(\boldsymbol{\lambda}) = \sup_{\boldsymbol{u}\in C} \boldsymbol{\lambda}\cdot\boldsymbol{u} \quad . \tag{2.10}$$

For $C = \{\boldsymbol{c}\}$, we obtain $\mathrm{I}_{\{\boldsymbol{c}\}}^*(\boldsymbol{\lambda}) = \boldsymbol{\lambda}\cdot\boldsymbol{c}$. For a box $B = \{\boldsymbol{u} : |u_j| \le \beta_j \text{ for all } j\}$, we obtain an ℓ_1-style conjugate $\mathrm{I}_B^*(\boldsymbol{\lambda}) = \sum_j \beta_j|\lambda_j|$. For a Euclidean ball $B' = \{\boldsymbol{u} : \|\boldsymbol{u}\|_2 \le \beta\}$, we obtain an ℓ_2-style conjugate, $\mathrm{I}_{B'}^*(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_2$.

If C is a convex hull of two closed convex sets C_1, C_2 then

$$\mathrm{I}_C^*(\boldsymbol{\lambda}) = \max\{\mathrm{I}_{C_1}^*(\boldsymbol{\lambda}), \mathrm{I}_{C_2}^*(\boldsymbol{\lambda})\} \quad . \tag{2.11}$$

(For a proof see Rockafellar, 1970, Corollary 16.5.1.) In particular, if C is a convex polytope with vertex set V then

$$\mathrm{I}_C^*(\boldsymbol{\lambda}) = \max_{\boldsymbol{u}\in V} \boldsymbol{\lambda}\cdot\boldsymbol{u} \quad . \tag{2.12}$$

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

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

$$\text{if } \varphi(\boldsymbol{u}) = a\psi(b\boldsymbol{u}+\boldsymbol{c}) \qquad \text{then } \varphi^*(\boldsymbol{\lambda}) = a\psi^*(\boldsymbol{\lambda}/(ab)) - \boldsymbol{\lambda}\cdot\boldsymbol{c}/b \tag{2.13}$$

$$\text{if } \varphi(\boldsymbol{u}) = \psi(\mathbf{A}\boldsymbol{u}) \qquad \text{then } \varphi^*(\boldsymbol{\lambda}) = \psi^*(\mathbf{A}^{-\top}\boldsymbol{\lambda}) \tag{2.14}$$

$$\text{if } \varphi(\boldsymbol{u}) = \textstyle\sum_j \varphi_j(u_j) \qquad \text{then } \varphi^*(\boldsymbol{\lambda}) = \textstyle\sum_j \varphi_j^*(\lambda_j) \tag{2.15}$$

where $a > 0$, $b \neq 0$, $\boldsymbol{c} \in \mathbb{R}^n$, $\mathbf{A}$ is an invertible square matrix, $\mathbf{A}^{-\top}$ denotes the transpose of the inverse of $\mathbf{A}$, and u_j, λ_j refer to the components of $\boldsymbol{u}, \boldsymbol{\lambda}$.

A convex function is called *polyhedral* if its epigraph is an intersection of a finite number of halfspaces. Proper polyhedral functions are always closed and their conjugates are also polyhedral. Examples of polyhedral functions include linear functions, the ℓ_1 norm, and box indicators.

Next, assume that $\varphi : \mathbb{R}^n \to (-\infty, \infty]$ can be written as a sum of two closed proper convex functions φ_1 and φ_2 such that one of the following conditions is satisfied: (i) $\mathrm{dom}\,\varphi_1 = \mathbb{R}^n$, (ii) $\mathrm{dom}\,\varphi_2 = \mathbb{R}^n$, or (iii) φ_1 and φ_2 are polyhedral and $\mathrm{dom}\,\varphi_1 \cap \mathrm{dom}\,\varphi_2 \neq \emptyset$. Then the conjugate φ^* is the *infimal convolution* of φ_1^* and φ_2^* (see Rockafellar, 1970, Theorem 20.1)

$$\varphi^*(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\lambda}'} \left[\varphi_1^*(\boldsymbol{\lambda}') + \varphi_2^*(\boldsymbol{\lambda} - \boldsymbol{\lambda}') \right] \quad . \tag{2.16}$$

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

Theorem 2.3 (Fenchel's Duality)**.** *Let* $\psi : \mathbb{R}^n \to (-\infty, \infty]$ *and* $\varphi : \mathbb{R}^m \to (-\infty, \infty]$ *be closed proper convex functions and* $\mathbf{A}$ *be a real-valued* $m \times n$ *matrix. Assume that* $\operatorname{dom}\psi^* = \mathbb{R}^n$ *or* $\operatorname{dom}\varphi = \mathbb{R}^m$*. Then*

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

We refer to the minimization over $\boldsymbol{u}$ as the primal problem and the maximization over $\boldsymbol{\lambda}$ as the dual problem. When no ambiguity arises, we also refer to the minimization of the negative dual objective as the dual problem. We call $\boldsymbol{u}$ a primal feasible point if the primal objective is finite at $\boldsymbol{u}$. If the primal has a feasible point, i.e., if its objective is proper, then we say that the primal is feasible. Similarly, we define feasibility for the dual.

2.5 Generalized Maximum Entropy

In this dissertation we study a generalized maxent problem

$$\min_{p \in \Delta} \Big[\mathrm{D}(p \parallel q_0) + \mathrm{U}(\mathbf{E}_p[\boldsymbol{f}]) \Big] \tag{2.17}$$

where $\mathrm{U} : \mathbb{R}^n \to (-\infty, \infty]$ is an arbitrary closed proper convex function. It is viewed as a *potential* for the maxent problem. We further assume that q_0 is positive on $\mathcal{X}$, i.e., $\mathrm{D}(p \parallel q_0)$ is finite for all $p \in \Delta$ (otherwise we could restrict $\mathcal{X}$ to the support of q_0), and there exists a distribution whose vector of feature expectations is a feasible point of U (this is typically satisfied by the empirical distribution). These two conditions imply that the problem (2.17) is feasible.

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

$$\left|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_p[f_j]\right| \le \beta_j \ .$$

Box constraints are represented by the potential $\mathrm{U}^{(1)}(\boldsymbol{u}) = \mathrm{I}(|\mathbf{E}_{\tilde{\pi}}[f_j] - u_j| \le \beta_j \text{ for all } j)$. Finally, as noted by Chen and Rosenfeld (2000) and Lebanon and Lafferty (2001), ℓ_2^2-

regularized maxent is obtained using the potential $\mathrm{U}^{(2)}(\boldsymbol{u}) = \|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \boldsymbol{u}\|_2^2/(2\alpha)$ which incurs an ℓ_2^2-style penalty for deviating from empirical averages.

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

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

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

To derive the dual of Eq. (2.17), define the matrix $\mathbf{F}$ with elements $F_{jx} = f_j(x)$ and use Fenchel's duality:

$$\begin{aligned}
\min_{p \in \Delta} \Big[\mathrm{D}(p \parallel q_0) + \mathrm{U}(\mathbf{E}_p[\boldsymbol{f}]) \Big] &= \min_{p \in \Delta} \Big[\mathrm{D}(p \parallel q_0) + \mathrm{U}(\mathbf{F}p) \Big] \\
&= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^J} \Big[-\ln \big(\textstyle\sum_{x \in \mathcal{X}} q_0(x) \exp\{(\mathbf{F}^\top \boldsymbol{\lambda})_x\} \big) - \mathrm{U}^*(-\boldsymbol{\lambda}) \Big] \qquad (2.18) \\
&= \sup_{\boldsymbol{\lambda} \in \mathbb{R}^J} \big[-\ln Z_{\boldsymbol{\lambda}} - \mathrm{U}^*(-\boldsymbol{\lambda}) \big] \ . \qquad (2.19)
\end{aligned}$$

In Eq. (2.18), we apply Theorem 2.3. We use $(\mathbf{F}^\top \boldsymbol{\lambda})_x$ to denote the entry of $\mathbf{F}^\top \boldsymbol{\lambda}$ indexed by x. In Eq. (2.19), we note that $(\mathbf{F}^\top \boldsymbol{\lambda})_x = \boldsymbol{\lambda} \cdot \boldsymbol{f}(x)$ and thus the expression inside the logarithm is the normalization constant of $q_{\boldsymbol{\lambda}}$. The maximization in Eq. (2.19) is the maxent dual. Its objective will be referred to as Q:

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

There are two formal differences between generalized maxent and basic maxent. The first difference is that the constraints of the basic primal (2.1) are stated relative to empirical expectations whereas the potential of the generalized primal (2.17) makes no reference to $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]$. This difference is only superficial. It is possible to hard-wire the distribution $\tilde{\pi}$ in the potential U, as we saw in the example of $\mathrm{U}^{(0)}(\boldsymbol{u})$. In the latter case, it would be more correct, but perhaps overly pedantic, to make the dependence of the potential on $\tilde{\pi}$ explicit and use the notation $\mathrm{U}^{(0)}(\boldsymbol{u}; \tilde{\pi})$.

The second difference, which seems more significant, is the difference between the duals. The objective of the basic dual (2.2) equals the log loss relative to the empirical distribution $\tilde{\pi}$, but the generalized dual contains no log-loss terms. We will see that the generalized dual can be expressed in terms of the log loss as well. In fact, it can be expressed in terms of the log loss relative to an arbitrary distribution, including the empirical distribution $\tilde{\pi}$ as well as the unknown distribution π.

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

Table 2.1. *Absolute and relative potentials, and their conjugates.*

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

2.5.1 Shifting

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

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

This function will be referred to as the *potential relative to* r or simply the *relative potential*. In contrast, the original potential U will be referred to as the *absolute potential*. In Table 2.1, we list the potentials discussed so far, alongside their versions relative to an arbitrary distribution r, and relative to $\tilde{\pi}$ in particular.

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

The conjugate of a relative potential, the *conjugate relative potential*, is obtained, according to Eq. (2.13), by adding a linear function to the conjugate of U:

$$\mathrm{U}^*_r(\boldsymbol{\lambda}) = \mathrm{U}^*(-\boldsymbol{\lambda})+\boldsymbol{\lambda}\cdot\mathbf{E}_r[\boldsymbol{f}] \ . \tag{2.20}$$

Table 2.1 lists $\mathrm{U}^{(0)*}$, $\mathrm{U}^{(1)*}$, $\mathrm{U}^{(2)*}$, and the conjugates of the corresponding relative potentials.

2.5.2 Generalized Dual as Minimization of a Regularized Log Loss

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

Comparing Eq. (2.20) with Eq. (2.3), we obtain

$$\mathrm{L}_r(\boldsymbol{\lambda}) + \mathrm{U}_r^*(\boldsymbol{\lambda}) = \ln Z_{\boldsymbol{\lambda}} + \mathrm{U}^*(-\boldsymbol{\lambda}) = -Q(\boldsymbol{\lambda}) \ . \tag{2.21}$$

Thus the maximization of $Q(\boldsymbol{\lambda})$ is equivalent to the minimization of $\mathrm{L}_r(\boldsymbol{\lambda}) + \mathrm{U}_r^*(\boldsymbol{\lambda})$. Setting $r = \tilde{\pi}$ we obtain a dual analogous to the basic dual (2.2):

$$\inf_{\boldsymbol{\lambda} \in \mathbb{R}^J} \left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^*(\boldsymbol{\lambda}) \right] \ . \tag{2.22}$$

From Eq. (2.21), it follows that the $\boldsymbol{\lambda}$ minimizing $\mathrm{L}_r(\boldsymbol{\lambda}) + \mathrm{U}_r^*(\boldsymbol{\lambda})$ does not depend on a particular choice of r. As a result, the minimizer of (2.22) is also the minimizer of $\mathrm{L}_\pi(\boldsymbol{\lambda}) + \mathrm{U}_\pi^*(\boldsymbol{\lambda})$. This observation will be used in Chapter 3 to prove performance guarantees.

The objective of Eq. (2.22) has two terms. The first of them is the empirical log loss. The second one is the regularization term penalizing "complex" solutions. The regularization term need not be non-negative and it does not necessarily increase with any norm of $\boldsymbol{\lambda}$. On the other hand, it is a proper closed convex function and if $\tilde{\pi}$ is feasible then by Fenchel's inequality the regularization is bounded from below by $-\mathrm{U}_{\tilde{\pi}}(\mathbf{0})$. From a Bayesian perspective, $\mathrm{U}_{\tilde{\pi}}^*$ corresponds to negative log of the prior, and minimizing $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^*(\boldsymbol{\lambda})$ is equivalent to maximizing the posterior.

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

In basic maxent, maxent with box constraints, and maxent with ℓ_2^2 regularization, it is natural to consider dual objectives relative to $\tilde{\pi}$. In other cases, the use of an absolute potential may be more natural, such as when applying the relative-entropy inequality constraints on categorical indicators (Eq. 2.7). In Chapter 6, we will also see that it is possible to define a meaningful absolute potential when the empirical distribution $\tilde{\pi}$ is not available, and we only have access to a sample from the biased

distribution. To capture this generality, we formulate generalized maxent without reference to the empirical distribution, using only the absolute potential.

2.5.3 Maxent Duality

We know from Eq. (2.19) that the generalized maxent primal and dual have equal *values*. In this section, we show the equivalence of the primal and dual *optimizers*. Specifically, we show that the maxent primal (2.17) is solved by the Gibbs distribution whose parameter vector $\boldsymbol{\lambda}$ solves the dual (possibly in a limit). This parallels the result of Della Pietra, Della Pietra, and Lafferty (1997) for basic maxent and gives additional motivation for the view of the dual objective as the regularized log loss.

Theorem 2.4 (Maxent Duality)**.** *Let q_0, U, P, Q be as above. Then*

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

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

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

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

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

Proof. Eq. (i) is a consequence of Fenchel's duality as was shown earlier. It remains to prove Eq. (ii). We will use an alternative expression for the dual objective. Let r be an arbitrary distribution. Combining Eqs. (2.4) and (2.21) yields

$$Q(\boldsymbol{\lambda}) = -\mathrm{D}(r \parallel q_{\boldsymbol{\lambda}}) + \mathrm{D}(r \parallel q_0) - \mathrm{U}_r^*(\boldsymbol{\lambda}) \ . \tag{2.23}$$

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

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

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

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

The left-hand side is non-negative by Fenchel's inequality, so $\mathrm{D}(\hat{p} \parallel q_t) \to 0$ by the

non-negativity of relative entropy. Therefore, by property (B2), every convergent subsequence of $q_1, q_2, \ldots$ has the limit $\hat{p}$. Since the q_t's come from the compact set Δ, we obtain $q_t \to \hat{p}$. ■

Thus, in order to solve the primal, it suffices to find a sequence of $\boldsymbol{\lambda}$'s maximizing the dual. This will be the goal of algorithms in Sections 4.1 and 4.2.

Chapter 3

Statistical Guarantees

All the justifications of maxent introduced in Chapter 1 view the set of constraints as part of the input specification. In this chapter, we explore how the choice of constraints influences the quality of the resulting models.

In Section 2.3, we saw several examples of "reasonable" constraint sets, motivated from different perspectives. A common motivation is to choose constraints that are likely to be satisfied by the true distribution. For example, in basic maxent, empirical averages are assumed to be good estimates of true expectations. This assumption is further refined by introducing inequality constraints, which allow for uncertainty in empirical estimates, based on the observation that true expectations should not be expected to match the empirical averages exactly, but only within some error bounds. Similarly, in discounting, the crucial observation is that the positive empirical counts of rare events overestimate the true probabilities, and thus they need to be scaled down. By introducing inequalities and discounting, we follow the intuition that the constraints should reflect our beliefs about the true distribution rather than simply summarize the empirical data.

This approach can be motivated by the max-min likelihood interpretation of maxent. According to the max-min likelihood interpretation, maxent optimizes the worst-case performance on distributions satisfying the constraints. If the constraints are too restrictive, we may miss the unknown true distribution, and nothing can be said about the performance of maxent. If the constraints are too weak then maxent is too conservative, optimizing against many unlikely distributions. Is it possible to say anything more specific about which constraints should yield better performance?

In this chapter, we develop theory addressing this very question. Specifically, we develop a quantitative understanding of how various choices of constraints influence the performance of maxent. As a result, we are able to derive novel instances of generalized maxent with favorable theoretical performance. From the dual perspective, our guarantees provide a principled method of choosing hyperparameters in various

regularization functions.

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

In linear models, ℓ_2^2, ℓ_1, and $\ell_1 + \ell_2^2$ regularization have been used under the names *ridge regression* (Hoerl and Kennard, 1970), *lasso regression* (Tibshirani, 1996), and *elastic nets* (Zou and Hastie, 2005), respectively. Lasso regression, in particular, has generated a lot of interest in recent statistical theory and practice. A frequently mentioned benefit of the lasso is its bias toward sparse solutions. The same bias is also present in ℓ_1-regularized maxent, but it is not our focus. We are interested in deriving performance guarantees. Similar guarantees are derived by Donoho and Johnstone (1994) for linear models with the lasso penalty. In a recent study, van de Geer (2006) derives non-asymptotic performance guarantees for a wide range of loss functions with ℓ_1 regularization, including log loss analyzed here. Van de Geer's results, qualitatively similar to ours, are derived independently of our work on ℓ_1-regularized maxent (Dudík et al., 2004). The relationship between the lasso approximation and the sparsest approximation in linear models is explored, for example, by Donoho and Elad (2003). In online learning literature, density estimation in exponential families is explored by Azoury and Warmuth (2001). Although our results resemble the "regret" bounds common in online learning, our analysis departs from the online setup, and exploits (in fact, relies on) statistical properties of the data.

3.1 Generalization Lemma

We start by deriving a lemma on which we base all of our generalization guarantees. This lemma will be referred to as the *generalization lemma*.

As a warm-up, consider maxent with box constraints. Its solution optimizes the regularized empirical log loss $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$. On the other hand, our goal is to do well on the true distribution π, to optimize the true log loss L_π. We express the true log loss in terms of the empirical log loss using Eq. (2.3):

$$\begin{aligned}
\mathrm{L}_\pi(\boldsymbol{\lambda}) &= \ln Z_{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \cdot \mathbf{E}_\pi[\boldsymbol{f}] \\
&= \ln Z_{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \cdot \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] + \boldsymbol{\lambda} \cdot (\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}]) \\
&= \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot (\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}]) \ . \qquad (3.1)
\end{aligned}$$

Unfortunately, $\mathbf{E}_\pi[\boldsymbol{f}]$ in Eq. (3.1) is unknown. However, assuming that $|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_\pi[f_j]| \leq \beta_j$, the inner product in Eq. (3.1) can be bounded as

$$\boldsymbol{\lambda} \cdot (\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}]) \leq \sum_{j \in \mathcal{J}} \beta_j |\lambda_j| \ .$$

Plugging in Eq. (3.1) yields

$$\mathrm{L}_\pi(\boldsymbol{\lambda}) \leq \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_{j \in \mathcal{J}} \beta_j |\lambda_j| \ . \qquad (3.2)$$

Thus, in this instance, the regularized empirical log loss is an upper bound on the true log loss. Therefore, by minimizing the regularized log loss we also minimize the guarantee on the generalization performance.

It is not a coincidence that the dual objective can be used to bound the true log loss. In general, when the potential is an arbitrary convex function, the inner product in Eq. (3.1) can be bounded by Fenchel's inequality

$$\boldsymbol{\lambda} \cdot (\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}]) \leq \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}]) \ , \qquad (3.3)$$

yielding a bound on $\mathrm{L}_\pi(\boldsymbol{\lambda})$:

$$\begin{aligned}
\mathrm{L}_\pi(\boldsymbol{\lambda}) &\leq \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}]) \\
&= \mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}]) \ . \qquad (3.4)
\end{aligned}$$

Thus, the dual objective combined with the potential of the true feature expectations bounds the true log loss. For the box potential, the term $\mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}])$ equals zero whenever the true expectation of each f_j differs from its empirical average by at most β_j,

yielding the same upper bound as in Eq. (3.2).

Since the potential of the true expectations is independent of $\boldsymbol{\lambda}$, the dual solution always optimizes an upper bound on the performance according to Eq. (3.4). Yet, the dual solution could be a poor approximation of π if this upper bound is too weak. A separate concern is that the choice of the feature set was poor and no Gibbs distribution provides a good model of π. We do not address the latter concern and compare the maxent density with the best Gibbs distribution. This is the tightest comparison we can hope for since maxent densities are constrained to take the form of Gibbs distributions.

Let $\hat{\boldsymbol{\lambda}}$ denote the solution of the dual Q.[1] In the generalization lemma below, we bound the difference between the log loss $\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})$ of $q_{\hat{\boldsymbol{\lambda}}}$ on the true distribution, and the log loss $\mathrm{L}_\pi(\boldsymbol{\lambda}^\star)$ of an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$ on π. In particular, the bound holds for the Gibbs distribution minimizing the true log loss L_π.

Lemma 3.1 (Generalization Lemma)**.** *Let $\hat{\boldsymbol{\lambda}}$ maximize Q. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$*

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

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

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

Proof. According to Eq. (2.21),

$$Q(\boldsymbol{\lambda}) = -\mathrm{L}_\pi(\boldsymbol{\lambda}) - \mathrm{U}^*_\pi(\boldsymbol{\lambda}) \ .$$

Since $\hat{\boldsymbol{\lambda}}$ maximizes $Q(\boldsymbol{\lambda})$, it also minimizes $\mathrm{L}_\pi(\boldsymbol{\lambda}) + \mathrm{U}^*_\pi(\boldsymbol{\lambda})$. Therefore, for an arbitrary $\boldsymbol{\lambda}^\star$,

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

Hence,

$$\begin{aligned} \mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{U}^*_\pi(\boldsymbol{\lambda}^\star) - \mathrm{U}^*_\pi(\hat{\boldsymbol{\lambda}}) \\ &= \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{U}^*(-\boldsymbol{\lambda}^\star) - \mathrm{U}^*(-\hat{\boldsymbol{\lambda}}) + (\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}) \cdot \mathbf{E}_\pi[\boldsymbol{f}] \ , \end{aligned} \tag{3.5}$$

where the last equality follows by shifting, i.e., Eq. (2.20). Now, similar to Eq. (3.3),

[1] We assume that the supremum of Q is attained at a finite $\hat{\boldsymbol{\lambda}}$, but the results generalize to cases when the supremum is attained only in a limit.

we bound the inner product on the right-hand side of Eq. (3.5) by Fenchel's inequality:

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

Plugging in Eq. (3.5) yields part (i) of the lemma. Part (ii) is obtained from part (i) by shifting. Similarly, part (iii) follows directly from Eq. (3.5) by shifting. ■

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

According to the Generalization Lemma(i,ii), the gap in performance between $q_{\hat{\boldsymbol{\lambda}}}$ and $q_{\boldsymbol{\lambda}^\star}$ depends on the potential of true feature expectations and on the (symmetrized) regularization of $\boldsymbol{\lambda}^\star$. This bound gives a concrete foundation to the view that the true distribution should satisfy the constraints. It suggests that we should ensure the low potential of true feature expectations. To obtain specific guarantees, we choose U to optimize the trade-off between the potential $\mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}])$ and the symmetrized regularization $\mathrm{U}^*(\boldsymbol{\lambda}^\star) + \mathrm{U}^*(-\boldsymbol{\lambda}^\star)$.

The guarantee of the Generalization Lemma(iii) is tighter than parts (i) and (ii), but it is more difficult to interpret, because of its dependence on $\hat{\boldsymbol{\lambda}}$, which is itself a random variable. To obtain interpretable bounds from part (iii), it is necessary to bound the deviation of $\hat{\boldsymbol{\lambda}}$ from the optimal $\boldsymbol{\lambda}^\star$ explicitly.

We now apply the Generalization Lemma to some specific cases of interest.

3.2 Indicator Potentials

First, we discuss the case which closely corresponds to the notion of potential as a constraint set. This is the case when U is an indicator of a closed convex set C, such as $\mathrm{U}^{(0)}$ and $\mathrm{U}^{(1)}$. The right-hand side of the Generalization Lemma(i) is then infinite unless $\mathbf{E}_\pi[\boldsymbol{f}]$ lies in C. To apply the Generalization Lemma(i), we ensure that $\mathbf{E}_\pi[\boldsymbol{f}] \in C$ with high probability. Therefore, we choose C as a confidence region for $\mathbf{E}_\pi[\boldsymbol{f}]$. If $\mathbf{E}_\pi[\boldsymbol{f}] \in C$ then for any Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}_C^*(\boldsymbol{\lambda}^\star) + \mathrm{I}_C^*(-\boldsymbol{\lambda}^\star) \ . \tag{3.6}$$

The expression $\mathrm{I}_C^*(\boldsymbol{\lambda}^\star) + \mathrm{I}_C^*(-\boldsymbol{\lambda}^\star)$ is by Eq. (2.10) equal to

$$\sup_{\boldsymbol{u} \in C}\left[\boldsymbol{\lambda}^\star \cdot \boldsymbol{u}\right] + \sup_{\boldsymbol{u} \in C}\left[-\boldsymbol{\lambda}^\star \cdot \boldsymbol{u}\right] = \sup_{\boldsymbol{u} \in C}\left[\boldsymbol{\lambda}^\star \cdot \boldsymbol{u}\right] - \inf_{\boldsymbol{u} \in C}\left[\boldsymbol{\lambda}^\star \cdot \boldsymbol{u}\right] \ .$$

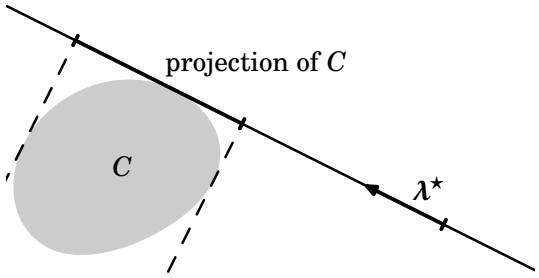


Figure 3.1. *Performance of indicator potentials.* The solution subject to constraints defining a confidence region C lags behind the optimal Gibbs distributions $q_{\boldsymbol{\lambda}^\star}$ by at most $\mathrm{I}^*_C(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_C(-\boldsymbol{\lambda}^\star)$. This amount is proportional to the projection of C onto a line parallel with $\boldsymbol{\lambda}^\star$. Thus, smaller confidence regions yield better performance guarantees.

In other words, $\mathrm{I}^*_C(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_C(-\boldsymbol{\lambda}^\star)$ is equal to the largest difference within the set of scalar products of $\boldsymbol{\lambda}^\star$ with points in C. For a fixed $\boldsymbol{\lambda}^\star$, this difference is proportional to the size of C's projection onto a line parallel with $\boldsymbol{\lambda}^\star$ (see Fig. 3.1). Thus, smaller confidence regions yield shorter projections, which in turn yield better performance guarantees.

A common method of obtaining confidence regions is to bound the difference between empirical averages and true expectations (see Appendix A). Before moving to specific examples, we state a general result for convex regions centered at empirical averages.

Theorem 3.2. *Assume that* $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}] \in \beta C_0$ *where* C_0 *is a closed convex set symmetric around the origin,* $\beta > 0$*, and* βC_0 *denotes* $\{\beta \boldsymbol{u} : \boldsymbol{u} \in C_0\}$*. Let* $\hat{\boldsymbol{\lambda}}$ *minimize* $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta \mathrm{I}^*_{C_0}(\boldsymbol{\lambda})$*. Then for an arbitrary Gibbs distribution* $q_{\boldsymbol{\lambda}^\star}$

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

Proof. Set $\mathrm{U}_{\tilde{\pi}}(\boldsymbol{u}) = \mathrm{I}_{\beta C_0}(\boldsymbol{u})$. By assumption $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}] \in \beta C_0$, and hence by the Generalization Lemma(ii)

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

Furthermore

$$\mathrm{I}^*_{\beta C_0}(\boldsymbol{\lambda}) = \sup_{\boldsymbol{u} \in \beta C_0} \boldsymbol{\lambda} \cdot \boldsymbol{u} = \sup_{\boldsymbol{u}' \in C_0} \boldsymbol{\lambda} \cdot \beta \boldsymbol{u}' = \beta \sup_{\boldsymbol{u}' \in C_0} \boldsymbol{\lambda} \cdot \boldsymbol{u}' = \beta \mathrm{I}^*_{C_0}(\boldsymbol{\lambda}) \ .$$

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

3.2.1 Maxent with ℓ_1 Regularization

Our first set of statistical guarantees concerns the box potential introduced in the previous sections. Recall that the box potential and the corresponding regularization are

$$\mathrm{U}_{\tilde{\pi}}^{(1)}(\boldsymbol{u}) = \mathrm{I}\big(|u_j| \le \beta_j \text{ for all } j\big) \ , \qquad \mathrm{U}_{\tilde{\pi}}^{(1)*}(\boldsymbol{\lambda}) = \sum_{j \in \mathcal{J}} \beta_j |\lambda_j| \ .$$

We will not be able to use Theorem 3.2 as is, because of the feature-dependent scaling β_j in $\mathrm{U}_{\tilde{\pi}}^{(1)}$. However, it is straightforward to verify that Theorem 3.2 can be modified as follows: If $|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_{\pi}[f_j]| \le \beta_j$ for all $j \in \mathcal{J}$ and $\hat{\boldsymbol{\lambda}}$ minimizes $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$ then

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_{\pi}(\boldsymbol{\lambda}^\star) + 2 \sum_{j \in \mathcal{J}} \beta_j |\lambda_j^\star|. \tag{3.7}$$

Thus, to bound the true loss $\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}})$ by Theorem 3.2, we need to find bounds β_j on $|\mathbf{E}_{\pi}[f_j] - \mathbf{E}_{\tilde{\pi}}[f_j]|$. For a finite set of bounded features, we can prove the following:

Theorem 3.3. *Assume that features f_j are bounded in $[0,1]$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$ with $\beta_j = \beta = \sqrt{\ln(2|\mathcal{J}|/\delta)/(2m)}$ for all j. Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,*

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

Proof. By Hoeffding's inequality (Theorem A.1), for a fixed j, the probability that $|\mathbf{E}_{\pi}[f_j] - \mathbf{E}_{\tilde{\pi}}[f_j]|$ exceeds β is at most $2e^{-2\beta^2 m} = \delta/|\mathcal{J}|$. By the union bound, the probability of this happening for any j is at most δ. The theorem now follows by Eq. (3.7). ■

Theorem 3.3 shows that the difference in performance between the distribution computed by minimizing ℓ_1-regularized log loss and the best Gibbs distribution becomes small rapidly as the number of samples m increases. Note that this difference depends only moderately on the number of features. Specifically, fix $\|\boldsymbol{\lambda}^\star\|_1$ and let the number of samples m grow to infinity. If at the same time the logarithm of the number of features grows slower than m, then the gap between the maxent solution and the best Gibbs distribution goes to zero. This is the case even as the best Gibbs distribution gradually improves due to more and more expressive feature sets.

The error bounds β_j in Theorem 3.3 are somewhat coarse, since they are identical for all the features. In practice, some features are more reliable than others, and the use of the reliability information should improve the estimates.

The next result improves error estimates β_j by incorporating feature specific in-

formation. As a starting point we use the approximation

$$\left|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_{\pi}[f_j]\right| = O\left(\sqrt{\mathbf{V}_{\pi}[f_j]/m}\right) \ , \tag{3.8}$$

where $\mathbf{V}_{\pi}[f_j]$ is the variance of f_j under π. Eq. (3.8) holds, by the Central Limit Theorem, with probability $1-\delta$ for any fixed δ. (For a statement of the Central Limit Theorem, see for example Lehmann and Casella, 1998, Theorem 8.9.) In order to turn Eq. (3.8) into the β_j settings with provable performance guarantees, we need to address two issues. First, the feature variances are not known, so they need to be estimated. We will use upper bounds obtained by McDiarmid's inequality (see Appendix A) for the empirical estimates of the variances

$$\mathbf{V}'_{\tilde{\pi}}[f_j] = \frac{m}{m-1}\mathbf{V}_{\tilde{\pi}}[f_j] = \frac{\sum_i \left(f_j(x_i) - \mathbf{E}_{\tilde{\pi}}[f_j]\right)^2}{m-1} = \frac{m\left(\mathbf{E}_{\tilde{\pi}}[f_j^2] - \mathbf{E}_{\tilde{\pi}}[f_j]^2\right)}{m-1} \ . \tag{3.9}$$

Second, Eq. (3.8) is an asymptotic statement with an unknown multiplicative constant. To obtain non-asymptotic bounds on $|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_{\pi}[f_j]|$ we will use Bernstein's inequality (see Appendix A).

We believe that the resulting settings of the β_j's are in practice more useful than the settings of Theorem 3.3 because they differentiate between features depending on the empirical-error estimates computed from the sample data. Motivated by these settings, in Chapter 5 we describe experiments that use $\beta_j = \beta_0\sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j]/m}$, where β_0 is a single tuning constant. This approach is equivalent to a common practice in statistics when the features are scaled to unit sample variances, resulting in transformed features $f_j^\circ(x) = f_j(x) \,/\, \sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j]}$ and a single feature-independent regularization parameter $\beta_j^\circ = \beta^\circ$. Theorem 3.4 below justifies this practice and also suggests replacing the sample variance by a slightly larger value $\mathbf{V}'_{\tilde{\pi}}[f_j] + O(1/\sqrt{m})$.

Theorem 3.4. *Assume that features f_j are bounded in $[0,1]$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j|\lambda_j|$ with*

$$\beta_j = \sqrt{\frac{2\ln(4|\mathcal{J}|/\delta)}{m}} \cdot \sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j] + \sqrt{\frac{\ln(2|\mathcal{J}|/\delta)}{2m}} + \frac{\ln(4|\mathcal{J}|/\delta)}{18m}} + \frac{\ln(4|\mathcal{J}|/\delta)}{3m} \ .$$

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

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_{\pi}(\boldsymbol{\lambda}^\star) + 2\sum_{j\in\mathcal{J}} \beta_j|\lambda_j^\star| \ .$$

Proof. Let

$$\beta'_j = \sqrt{\frac{\ln(4|\mathcal{J}|/\delta)}{3m}} \cdot \sqrt{6\mathbf{V}_\pi[f_j] + \frac{\ln(4|\mathcal{J}|/\delta)}{3m}} + \frac{\ln(4|\mathcal{J}|/\delta)}{3m} \quad .$$

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

$$\left|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_\pi[f_j]\right| \leq \beta'_j \leq \beta_j$$

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

Consider a fixed j and let $\varepsilon = \ln(4|\mathcal{J}|/\delta)/3m$. Thus,

$$\begin{aligned}
\beta'_j &= \sqrt{\varepsilon}\left(\sqrt{6\mathbf{V}_\pi[f_j] + \varepsilon} + \sqrt{\varepsilon}\right) \\
\beta_j &= \sqrt{6\varepsilon}\sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j] + \sqrt{\frac{\ln(2|\mathcal{J}|/\delta)}{2m}} + \frac{\varepsilon}{6}} + \varepsilon \\
&= \sqrt{\varepsilon}\left(\sqrt{6\left[\mathbf{V}'_{\tilde{\pi}}[f_j] + \sqrt{\ln(2|\mathcal{J}|/\delta)/(2m)}\right] + \varepsilon} + \sqrt{\varepsilon}\right) \quad .
\end{aligned}$$

By Bernstein's inequality (Theorem A.2)

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

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

$$\mathbf{V}_\pi[f_j] \geq \mathbf{V}'_{\tilde{\pi}}[f_j] + \sqrt{\frac{\ln(2|\mathcal{J}|/\delta)}{2m}} \quad .$$

We will use McDiarmid's inequality (Theorem A.3) for the function

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

Note that $\mathbf{V}'_{\tilde{\pi}}[f_j] = s\big(f_j(x_1), f_j(x_2), \ldots, f_j(x_m)\big)$ and $\mathbf{E}\big[\mathbf{V}'_{\tilde{\pi}}[f_j]\big] = \mathbf{V}_\pi[f_j]$. To apply McDiarmid's inequality, we need to bound

$$\sup_{y_1,\ldots,y_m,y'_i \in [0,1]} \left|s(y_1,\ldots,y_m) - s(y_1,\ldots,y_{i-1},y'_i,y_{i+1},\ldots,y_m)\right| \tag{3.10}$$

for every i. By symmetry, it suffices to consider a single index i. Fix i, use y to denote y_i, and y' to denote y'_i. Let $S_{m-1} = y_1 + \cdots + y_{i-1} + y_{i+1} + \cdots + y_m$. Then the difference inside the absolute value of Eq. (3.10) is

$$\begin{aligned}
&s(y_1,\ldots,y_{i-1},y,y_{i+1},\ldots,y_m) - s(y_1,\ldots,y_{i-1},y',y_{i+1},\ldots,y_m) \\
&\quad = \frac{y^2 - y'^2}{m-1} - \frac{(S_{m-1}+y)^2 - (S_{m-1}+y')^2}{m(m-1)} \\
&\quad = \frac{my^2 - my'^2 - S_{m-1}^2 - 2S_{m-1}y - y^2 + S_{m-1}^2 + 2S_{m-1}y' + y'^2}{m(m-1)} \\
&\quad = \frac{1}{m}\left[y^2 - y'^2 - 2\frac{S_{m-1}}{m-1}y + 2\frac{S_{m-1}}{m-1}y'\right] \\
&\quad = \frac{1}{m}\left[\left(y - \frac{S_{m-1}}{m-1}\right)^2 - \left(y' - \frac{S_{m-1}}{m-1}\right)^2\right] \quad . \qquad (3.11)
\end{aligned}$$

Note that the value inside the brackets of Eq. (3.11) is bounded in $[-1,1]$ because $y, y' \in [0,1]$ and $S_{m-1} \in [0, m-1]$. Plugging in Eq. (3.10) yields, for every i,

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

Thus, by McDiarmid's inequality,

$$\begin{aligned}
\mathbf{P}\left(\mathbf{V}_\pi[f_j] \geq \mathbf{V}'_{\tilde{\pi}}[f_j] + \sqrt{\frac{\ln(2|\mathcal{J}|/\delta)}{2m}}\right) &\leq \exp\left\{\frac{-2\cdot\left[\ln(2|\mathcal{J}|/\delta)/2m\right]}{m\cdot(1/m)^2}\right\} \\
&= \exp\left\{-\ln(2|\mathcal{J}|/\delta)\right\} = \delta/(2|\mathcal{J}|) \quad .
\end{aligned}$$

Hence, $\beta'_j \geq \beta_j$ with probability at most $\delta/(2|\mathcal{J}|)$, completing the proof. ■

An often cited characteristic of ℓ_1 regularization is that it induces sparsity (Tibshirani, 1996). We mention one particular aspect of sparsity which is easy to check for ℓ_1 regularization. We say that a solution $\hat{\boldsymbol{\lambda}}$ of an optimization problem is *robustly sparse* if all of its zero-valued components remain zero under perturbations of parameters. The definition of robust sparsity states that the components of $\hat{\boldsymbol{\lambda}}$ are never zero just by a lucky coincidence (in the choice of parameters). To see how ℓ_1 regularization induces this property, notice that its partial derivatives are discontinuous at $\lambda_j = 0$. As a consequence, if the regularized log loss is uniquely minimized at a point where the j_0-th component $\hat{\lambda}_{j_0}$ equals zero, then the optimal $\hat{\lambda}_{j_0}$ will remain zero even if the parameters β_j and the expectations $\mathbf{E}_{\tilde{\pi}}[f_j]$ are slightly perturbed.

So far we have considered features bounded in $[0,1]$. The results of this section, however, easily generalize to feature classes that are bounded in arbitrary finite intervals. Note that features are bounded whenever the sample space is finite. To

obtain guarantees, it suffices to scale the β_j's by the individual feature ranges and use the previous results. We refer to the size of the range as the *diameter* and use the notation $D(f_j)$. Specifically, the diameter of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$D(f) = \sup_{x,x'} \left| f(x) - f(x') \right| \ .$$

For the sake of completeness, we state versions of Theorems 3.3 and 3.4 which include the dependence on feature diameters.

Theorem 3.5. *Assume that features f_j are of bounded diameters. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$ with either of the following settings:*

$$\beta_j^{\text{Hoeffding}} = D(f_j) \sqrt{\frac{\ln(2|\mathcal{J}|/\delta)}{2m}} \tag{i}$$

$$\beta_j^{\text{Bernstein}} = D(f_j) \sqrt{\frac{\ln(4|\mathcal{J}|/\delta)}{2m}} \cdot \sqrt{\frac{4\mathbf{V}'_{\tilde{\pi}}[f_j]}{D(f_j)^2} + \sqrt{\frac{8\ln(2|\mathcal{J}|/\delta)}{m}} + \frac{2\ln(4|\mathcal{J}|/\delta)}{9m}} + D(f_j) \frac{\ln(4|\mathcal{J}|/\delta)}{3m} \ . \tag{ii}$$

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

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2 \sum_{j \in \mathcal{J}} \beta_j |\lambda_j^\star| \ .$$

The results of this section bound the performance of ℓ_1-regularized maxent in terms of feature diameters, feature variances, the logarithm of the feature-set size, and the norm $\|\boldsymbol{\lambda}\|_1$. Feature diameters, feature variances, and the logarithm of the feature-set size can be viewed as measures of the feature complexity, whereas the norm $\|\boldsymbol{\lambda}\|_1$ is a measure of the Gibbs-distribution complexity. In the next sections, we derive alternative complexity measures.

3.2.2 Maxent with Polyhedral Regularization

In this section, we consider potentials which are indicator functions of polytopes. The simplest case is the box indicator $\mathrm{U}^{(1)}$, explored in Section 3.2.1. However, when additional knowledge about the structure of the feature space is available, we show that other polytopes yield tighter confidence regions and hence better performance guarantees.

Specifically, when values of $\boldsymbol{f}(x)$ lie inside a polytope with a possibly very large number of facets then a symmetrized version of this polytope can be used as a prototype for the confidence region. For example, suppose that values $\boldsymbol{f}(x)$ lie inside the

polytope[2] $\{\boldsymbol{u} \in \mathbb{R}^\partial : a_k \le \boldsymbol{\eta}_k \cdot \boldsymbol{u} \le b_k \text{ for all } k \in \mathcal{K}\}$ where $\boldsymbol{\eta}_k \in \mathbb{R}^\partial, a_k \in \mathbb{R}, b_k \in \mathbb{R}$. Then the following holds:

Theorem 3.6. *Let $\boldsymbol{\eta}_k, a_k, b_k$ be as above. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta \mathrm{I}^*_{C_0}(\boldsymbol{\lambda})$ with $\beta = \sqrt{\ln(2|\mathcal{K}|/\delta)/(2m)}$ and $C_0 = \{\boldsymbol{u} : |\boldsymbol{\eta}_k \cdot \boldsymbol{u}| \le b_k - a_k \text{ for all } k\}$. Then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_{C_0}(\boldsymbol{\lambda}^\star) \sqrt{\frac{2\ln(2|\mathcal{K}|/\delta)}{m}} \ .$$

Proof. By Hoeffding's inequality, for a fixed k, the probability that $|\boldsymbol{\eta}_k \cdot (\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}])|$ exceeds $\beta(b_k - a_k)$ is at most $2e^{-2\beta^2 m} = \delta/|\mathcal{K}|$. By the union bound, the probability of this happening for any k is at most δ. Thus, $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}] \in \beta C_0$ with probability at least $1-\delta$ and the claim follows from Theorem 3.2. ■

Remark. Instead of applying Hoeffding's inequality, it is possible to incorporate information about variances of random variables $\boldsymbol{\eta}_k \cdot \boldsymbol{f}$ and apply Bernstein's inequality, similar to Theorem 3.4.

The performance bound of Theorem 3.6 decreases as $1/\sqrt{m}$ with an increasing number of samples and grows only logarithmically with the number of facets of the bounding polytope. Thus, bounding polytopes can have a very large number of facets and still yield good bounds for moderate sample sizes. When deciding between several polytopes based on this bound, the increase in the number of facets should be weighed against the decrease in the regularization $\mathrm{I}^*_{C_0}$ as we demonstrate in the following examples.

Linear and Quadratic Features

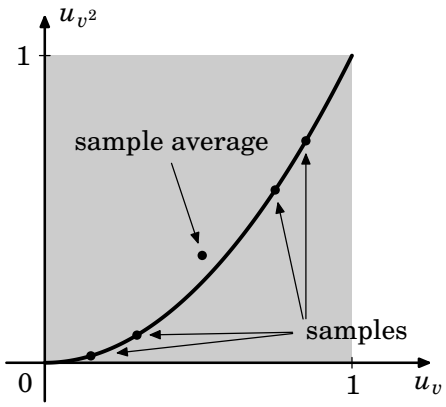
As a specific application, consider linear and quadratic features derived from variables $\mathcal{V}$. For simplicity assume that the variables are scaled to take values in $[0,1]$. Thus, both $v(x) \in [0,1]$ and $v^2(x) \in [0,1]$ for all $v \in \mathcal{V}$. Box constraints yield the guarantee

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \|\boldsymbol{\lambda}^\star\|_1 \sqrt{\frac{2\ln(4|\mathcal{V}|/\delta)}{m}} \ .$$

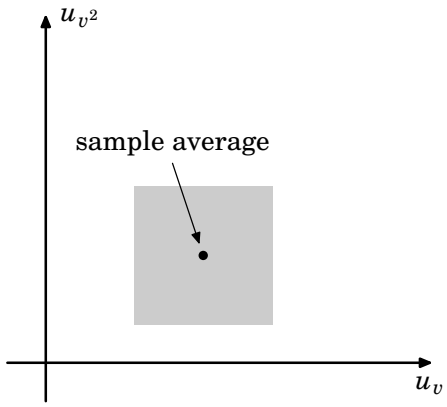
The bounding polytope corresponding to box constraints is depicted in Fig. 3.2(a). It is derived from the bounding inequalities

$$0 \le v \le 1 \ , \quad 0 \le v^2 \le 1 \ ,$$

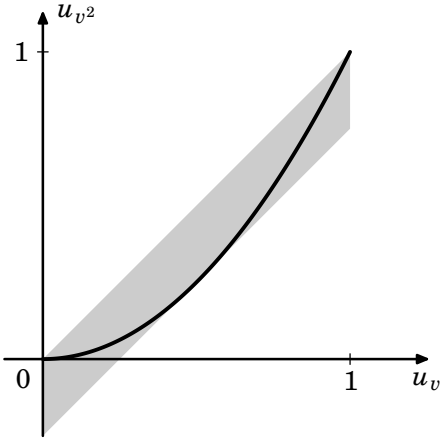
[2] For technical reasons, we represent polytopes as intersections of bands rather than intersections of halfspaces.



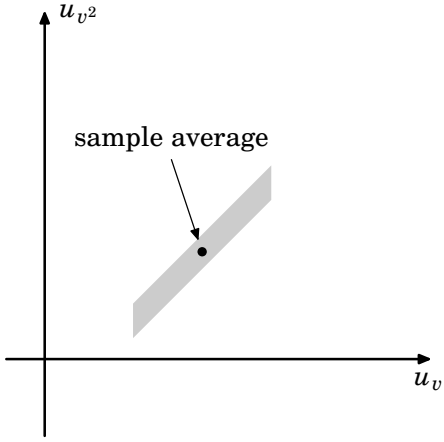
(a) Bounding box.



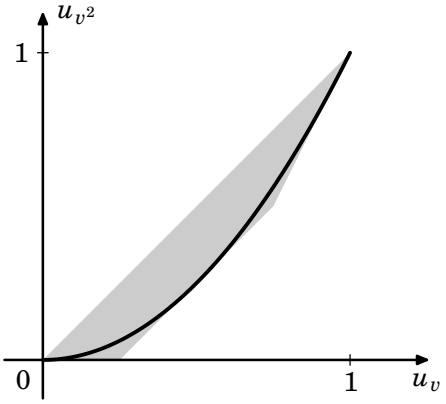
(b) Box-shaped confidence region.



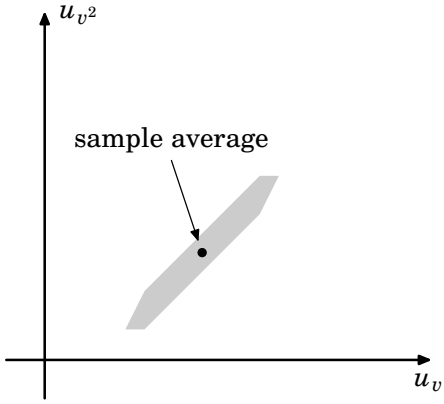
(c) Bounding polytope.



(d) Confidence region derived from the polytope (c).



(e) Bounding polytope.



(f) Confidence region derived from the polytope (e).

Figure 3.2. *Examples of indicator potentials for linear and quadratic features.*

which yield the prototype polytope

$$C_0 = \{\boldsymbol{u} : |u_v| \le 1,\ |u_{v^2}| \le 1 \text{ for all } v\} \ . \qquad \text{(see Fig. 3.2b)}$$

Noting that the pairs $(v(x), v^2(x))$ lie only on a thin sliver inside the box $[0,1]\times[0,1]$, we can instead consider tighter bounding inequalities

$$0 \le v \le 1 \ , \quad -\frac{1}{4} \le v^2 - v \le 0 \ , \qquad \text{(see Fig. 3.2c)}$$

yielding

$$C_0' = \left\{\boldsymbol{u} : \left|u_v\right| \le 1,\ \left|u_{v^2} - u_v\right| \le \frac{1}{4} \text{ for all } v\right\} \ , \qquad \text{(see Fig. 3.2d)}$$

and the guarantee

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_{C_0'}(\boldsymbol{\lambda}^\star)\sqrt{\frac{2\ln(4|\mathcal{V}|/\delta)}{m}} \ .$$

In this case, it is possible to derive $\mathrm{I}^*_{C_0'}$ explicitly by Eq. (2.12). For a single variable v, the polytope C_0' defined on components u_v and u_{v^2} has vertices $(-1,-5/4)$, $(-1,-3/4)$, $(1,5/4)$, $(1,3/4)$. Thus

$$\begin{aligned}
\mathrm{I}^*_{C_0';v}(\lambda_v, \lambda_{v^2}) &= \max\left\{-\lambda_v - \frac{5}{4}\lambda_{v^2},\ -\lambda_v - \frac{3}{4}\lambda_{v^2},\ \lambda_v + \frac{5}{4}\lambda_{v^2},\ \lambda_v + \frac{3}{4}\lambda_{v^2}\right\} \\
&= \max\left\{\left|\lambda_v + \frac{5}{4}\lambda_{v^2}\right|,\ \left|\lambda_v + \frac{3}{4}\lambda_{v^2}\right|\right\} \\
&= \left|\lambda_v + \lambda_{v^2}\right| + \left|\frac{1}{4}\lambda_{v^2}\right| \ ,
\end{aligned}$$

where the last inequality follows from the identity

$$\max\{|a|, |b|\} = \frac{|a-b|}{2} + \frac{|a+b|}{2} \ .$$

Summing $\mathrm{I}^*_{C_0';v}$ across all variables v, we obtain

$$\mathrm{I}^*_{C_0'}(\boldsymbol{\lambda}) = \sum_{v\in\mathcal{V}} \left(\left|\lambda_v + \lambda_{v^2}\right| + \left|\frac{1}{4}\lambda_{v^2}\right|\right) \ .$$

Note that $\mathrm{I}^*_{C_0'}(\boldsymbol{\lambda})$ may be up to eight times smaller than $\|\boldsymbol{\lambda}\|_1$ (if $\lambda_v = -\lambda_{v^2}$ for all v) while $\mathrm{I}^*_{C_0'}(\boldsymbol{\lambda})$ is at most 1.25-times larger than $\|\boldsymbol{\lambda}\|_1$ (if $\lambda_v = 0$ for all v). Thus, compared with the box potential, the bound may decrease up to eight times, or increase 1.25 times. The introduced improvement would require a 64-fold increase in the number of training samples using ℓ_1 regularization, whereas in the worst case, we

perform as well as ℓ_1 regularization with about 1.56-times fewer samples.

Of course, it is possible to construct even tighter bounding polytopes, which lie strictly inside the box $[0,1]\times[0,1]$, at the cost of enlarging the number of constraints. For example, we may consider the bounds

$$0 \le v^2 \le 1 \ , \quad -\frac{1}{4} \le v^2 - v \le 0 \ , \quad -1 \le v^2 - 2v \le 0 \ , \qquad \text{(see Fig. 3.2e)}$$

yielding

$$C_0'' = \left\{ \boldsymbol{u} : \left|u_{v^2}\right| \le 1, \left|u_{v^2} - u_v\right| \le \frac{1}{4}, \left|u_{v^2} - 2u_v\right| \le 1 \text{ for all } v \right\} \ , \qquad \text{(see Fig. 3.2f)}$$

and the guarantee

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_{C_0''}(\boldsymbol{\lambda}^\star)\sqrt{\frac{2\ln(6|\mathcal{V}|/\delta)}{m}} \ . \tag{3.12}$$

In this case, the relative increase of the bound due to a larger size of $\mathcal{K}$ is

$$\sqrt{\frac{\ln(6|\mathcal{V}|/\delta)}{\ln(4|\mathcal{V}|/\delta)}} = \sqrt{1 + \frac{\ln 1.5}{\ln(4|\mathcal{V}|/\delta)}} \ ,$$

which is close to one for moderate sizes of $\mathcal{V}$, whereas the decrease due to a tighter confidence region may still be eightfold compared with the box potential.

The specific form of $\mathrm{I}^*_{C_0''}$ can be derived by noticing that for a single variable v, the vertices of C_0'' defined on components u_v and u_{v^2} have coordinates $\pm(1,1)$, $\pm(3/4,1)$, $\pm(3/4,1/2)$. Similar to $\mathrm{I}^*_{C_0'}$, we can then derive

$$\mathrm{I}^*_{C_0''}(\boldsymbol{\lambda}) = \sum_{v\in\mathcal{V}} \left(\left| \frac{3}{4}\lambda_{v^2} + \frac{3}{4}\lambda_v \right| + \left| \frac{1}{4}\lambda_{v^2} + \frac{1}{8}\lambda_v \right| + \left| \frac{1}{8}\lambda_v \right| \right) \ .$$

Linear, Quadratic, and Product Features

In this example, we expand the feature set to include also product features $f_{vw}(x) = v(x)w(x)$ where $v, w \in \mathcal{V}$. Instead of the single inequality

$$0 \le vw \le 1 \ ,$$

we can for example consider

$$\begin{aligned} 0 \le vw \le 1 \qquad & \qquad -\frac{1}{2} \le vw - \frac{v+w}{2} \le 0 \\ -1 \le vw - v \le 0 \qquad & \qquad -\frac{1}{2} \le vw - \frac{v^2+w^2}{2} \le 0 \\ -1 \le vw - w \le 0 \qquad & \qquad -\frac{1}{2} \le vw - \frac{v+w}{2} + \frac{v^2+w^2}{2} \le 0 \ . \\ -1 \le vw - v - w \le 0 \qquad & \end{aligned}$$

Similar to the previous example, a constant-factor increase in the number of constraints $|\mathcal{K}|$ yields only a slight relative increase in the generalization bound for a moderate number of variables. This is outweighed by the decrease of the bound due to a tighter confidence region.

3.2.3 Maxent with ℓ_2 Regularization

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

$$\mathrm{U}_{\tilde{\pi}}^{(\sqrt{2})}(\boldsymbol{u}) = \mathrm{I}\big(\|\boldsymbol{u}\|_2 \le \beta\big) \ , \qquad \mathrm{U}_{\tilde{\pi}}^{(\sqrt{2})*}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_2 \ .$$

We obtain performance guarantees using the same technique as in the previous sections: we bound the deviation $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2$ and then apply Theorem 3.2.

As the first step we bound the *expectation* of the deviation $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2$. Then we use McDiarmid's inequality to obtain a probabilistic bound on $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2$.

By Jensen's inequality,

$$\begin{aligned}\mathbf{E}\big[\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2\big] &= \mathbf{E}\left[\sqrt{\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2^2}\right] \\ &\le \sqrt{\mathbf{E}\big[\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2^2\big]} = \sqrt{\frac{\mathrm{tr}\,\boldsymbol{\Sigma}}{m}} \end{aligned} \tag{3.13}$$

where $\boldsymbol{\Sigma}$ is the feature covariance matrix with elements $\Sigma_{jj'} = \mathbf{E}_{\pi}[f_j f_{j'}] - \mathbf{E}_{\pi}[f_j]\mathbf{E}_{\pi}[f_{j'}]$. Thus, to bound the expectation of $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_2$ it suffices to bound the trace of the feature covariance matrix.

Lemma 3.7. *Let* $D_2(\boldsymbol{f}) = \sup_{x,x' \in \mathcal{X}} \|\boldsymbol{f}(x) - \boldsymbol{f}(x')\|_2$ *be the* ℓ_2 *diameter of* $\boldsymbol{f}$ *and let* $\boldsymbol{\Sigma}$ *denote the feature covariance matrix. Then* $\mathrm{tr}\,\boldsymbol{\Sigma} \le D_2(\boldsymbol{f})^2/2$.

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

$$\begin{aligned}\mathbf{E}\big[\|\boldsymbol{f} - \boldsymbol{f}'\|_2^2\big] &= \mathbf{E}[\boldsymbol{f} \cdot \boldsymbol{f}] - 2\mathbf{E}[\boldsymbol{f}] \cdot \mathbf{E}[\boldsymbol{f}'] + \mathbf{E}[\boldsymbol{f}' \cdot \boldsymbol{f}'] \\ &= 2\mathbf{E}[\boldsymbol{f} \cdot \boldsymbol{f}] - 2\mathbf{E}[\boldsymbol{f}] \cdot \mathbf{E}[\boldsymbol{f}] \\ &= 2\sum_{j \in \mathcal{J}} \Big[\mathbf{E}[f_j^2] - (\mathbf{E}[f_j])^2\Big] = 2\,\mathrm{tr}\,\boldsymbol{\Sigma} \ . \end{aligned}$$

Since $\|\boldsymbol{f} - \boldsymbol{f}'\|_2 \le D_2(\boldsymbol{f})$, we obtain $\mathrm{tr}\,\boldsymbol{\Sigma} \le D_2(\boldsymbol{f})^2/2$. ∎

Now we can use McDiarmid's inequality to prove an ℓ_2 version of Hoeffding's inequality.

Lemma 3.8. *Let $D_2(\boldsymbol{f})$ be the ℓ_2 diameter of $\boldsymbol{f}$ and let $\delta > 0$. Then with probability at least $1-\delta$*

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

Proof. Consider independent samples $X_1, \ldots, X_m$ distributed according to π and the random variable $\boldsymbol{u}(X_1, \ldots, X_m) = \sum_i (\boldsymbol{f}(X_i) - \mathbf{E}_{\pi}[\boldsymbol{f}]) = m(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}])$. We will bound $\mathbf{E}[\|\boldsymbol{u}\|_2]$ and use McDiarmid's inequality (Theorem A.3) to show that

$$\mathbf{P}\left(\|\boldsymbol{u}\|_2 - \mathbf{E}[\|\boldsymbol{u}\|_2] \geq D_2(\boldsymbol{f})\sqrt{m\ln(1/\delta)/2}\right) \leq \delta \ . \tag{3.14}$$

By Eq. (3.13) and Lemma 3.7, we obtain

$$\mathbf{E}[\|\boldsymbol{u}\|_2] \leq \sqrt{m \operatorname{tr} \boldsymbol{\Sigma}} \leq D_2(\boldsymbol{f})\sqrt{m/2} \ .$$

Now, by the triangle inequality,

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

and Eq. (3.14) follows by McDiarmid's inequality. ∎

Finally, we can derive a guarantee on the performance of ℓ_2-regularized maxent.

Theorem 3.9. *Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\|\boldsymbol{\lambda}\|_2$ with*

$$\beta = D_2(\boldsymbol{f})\left[1 + \sqrt{\ln(1/\delta)}\right]/\sqrt{2m} \ .$$

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

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

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

An analogous bound can also be obtained for ℓ_1-regularized maxent in terms of

the ℓ_∞ diameter of the feature space

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

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

When non-overlapping groups of features can be bounded separately in the ℓ_2 norm, Lemma 3.8 can be used to bound the ℓ_2-norm deviation in each specific group, and the union bound can be used to bound probability of deviation in any group. As a result, we obtain guarantees for the regularization

$$\beta_1\|\boldsymbol{\lambda}_1\|_2 + \beta_2\|\boldsymbol{\lambda}_2\|_2 \cdots + \beta_G\|\boldsymbol{\lambda}_G\|_2 \ .$$

Here, we used $\boldsymbol{\lambda}_g$, $g = 1,\ldots,G$, to denote groups of parameters that correspond to the respective groups of features. According to Lemma 3.8, we should set $\beta_g \propto D_2(\boldsymbol{f}_g)/\sqrt{m}$. When each group consists of exactly one feature, we obtain ℓ_1 regularization. In the general case, we obtain the regularization known from linear models as the *group lasso* (Yuan and Lin, 2006). According to our guarantees, we benefit from partitioning the variables into groups as long as

$$\sum_{g=1}^{G} \|\boldsymbol{\lambda}_g^\star\|_2 D_2(\boldsymbol{f}_g)\sqrt{\ln G} \le \|\boldsymbol{\lambda}^\star\|_2 D_2(\boldsymbol{f}) \ . \tag{3.15}$$

The leading $\sqrt{\ln G}$ on the left-hand side comes from the union bound across the individual groups.[3] Eq. (3.15) holds, for example, when groups are uncorrelated, and

[3] More precisely, we benefit from partitioning the variables into groups if

$$\sum_{g=1}^{G} \|\boldsymbol{\lambda}_g^\star\|_2 D_2(\boldsymbol{f}_g)\Big(\sqrt{2} + \sqrt{2\ln(G/\delta)}\Big) \le \|\boldsymbol{\lambda}^\star\|_2 D_2(\boldsymbol{f})\Big(\sqrt{2} + \sqrt{2\ln(1/\delta)}\Big) \ . \tag{3.16}$$

Assuming that the number of groups is large enough, specifically, $\ln G \ge \frac{\ln(1/\delta)}{\ln(1/\delta)-1}$, we obtain that

only a small proportion of them is relevant, as we discuss next.

Without loss of generality assume that $D_2(\boldsymbol{f}_1) = D_2(\boldsymbol{f}_2) = \cdots = D_2(\boldsymbol{f}_G)$. Lack of correlation between groups means that $D_2(\boldsymbol{f})^2 \approx \sum_g D_2(\boldsymbol{f}_g)^2$, i.e., $D_2(\boldsymbol{f}_g) = D_2(\boldsymbol{f})/\sqrt{G}$. Assume that only the groups $1,\ldots,G^\star$ are relevant, where $G^\star \leq G$, and for simplicity assume that they are equally relevant in the sense that $\|\boldsymbol{\lambda}_1^\star\|_2 = \cdots = \|\boldsymbol{\lambda}_{G^\star}^\star\|_2$ (the remaining parameters are zero). Thus, $\|\boldsymbol{\lambda}_g^\star\|_2 = \|\boldsymbol{\lambda}^\star\|_2/\sqrt{G^\star}$ for $g = 1,\ldots,G^\star$. Plugging these in the left-hand side of Eq. (3.15), we obtain

$$\begin{aligned}\sum_{g=1}^{G} \|\boldsymbol{\lambda}_g^\star\|_2 D_2(\boldsymbol{f}_g)\sqrt{\ln G} &= \frac{G^\star \|\boldsymbol{\lambda}^\star\|_2}{\sqrt{G^\star}} \cdot \frac{D_2(\boldsymbol{f})\sqrt{\ln G}}{\sqrt{G}} \\ &= \|\boldsymbol{\lambda}^\star\|_2 D_2(\boldsymbol{f}) \sqrt{\frac{G^\star \ln G}{G}} \quad .\end{aligned}$$

Thus Eq. (3.15) is satisfied if $G^\star \leq G/\ln G$, i.e., if the relevant groups form no more than a logarithmic fraction of all groups.

Consider sparsity-inducing properties of ℓ_2 regularization. Since the sole discontinuity of the derivative of the ℓ_2-norm is at zero, there are only two sparsity levels: either all coordinates of $\hat{\boldsymbol{\lambda}}$ are zero or none of them are. When groups of parameters are regularized separately, this means that either all parameters in a given group $\hat{\boldsymbol{\lambda}}_g$ are zero or none of them are. This can be viewed as a group-level version of the sparsity-inducing property of the ℓ_1 regularization, hence the name "group lasso."

Generalizations

The previous results for ℓ_2-regularized maxent immediately generalize to the cases where the values $\boldsymbol{f}(x)$ belong to an arbitrary Hilbert space $\mathcal{H}$. Parameter vectors $\boldsymbol{\lambda}$ are then taken from $\mathcal{H}$ as well, and the standard inner product and the ℓ_2 norm are replaced by their Hilbert-space equivalents. In machine learning, the most prominent examples of Hilbert spaces are reproducing kernel Hilbert spaces, used heavily in support vector machine literature (see for example Schölkopf and Smola, 2002).

A separate line of generalizations arises by replacing the ℓ_2-ball constraints by ellipsoid constraints. These are represented using a positive definite matrix $\mathbf{A}$, defining the potential and regularization

$$\mathrm{U}_{\tilde{\pi}}(\boldsymbol{u}) = \mathrm{I}\Big(\sqrt{\boldsymbol{u}^\top \mathbf{A} \boldsymbol{u}} \leq \beta\Big) \quad , \qquad \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) = \beta\sqrt{\boldsymbol{\lambda}^\top \mathbf{A}^{-1} \boldsymbol{\lambda}} \quad .$$

Eq. (3.16) follows from Eq. (3.15). Thus, Eq. (3.15) poses stronger requirements than necessary, but it simplifies the exposition.

Ellipsoid indicators can be reduced to the ℓ_2-ball indicator by the transformation

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

where $\mathbf{A}^{1/2}$ is the unique symmetric positive definite matrix such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. As a result, we obtain guarantees for ellipsoid potentials analogous to those for the ℓ_2-ball potential, with $D_2(\boldsymbol{f})$ replaced by

$$D_{\mathbf{A}}(\boldsymbol{f}) = \sup_{x,x' \in \mathcal{X}} \sqrt{\big(\boldsymbol{f}(x) - \boldsymbol{f}(x')\big)^\top \mathbf{A} \big(\boldsymbol{f}(x) - \boldsymbol{f}(x')\big)} \ .$$

The previous can be generalized even further by considering a pair of conjugate norms $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{A}^*}$, and the potential and regularization

$$\mathrm{U}_{\tilde{\pi}}(\boldsymbol{u}) = \mathrm{I}\big(\|\boldsymbol{u}\|_{\mathcal{A}} \le \beta\big) \ , \qquad \mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_{\mathcal{A}^*} \ .$$

Unfortunately, general bounds on deviations $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_{\mathcal{A}}$ are not available, so they need to be derived explicitly for specific norms. Box and polyhedral potentials of previous sections are examples of norm indicators.[4] The specific bounds were obtained by directly bounding the deviations of averages from expectations. In this section, we have explored an alternative, two-step approach. The first step was an upper-bound on the expected deviation

$$\mathbf{E}\big[\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_{\pi}[\boldsymbol{f}]\|_{\mathcal{A}}\big] \ ,$$

which was in our case derived from specific properties of the ℓ_2 norm. The second step was an application of McDiarmid's inequality (as in Lemma 3.8) and it required only the triangular inequality, which makes it applicable to arbitrary norms.

3.3 Smooth Potentials

The guarantees we have derived for indicator potentials have many favorable properties. Most notably, they provide regularization settings that achieve good performance compared with arbitrary Gibbs distributions. However, in certain situations there are computational and statistical reasons to use other types of potentials.

First, the indicator potentials may lead to multiple vectors $\hat{\boldsymbol{\lambda}}$ specifying the unique maxent distribution. When features are linearly dependent, there will be infinitely

[4] To obtain norms, we need to exclude degeneracies such as zero-width and infinite-width boxes and polytopes. More precisely, all symmetric bounded closed convex sets with non-empty interior can be viewed as norm-one balls (Rockafellar, 1970, Theorem 15.2).

many $\boldsymbol{\lambda}$'s specifying every distribution. This may be problematic if $\hat{\boldsymbol{\lambda}}$ is used for extrapolation into new sample spaces in which the features need no longer be linearly dependent. In species distribution modeling, the extrapolation is used, for example, to assess impact of changes in the environment on species distributions. The uniqueness of $\hat{\boldsymbol{\lambda}}$ can be achieved by introducing strictly convex regularization functions. Strictly convex regularizations allow us to prioritize among otherwise equal solutions. For example, we might prefer solutions that spread the parameter weights across a larger group of features, rather than rely on a single feature. Strictly convex regularization functions correspond to smooth potentials.[5] Unfortunately, indicator potentials I_C are not smooth, because they have an "edge" at the boundary of C.

The second concern is the lack of smoothness of the regularization function. While smoothness is not necessary for efficient convex optimization, many existing techniques, such as the Newton method, rely on the existence of second derivatives. The derivatives of I_C^* are discontinuous at zero, and hence the second derivatives are not defined at zero. This problem can be prevented by using smooth approximations of I_C^*.

The final reason to deviate from indicator potentials is to ensure feasibility of the maxent primal. The maxent primal is always feasible when indicators are derived from empirical distributions. However, when indicators are derived by other means, such as from feature averages sampled at a different resolution, it may be difficult to guarantee feasibility. The problem is that the indicator potential may be infinite for all possible vectors of feature expectations realizable on a given sample space. This can be prevented by introducing finite-valued potentials.

Finite-valued potentials also yield guarantees on the expected performance of maxent. For example, the right-hand side of the Generalization Lemma(i) includes the term $\mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}])$. If the potential U is derived from empirical data (such as the previously mentioned potentials, derived from $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]$), then the term $\mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}])$ is a random variable. If $\mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}])$ is infinite with a non-zero probability, then the Generalization Lemma(i) cannot be used to prove any guarantees on the expected performance. On the other hand, if U is always finite and the expectation of $\mathrm{U}(\mathbf{E}_\pi[\boldsymbol{f}])$ is finite, then we obtain guarantees on the expected performance of maxent.

In this section, we first examine a smooth approximation of ℓ_1 regularization and then turn to examples derived from ℓ_2^2 regularization.

[5] More precisely, *essentially strictly convex* regularization functions are derived from *essentially smooth* potentials. For the definition and the correspondence of essential smoothness and essential strict convexity see Rockafellar (1970), Section 26.

3.3.1 Maxent with Smoothed ℓ_1 Regularization

We analyze a smooth approximation to ℓ_1-regularization, similar to one used by Dekel et al. (2003),

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}(\boldsymbol{\lambda}) = \sum_{j \in \mathcal{J}} \alpha_j \beta_j \ln \cosh(\lambda_j/\alpha_j) = \sum_{j \in \mathcal{J}} \alpha_j \beta_j \ln\Big(\frac{e^{\lambda_j/\alpha_j} + e^{-\lambda_j/\alpha_j}}{2}\Big) \ .$$

Constants $\alpha_j > 0$ control the tightness of fit to the ℓ_1 norm while constants $\beta_j \geq 0$ control scaling (see Fig. 3.3). Note that $\cosh x \leq e^{|x|}$. Hence

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}(\boldsymbol{\lambda}) \leq \sum_{j \in \mathcal{J}} \alpha_j \beta_j \ln e^{|\lambda_j|/\alpha_j} = \sum_{j \in \mathcal{J}} \alpha_j \beta_j |\lambda_j|/\alpha_j = \sum_{j \in \mathcal{J}} \beta_j |\lambda_j| \ . \tag{3.17}$$

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

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}(\boldsymbol{u}) = \textstyle\sum_j \alpha_j \beta_j \mathrm{D}\Big(\frac{1 + u_j/\beta_j}{2} \,\Big\|\, \frac{1}{2}\Big)$$

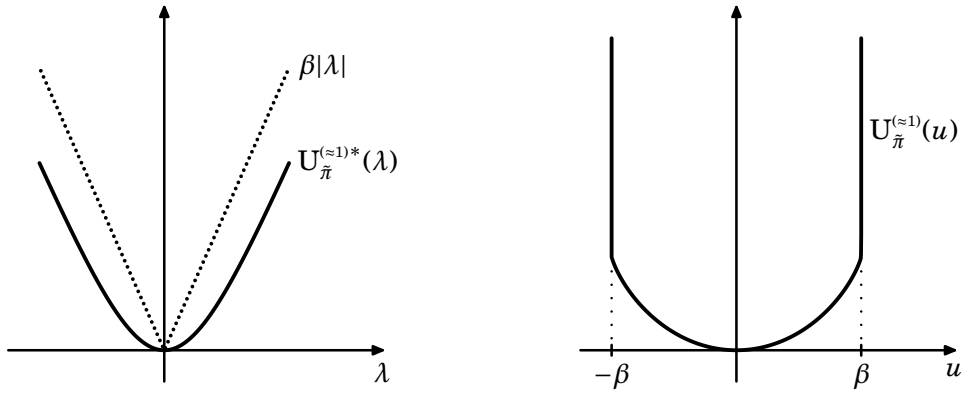
where, for $a, b \in [0,1]$, $\mathrm{D}(a \parallel b)$ is a shorthand for $\mathrm{D}((a, 1-a) \parallel (b, 1-b))$. To derive $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$, notice that $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}$ decomposes into a sum of functions of individual coordinates, so it suffices to derive a single coordinate potential $\mathrm{U}_{\tilde{\pi},j}^{(\approx 1)}$:

$$\begin{aligned}
\mathrm{U}_{\tilde{\pi},j}^{(\approx 1)}(u_j) &= \sup_{\lambda_j} \left[u_j \lambda_j - \alpha_j \beta_j \ln\left(\frac{e^{\lambda_j/\alpha_j} + e^{-\lambda_j/\alpha_j}}{2}\right)\right] \\
&= \alpha_j \beta_j \sup_{\lambda_j} \left[u_j \cdot \frac{\lambda_j}{\alpha_j \beta_j} - \ln\left(\frac{1}{2}\exp\left\{\beta_j \cdot \frac{\lambda_j}{\alpha_j \beta_j}\right\} + \frac{1}{2}\exp\left\{-\beta_j \cdot \frac{\lambda_j}{\alpha_j \beta_j}\right\}\right)\right] \\
&= \alpha_j \beta_j \sup_{\lambda'_j := \lambda_j/\alpha_j\beta_j} \left[u_j \lambda'_j - \ln\left(\frac{1}{2}e^{\beta_j \lambda'_j} + \frac{1}{2}e^{-\beta_j \lambda'_j}\right)\right] \qquad (3.18) \\
&= \alpha_j \beta_j \mathrm{D}\left(\left(\frac{1 + u_j/\beta_j}{2}, \frac{1 - u_j/\beta_j}{2}\right) \,\middle\|\, \left(\frac{1}{2}, \frac{1}{2}\right)\right) \ . \qquad (3.19)
\end{aligned}$$

Eq. (3.18) follows by a change of variables. The maximization in Eq. (3.18) takes the form of a basic-maxent dual over a two-point space, say $\mathcal{X} = \{0,1\}$, with a single feature $f(0) = \beta_j$, $f(1) = -\beta_j$, and the empirical expectation $\mathbf{E}_{\tilde{\pi}}[f] = u_j$. Thus, by maxent duality, the value of the supremum equals $\mathrm{D}(p \parallel (1/2, 1/2))$, where p comes from a closure of the set of Gibbs distributions and $\mathbf{E}_p[f] = u_j$. However, the only distribution on $\mathcal{X}$ that satisfies the expectation constraint is

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

hence Eq. (3.19) follows.



(a) Regularization $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}$. **(b)** Potential $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$.

Figure 3.3. *Smoothed ℓ_1 regularization and the corresponding potential.*

The potential $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$ can be viewed as a smooth upper bound on the box potential $\mathrm{U}_{\tilde{\pi}}^{(1)}$ in the sense that the gradient of $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$ is continuous on the interior of the effective domain of $\mathrm{U}_{\tilde{\pi}}^{(1)}$ and the norm of the gradient approaches ∞ on the border (see Fig. 3.3). Note that if $|u_j| \le \beta_j$ for all j then $\mathrm{D}\left(\frac{1+u_j/\beta_j}{2} \,\middle\|\, \frac{1}{2}\right) \le \mathrm{D}\left(0 \,\middle\|\, \frac{1}{2}\right) = \ln 2$ and hence

$$\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}(\boldsymbol{u}) \le (\ln 2) \sum_j \alpha_j \beta_j \ . \tag{3.20}$$

Applying bounds (3.17) and (3.20) in the Generalization Lemma(ii), we obtain an analog of Eq. (3.7).

Theorem 3.10. *Assume that for each j, $|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_{\pi}[f_j]| \le \beta_j$. Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}(\boldsymbol{\lambda})$. Then for an arbitrary Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$*

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

To obtain guarantees analogous to those for ℓ_1-regularized maxent, it suffices to choose sufficiently small α_j's. For example, in order to perform well relative to distributions $q_{\boldsymbol{\lambda}^\star}$ with $\sum_j \beta_j |\lambda_j^\star| \le L_1$, it suffices to set $\alpha_j = (\varepsilon L_1)/(n \beta_j \ln 2)$. Then

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_{\pi}(\boldsymbol{\lambda}^\star) + 2(1+\varepsilon) L_1 \ .$$

For example, we can derive an analog of Theorem 3.3. We relax the constraint that features are bounded in $[0,1]$ and, instead, provide a guarantee in terms of the ℓ_∞ diameter of the feature space.

Theorem 3.11. *Let $\delta, \varepsilon, L_1 > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\beta\sum_j \ln\cosh(\lambda_j/\alpha)$ with*

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

Then with probability at least $1 - \delta$

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

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

According to Theorem 3.11, the guarantees for smoothed ℓ_1 regularization converge to those for ℓ_1 regularization as $\alpha_j \to 0$. At the same time the objective becomes less smooth in some regions and less convex (more flat) in other regions. This has a negative impact on the convergence properties of many convex-optimization methods. For example, the number of iterations of gradient descent increases with increasing condition number of the Hessian. In our case, this condition number increases as $\alpha_j \to 0$. Similarly, the number of iterations of Newton's method depends on the condition number and the Lipschitz constant of the Hessian, both of which increase as $\alpha_j \to 0$. Thus, in choosing α_j, we trade an improvement in the performance guarantees for an increase in the running time.

Generalizations

Although we have worked with explicit forms of $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)}$ and $\mathrm{U}_{\tilde{\pi}}^{(\approx 1)*}$, the only properties used in the proofs were upper and lower bounds placing coordinate potentials $\mathrm{U}_{\tilde{\pi},j}^{(\approx 1)}$ between the displaced versions of $\mathrm{U}_{\tilde{\pi},j}^{(1)}$:

$$\mathrm{I}\big(|u_j| \leq \beta_j\big) \leq \mathrm{U}_{\tilde{\pi},j}^{(\approx 1)}(u_j) \leq \mathrm{I}\big(|u_j| \leq \beta_j\big) + \alpha_j \beta_j \ln 2 \ .$$

Equivalently, it is possible to consider bounds on the conjugates:

$$\beta_j |\lambda_j| - \alpha_j \beta_j \ln 2 \leq \mathrm{U}_{\tilde{\pi},j}^{(\approx 1)*}(\lambda_j) \leq \beta_j |\lambda_j| \ .$$

(The lower bound was not proved, but it is straightforward to derive from the inequality $\cosh x \geq e^{|x|}/2$.)

The previous bounds (or similar) are satisfied by a large class of smooth approxi-

mations to ℓ_1 regularization. For example, Lee et al. (2006) consider

$$\begin{aligned}
\mathrm{U}_{\tilde{\pi},j}(u_j) &= \mathrm{I}\big(|u_j| \le \beta_j\big) - \alpha_j \sqrt{\beta_j^2 - u_j^2} \\
\mathrm{U}^*_{\tilde{\pi},j}(\lambda_j) &= \beta_j \sqrt{\lambda_j^2 + \alpha_j^2} \ .
\end{aligned}$$

For this potential and regularization, it is straightforward to show that

$$\begin{aligned}
\mathrm{I}\big(|u_j| \le \beta_j\big) - \alpha_j\beta_j \ \le\ \mathrm{U}_{\tilde{\pi},j}(u_j) \ &\le\ \mathrm{I}\big(|u_j| \le \beta_j\big) \\
\beta_j|\lambda_j| \ \le\ \mathrm{U}^*_{\tilde{\pi},j}(\lambda_j) \ &\le\ \beta_j|\lambda_j| + \alpha_j\beta_j \ ,
\end{aligned}$$

which yields guarantees similar to those obtained for $\mathrm{U}^{(\approx 1)*}_{\tilde{\pi}}$.

3.3.2 Maxent with ℓ_2^2 Regularization

So far we have considered potentials that take the form of an indicator function or its smooth approximation. In this section we present results for the ℓ_2^2 potential $\mathrm{U}^{(2)}_{\tilde{\pi}}$ of Section 2.5 and the corresponding conjugate $\mathrm{U}^{(2)*}_{\tilde{\pi}}$:

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

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

The main difference from the previously considered potentials is that $\mathrm{U}^{(2)}_{\tilde{\pi}}$ is finitely-valued. As a result, the primal will always be feasible. Another consequence of the finitely-valued potential is that it is possible to derive guarantees on the expected performance (in addition to probabilistic guarantees).

In the previous sections, we obtained guarantees by optimizing tuning constants. Here, we will not be able to optimize the tuning constant uniformly across all $\boldsymbol{\lambda}^\star$. Our guarantees will require an *a priori* bound on $\|\boldsymbol{\lambda}^\star\|_2$. This is analogous to the guarantees derived by Zhang (2005) for the expected performance of conditional maxent. However, we are able to obtain a better multiplicative constant.

Note that we could derive expectation guarantees by simply applying the Generalization Lemma(ii) and taking the expectation over a random sample:

$$\begin{aligned}
\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]\|_2^2}{\alpha} + \alpha\|\boldsymbol{\lambda}^\star\|_2^2 \qquad (3.21) \\
\mathbf{E}\big[\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}})\big] &\le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\mathrm{tr}\boldsymbol{\Sigma}}{\alpha m} + \alpha\|\boldsymbol{\lambda}^\star\|_2^2 \ .
\end{aligned}$$

Here, $\boldsymbol{\Sigma}$ is the feature covariance matrix (similar to Eq. 3.13). We improve this guar-

antee by using the Generalization Lemma(iii) with $q_{\boldsymbol{\lambda}^\star}$ chosen to minimize $\mathrm{L}_\pi(\boldsymbol{\lambda}) + \mathrm{U}^{(2)*}_{\tilde{\pi}}(\boldsymbol{\lambda})$, and explicitly bounding $(\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}) \cdot (\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}])$ by a stability result similar to Zhang (2005).

Lemma 3.12. *Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ where $\alpha > 0$. Then for every $q_{\boldsymbol{\lambda}^\star}$*

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

Proof. By assumption

$$\hat{\boldsymbol{\lambda}} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left[\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2\right] \ .$$

Further, let

$$\boldsymbol{\lambda}^{\star\star} = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left[\mathrm{L}_\pi(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2\right] \ .$$

As the first step, we show that

$$\left\|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\right\|_2 \le \frac{\|\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]\|_2}{\alpha} \ . \tag{3.22}$$

Assume that $\boldsymbol{\lambda}^{\star\star} \neq \hat{\boldsymbol{\lambda}}$ (otherwise Eq. 3.22 holds). Let $g(\boldsymbol{\lambda})$ denote $\ln Z_{\boldsymbol{\lambda}}$. This is the cumulant or the log partition function of the exponential family, which is convex in $\boldsymbol{\lambda}$ (Kapur and Kesavan, 1992). By convexity of the cumulant $g(\boldsymbol{\lambda})$ and the squared norm $\alpha\|\boldsymbol{\lambda}\|_2^2/2$, the gradients of

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

at their respective minima must equal zero:

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

Taking the difference yields

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

Taking inner product of both sides with $(\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}})$, we obtain

$$\begin{aligned}
\alpha\|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2^2 &= -(\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\nabla g(\boldsymbol{\lambda}^{\star\star}) - \nabla g(\hat{\boldsymbol{\lambda}})) + (\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]) \\
&\le (\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}) \cdot (\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]) \qquad (3.23) \\
&\le \|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2 \|\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]\|_2 \ . \qquad (3.24)
\end{aligned}$$

Eq. (3.23) follows because, by convexity of $g(\boldsymbol{\lambda})$, for all $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$

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

Eq. (3.24) follows by the Cauchy-Schwartz inequality. Dividing (3.24) by $\alpha \|\boldsymbol{\lambda}^{\star\star} - \hat{\boldsymbol{\lambda}}\|_2$ we obtain Eq. (3.22). Now, by the Generalization Lemma(ii), the Cauchy-Schwartz inequality, Eq. (3.22) and the optimality of $\boldsymbol{\lambda}^{\star\star}$, we obtain

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

Lemma 3.12 improves on Eq. (3.21) in the leading constant of $\|\boldsymbol{\lambda}^\star\|_2^2$ which is $\alpha/2$ instead of α. Taking the expectation over a random sample and bounding the trace of $\boldsymbol{\Sigma}$ in terms of the ℓ_2 diameter (see Lemma 3.7), we obtain an expectation guarantee. We can also use Lemma 3.8 to bound $\|\mathbf{E}_\pi[\boldsymbol{f}] - \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]\|_2^2$ with high probability, and obtain a probabilistic guarantee. The two results are presented in Theorem 3.13 with the tradeoff between the guarantees controlled by the parameter s.

Theorem 3.13. *Let $L_2, s > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \alpha \|\boldsymbol{\lambda}\|_2^2/2$ with*

$$\alpha = \frac{s D_2(\boldsymbol{f})}{L_2 \sqrt{m}} \ .$$

Then

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

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

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

The bounds of Theorem 3.13 are similar to the probabilistic guarantees for ℓ_2-regularized maxent. As mentioned earlier, they differ in the crucial fact that the norm $\|\boldsymbol{\lambda}^\star\|_2$ needs to be bounded *a priori* by a constant L_2. It is this constant rather than a possibly smaller norm $\|\boldsymbol{\lambda}^\star\|_2$ that enters the bound.

Similar to the bounds for ℓ_2-regularization, the bounds in this section generalize to arbitrary Hilbert-space and ellipsoid regularizations. The guarantees also generalize to regularizations obtained by taking squares of arbitrary norms as long as the corresponding concentration and expectation bounds are available.

Maxent with ℓ_2 Regularization versus ℓ_2^2 Regularization

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

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

$$\Lambda^{(\sqrt{2}),\beta} = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}^J}[\mathrm{L}_{\tilde{\pi}}(q_{\boldsymbol{\lambda}}) + \beta\|\boldsymbol{\lambda}\|_2] \tag{3.25}$$

$$\Lambda^{(2),\alpha} = \operatorname*{argmin}_{\boldsymbol{\lambda}\in\mathbb{R}^J}[\mathrm{L}_{\tilde{\pi}}(q_{\boldsymbol{\lambda}}) + \alpha\|\boldsymbol{\lambda}\|_2^2/2] \ . \tag{3.26}$$

If $\beta, \alpha > 0$ then $\Lambda^{(\sqrt{2}),\beta}$ and $\Lambda^{(2),\alpha}$ are non-empty because the objectives are lower semi-continuous and approach infinity as $\|\boldsymbol{\lambda}\|_2$ increases. For $\beta = 0$ and $\alpha = 0$, Eqs. (3.25) and (3.26) reduce to basic maxent. Thus, $\Lambda^{(\sqrt{2}),0}$ and $\Lambda^{(2),0}$ contain the $\boldsymbol{\lambda}$'s for which $\mathbf{E}_{q_{\boldsymbol{\lambda}}}[\boldsymbol{f}] = \mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]$. This set will be empty if the basic maxent solutions are attained only in a limit.

Theorem 3.14. *Let* $\Lambda^{(\sqrt{2})} = \bigcup_{\beta\in[0,\infty]} \Lambda^{(\sqrt{2}),\beta}$ *and* $\Lambda^{(2)} = \bigcup_{\alpha\in[0,\infty]} \Lambda^{(2),\alpha}$. *Then* $\Lambda^{(\sqrt{2})} = \Lambda^{(2)}$.

Proof. First note that $\Lambda^{(\sqrt{2}),\infty} = \Lambda^{(2),\infty} = \{\mathbf{0}\}$. Next, we will show that $\Lambda^{(\sqrt{2})} \setminus \{\mathbf{0}\} = \Lambda^{(2)} \setminus \{\mathbf{0}\}$. Taking derivatives in Eqs. (3.25) and (3.26), we obtain that $\boldsymbol{\lambda} \in \Lambda^{(\sqrt{2}),\beta} \setminus \{\mathbf{0}\}$ if and only if

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

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

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

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

The proof of Theorem 3.14 rests on the fact that the contours of the regularization functions $\|\boldsymbol{\lambda}\|_2$ and $\|\boldsymbol{\lambda}\|_2^2$ coincide. We could easily extend the proof to include the

equivalence of $\Lambda^{(\sqrt{2})}$, $\Lambda^{(2)}$ with the set of solutions to the problem

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^J} \left[\mathrm{L}_{\tilde{\pi}}(q_{\boldsymbol{\lambda}}) + \mathrm{I}\big(\|\boldsymbol{\lambda}\|_2 \le 1/\gamma\big)\right]$$

where $\gamma \in [0,\infty]$. Similarly, one could show the equivalence of the solutions with regularizations $\beta\|\boldsymbol{\lambda}\|_1$, $\alpha\|\boldsymbol{\lambda}\|_1^2/2$ and $\mathrm{I}(\|\boldsymbol{\lambda}\|_1 \le 1/\gamma)$.

The main implication of Theorem 3.14 is for maxent density estimation with selection of regularization parameters by the minimization of the held-out or cross-validated empirical error. In those cases, maxent versions with ℓ_2, ℓ_2^2 (and ℓ_2-ball indicator) regularization yield the same solution. Thus, we prefer to use the computationally least intensive method. This will typically be ℓ_2^2-regularized maxent whose potential and regularization are smooth.

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

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

Finally, we consider regularization that has both ℓ_1-style and ℓ_2^2-style terms. We will be able to derive both the expectation and probabilistic guarantees as in the case of ℓ_2^2 regularization, while retaining sparsity-inducing properties (and some generalization properties) of ℓ_1 regularization. To simplify the discussion, we apply the same regularization parameters β and α across all coordinates:

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

Here α and β are positive constants, and $|x|_+ = \max\{0, x\}$ denotes the positive part of x. To derive $\mathrm{U}_{\tilde{\pi}}^{(1+2)}$, notice that $\mathrm{U}_{\tilde{\pi}}^{(1+2)*}$ decomposes into a sum of functions of individual coordinates, so it suffices to derive a single coordinate potential $\mathrm{U}_{\tilde{\pi},j}^{(1+2)}$:

$$\begin{aligned}
\mathrm{U}_{\tilde{\pi},j}^{(1+2)}(u_j) &= \sup_{\lambda_j} \left(u_j\lambda_j - \beta|\lambda_j| - \frac{\alpha\lambda_j^2}{2} \right) \\
&= \sup_{\lambda_j : u_j\lambda_j = |u_j||\lambda_j|} \left(\frac{\alpha}{2} \cdot |\lambda_j| \cdot \left[\frac{2\big(|u_j| - \beta\big)}{\alpha} - |\lambda_j| \right] \right) \quad . \qquad (3.27)
\end{aligned}$$

In Eq. (3.27) we note that for each pair $\pm\lambda_j$, it suffices to consider the value whose sign agrees with u_j. Next, distinguish two cases. First, if $|u_j| \le \beta$ then the bracketed

expression is non-positive, hence the supremum is attained at $\lambda_j = 0$ and its value equals 0. Second, for $|u_j| > \beta$, the supremum is attained when $|\lambda_j| = (|u_j| - \beta)/\alpha$, in which case its value equals $(|u_j| - \beta)^2/(2\alpha)$, completing the derivation of (3.27). Using similar techniques as in the previous sections we can derive the following theorem.

Theorem 3.15. *Let $\delta, L_2 > 0$, and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta\|\boldsymbol{\lambda}\|_1 + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ with $\alpha = D_2(\boldsymbol{f})\min\{1/\sqrt{2}, \sqrt{m\delta}\} \,/\, (2L_2\sqrt{m})$ and $\beta = D_\infty(\boldsymbol{f})\sqrt{\ln(2|\mathcal{J}|/\delta)/(2m)}$. Then*

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

and with probability at least $1-\delta$

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

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

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

hence $\mathbf{E}\big[\mathrm{U}_{\tilde{\pi}}^{(1+2)}(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_\pi[\boldsymbol{f}])\big] \le \delta D_2(\boldsymbol{f})^2/(2\alpha)$. On the other hand, we can bound the trace of the feature covariance matrix by Lemma 3.7 and obtain

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

Hence

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

and the first claim follows. ∎

Setting $\delta = s/m$, we bound the difference in performance between the maxent distribution and any Gibbs distribution of a bounded weight vector by

$$O\left(\frac{\|\boldsymbol{\lambda}^\star\|_1 D_\infty(\boldsymbol{f})\sqrt{\ln(2m|\mathcal{J}|/s)} + L_2 D_2(\boldsymbol{f})\sqrt{s}}{\sqrt{m}}\right) \ .$$

Now the constant s can be tuned to achieve the optimal tradeoff between $\|\boldsymbol{\lambda}^\star\|_1 D_\infty(\boldsymbol{f})$ and $L_2 D_2(\boldsymbol{f})$. Notice that the sparsity inducing properties of ℓ_1 regularization are preserved in $\ell_1 + \ell_2^2$ regularization because partial derivatives of $\beta\|\boldsymbol{\lambda}\|_1 + \alpha\|\boldsymbol{\lambda}\|_2^2/2$ are

discontinuous at zero.

Finally, we point out how the results in this section generalize to regularizations defined by arbitrary pairs of norms $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{B}}$ as long as the corresponding concentration and expectation bounds are available. Specifically, consider the regularization

$$\mathrm{U}^*_{\tilde{\pi}}(\boldsymbol{\lambda}) = \beta\|\boldsymbol{\lambda}\|_{\mathcal{B}^*} + \frac{\alpha\|\boldsymbol{\lambda}\|^2_{\mathcal{A}^*}}{2}$$

where $\|\cdot\|_{\mathcal{A}^*}$ and $\|\cdot\|_{\mathcal{B}^*}$ are norms dual to $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{B}}$. The potential can then be bounded above using Eq. (2.16),

$$\begin{aligned}
\mathrm{U}_{\tilde{\pi}}(\boldsymbol{u}) &= \inf_{\boldsymbol{u}'}\left[\mathrm{I}\big(\|\boldsymbol{u}'\|_{\mathcal{B}} \le \beta\big) + \frac{\|\boldsymbol{u}-\boldsymbol{u}'\|^2_{\mathcal{A}}}{2\alpha}\right] \\
&\le \min_{\boldsymbol{u}'\in\{\mathbf{0},\boldsymbol{u}\}}\left[\mathrm{I}\big(\|\boldsymbol{u}'\|_{\mathcal{B}} \le \beta\big) + \frac{\|\boldsymbol{u}-\boldsymbol{u}'\|^2_{\mathcal{A}}}{2\alpha}\right] \\
&= \min\left\{\mathrm{I}\big(\|\boldsymbol{u}\|_{\mathcal{B}} \le \beta\big), \frac{\|\boldsymbol{u}\|^2_{\mathcal{A}}}{2\alpha}\right\}\ , \qquad (3.28)
\end{aligned}$$

yielding guarantees similar to Theorem 3.15.

3.4 Infinite Feature Classes

So far we have considered the generalization properties of maxent on finite feature classes bounded in the ℓ_∞ norm as well as the generalization properties on possibly infinite feature classes bounded in the ℓ_2 norm. In the former case, we have found that the ℓ_1 regularized maxent generalizes well if the number of features is smaller than an exponential of the number of samples. In the latter case, the ℓ_2-regularized maxent generalizes well regardless of the dimensionality. While these requirements seem rather modest, there are several interesting feature classes for which the previous results do not give satisfying guarantees.

For example, consider threshold features, hinge features, and decision paths derived from a set of variables $\mathcal{V}$. If we consider the continuum of possible thresholds for each variable $v \in \mathcal{V}$ then these feature classes have infinite sizes and infinite ℓ_2 diameters. Even if we consider only one threshold value between consecutive pairs of values that the variables v attain on $\mathcal{X}$, the number of features and the ℓ_2 diameters of the corresponding feature spaces will depend on the size of the sample space $\mathcal{X}$, which may be significantly larger than the number of samples m.

In this section we prove guarantees that do not require bounded ℓ_2 diameters and do not depend explicitly on the number of features or the size of $\mathcal{X}$. They will allow working with potentially infinite feature classes including classes of threshold

features, hinge features, and decision paths. Our approach is based on ℓ_1-regularized maxent. To bound the empirical errors we use a set of uniform convergence results known as Vapnik-Chervonenkis theory (VC theory).

3.4.1 VC bounds

VC theory was extensively developed by Vapnik and Chervonenkis (1968, 1971, 1974). Before proving specific results, we define some relevant concepts, following the exposition of Devroye et al. (1996), Chapters 12 and 13.

First, we define the *growth function* $s(\mathcal{F}, m)$, for a set of binary features $\mathcal{F}$ and the number of samples m, as the largest number of distinct labelings assigned by features in $\mathcal{F}$ to any set of m samples. In symbols:

$$s(\mathcal{F}, m) = \max_{x_1,\ldots,x_m \in \mathcal{X}} \left| \left\{ \big(f(x_1), \ldots, f(x_m)\big) : f \in \mathcal{F} \right\} \right| \ .$$

The *VC dimension* of $\mathcal{F}$ is the largest number of samples for which all possible labelings exist:

$$d(\mathcal{F}) = \max\{m : s(\mathcal{F}, m) = 2^m\} \ .$$

The growth function can be bounded in terms of the VC dimension by Sauer's lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972):

$$s(\mathcal{F}, m) \le \sum_{i=0}^{d(\mathcal{F})} \binom{m}{i} \ .$$

If $m > 2d(\mathcal{F})$ then the right-hand side of Sauer's lemma can be further bounded (see Devroye et al., 1996, Theorem 13.3), yielding the simpler inequality

$$\ln s(\mathcal{F}, m) \le d(\mathcal{F}) \ln(em/d(\mathcal{F})) \ . \tag{3.29}$$

A central result of VC theory is the uniform convergence of empirical averages of feature classes with finite VC dimension to their means, regardless of the actual number of features. Compared with Theorem 3.3, the number of features $|\mathcal{J}|$ is typically replaced by the growth function s, and $\ln|\mathcal{J}|$ is replaced by the VC dimension. For example, we can derive the following result:

Theorem 3.16. *Let $\mathcal{F}$ be a set of binary features indexed by $j \in \mathcal{J}$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j |\lambda_j|$ with $\beta_j = \beta = \sqrt{[\ln s(\mathcal{F}, m^2) + \ln(1/\delta) + \ln(4e^8)]/(2m)}$ for all j. Then*

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

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\|\boldsymbol{\lambda}^\star\|_1 \sqrt{\frac{\ln s(\mathcal{F}, m^2) + \ln(1/\delta) + \ln(4e^8)}{2m}} \ .$$

Proof. By Theorem A.5, we obtain that $|\mathbf{E}_\pi[f_j] - \mathbf{E}_{\tilde{\pi}}[f_j]| \le \beta$ for all the f_j's simultaneously, with probability at least $1-\delta$. The statement of the theorem now follows by Theorem 3.2. ■

Next, we consider a few applications of Theorem 3.16 for specific feature classes.

Example 3.17. *Half-spaces.* Consider the set of binary features defined as indicators of half-spaces using at most γ variables; for example, threshold features are half-spaces in a single variable. For any fixed γ-tuple of variables, there exist at most $2m^\gamma$ labelings induced by half-spaces in these variables (see, for example, Devroye et al., 1996, Corollary 13.1). Summing over all possible γ-tuples yields

$$s(\mathcal{F}, m) \le \binom{|\mathcal{V}|}{\gamma} 2m^\gamma \le 2|\mathcal{V}|^\gamma m^\gamma \ .$$

For the β_j's chosen according to Theorem 3.16, the performance of the half-space features can be bounded as

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + 2\|\boldsymbol{\lambda}^\star\|_1 \sqrt{\frac{\gamma \ln(m^2|\mathcal{V}|) + \ln(1/\delta) + \ln(8e^8)}{2m}} \ .$$

Note that even though the number of features is potentially infinite, we get a meaningful bound as long as $\gamma \ln(m^2|\mathcal{V}|) = o(m)$. Thus, if γ is fixed, as in threshold features, then the log of the number of variables specifies the feature complexity in a similar way as the log of the number of features in Section 3.2.1.

Example 3.18. *Spaces of bounded VC dimension.* If $d(\mathcal{F})$ is finite then the growth function can be bounded by Eq. (3.29) for $m > 2d(\mathcal{F})$. Theorem 3.16 then yields meaningful bounds as long as $d(\mathcal{F}) \ln(em^2/d(\mathcal{F})) = o(m)$. Examples of binary features with bounded VC dimension include half-spaces, Euclidean balls and ellipses, defined on at most γ variables from the set $\mathcal{V}$. VC dimensions of these classes, similar to VC dimensions of half-spaces, depend polynomially on γ and logarithmically on $|\mathcal{V}|$.

Example 3.19. *Decision paths.* Threshold features over the variables $\mathcal{V}$ define at most $2m|\mathcal{V}|$ distinct labelings on any m examples. Labelings by decision paths of length ℓ are conjunctions of ℓ threshold-feature labelings. Thus, the number of label-

ings they induce is at most

$$\binom{2m|\mathcal{V}|}{\ell} \leq (2m|\mathcal{V}|)^{\ell} \ .$$

Notice that threshold features are decision paths of length one. For the β_j's chosen by Theorem 3.16, the performance of threshold features and decision paths can be bounded as

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_{\pi}(\boldsymbol{\lambda}^{\star}) + 2\|\boldsymbol{\lambda}^{\star}\|_1 \sqrt{\frac{\ell \ln(2m^2|\mathcal{V}|) + \ln(1/\delta) + \ln(4e^8)}{2m}} \ .$$

Thus, we get a meaningful bound as long as $\ell \ln(2m^2|\mathcal{V}|) = o(m)$. Similarly to half-spaces, if the path length ℓ is fixed then the log of the number of variables specifies the feature complexity.

Example 3.20. *Conjunctions of spaces of bounded VC dimension.* Similar to the view of decision paths as conjunctions of threshold features, it is possible to consider conjunctions of classes of bounded VC dimension such as those mentioned in Example 3.18.

Among infinite feature classes introduced in Section 2.2, threshold features and decision paths are the only examples of binary-valued features, and therefore the only examples that fit directly in VC theory. Interesting examples of real-valued infinite feature classes are hinge features and splines. We will show that their performance, as well as performance of arbitrary features with a finite "total variation," can be bounded using the results for threshold features and decision paths. Before proving guarantees for infinite classes of real-valued features, we analyze ℓ_1 regularization of threshold features and decision paths in more detail.

3.4.2 ℓ_1 Regularization of Threshold Features

We begin with a more detailed analysis of threshold features. Threshold features are highly expressive, because they can model arbitrary additive responses to variables (i.e., arbitrary responses without pairwise or higher-order interactions). We analyze the cost of this generality.

Suppose we are given an arbitrary response function to a single variable v. What is the penalty incurred by using threshold features to express this function?

Consider a simple model q_g that depends on a single variable $v \in \mathcal{V}$ and is parameterized by a response $g : \mathbb{R} \to \mathbb{R}$,

$$q_g(x) = q_0(x) e^{g(v(x))} / Z_g \ ,$$

where Z_g ensures that the probabilities sum to one over $\mathcal{X}$. Threshold features can be used to represent q_g, because the space $\mathcal{X}$ is finite, and therefore g can be modeled on $\mathcal{X}$ exactly as a finite sum of step functions.

To determine an explicit representation of g, assume that $|\mathcal{X}| = N$, and sort the values $v(x)$, $x \in \mathcal{X}$, in an increasing order, calling them $\theta_0, \ldots, \theta_{N-1}$. Step functions $\mathbb{1}(t; t \geq \theta_i)$, used to derive threshold features, can be used to represent g as

$$g(t) = g(\theta_0) + \sum_{i=1}^{N-1} \Big(g(\theta_i) - g(\theta_{i-1}) \Big) \mathbb{1}(t; t \geq \theta_i) \ . \tag{3.30}$$

The constant $g(\theta_0)$ is normalized out of the exponent, whereas the differences $g(\theta_i) - g(\theta_{i-1})$ correspond to the coefficients of threshold features $\mathbb{1}(x; v(x) \geq \theta_i)$. Using ℓ_1-regularized maxent with threshold features for the variable v, and with β set according to Example 3.19, we obtain by Theorem 3.16

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \leq \mathrm{L}_\pi(g) + 2\beta \sum_{i=1}^{N-1} |g(\theta_i) - g(\theta_{i-1})| \ ,$$

where we used $\mathrm{L}_\pi(g)$ to denote the log loss

$$-\mathbf{E}_\pi \left[\ln \frac{q_g}{q_0} \right] \ .$$

Thus, the response g is penalized by

$$\sum_{i=1}^{N-1} |g(\theta_i) - g(\theta_{i-1})| \ .$$

This expression measures how much the value of g changes over the domain of g. In our specific case, this expression corresponds to the "total variation" of g. To be more precise, the *total variation* of a function $g : [t_{\min}, t_{\max}] \to \mathbb{R}$ is defined as

$$V(g) = \sup_{\substack{t_0 < t_1 < \cdots < t_n \\ n \in \mathbb{N}, t_0 = t_{\min}, t_n = t_{\max}}} \left[\sum_{i=1}^{n} |g(t_i) - g(t_{i-1})| \right] \ . \tag{3.31}$$

If g is continuously differentiable then it can be shown that

$$V(g) = \int_{t_{\min}}^{t_{\max}} \left| \frac{\mathrm{d}g(t)}{\mathrm{d}t} \right| \mathrm{d}t \ .$$

Thus, we have argued that threshold features in a single variable minimize

$$\mathrm{L}_{\tilde{\pi}}(g) + \beta \int_{t_{\min}}^{t_{\max}} \left| \frac{\mathrm{d}g(t)}{\mathrm{d}t} \right| \mathrm{d}t$$

over responses g for which the integral is defined. Total-variation regularization has been successfully applied, for example, in image restoration (Strong and Chan, 2003). The approach is similar to smoothing splines (see, for example, Wahba, 1990; or Hastie et al., 2001, Chapter 5), which minimize

$$\mathrm{L}_{\tilde{\pi}}(g) + \beta \int_{t_{\min}}^{t_{\max}} \left(\frac{\mathrm{d}^2 g(t)}{\mathrm{d}t^2} \right)^2 \mathrm{d}t \ .$$

Thus, responses obtained by the ℓ_1-regularized maxent with threshold features can be viewed as the ℓ_1 versions of solutions obtained by smoothing splines.

3.4.3 ℓ_1 Regularization of Decision Paths

The previous observations can be generalized to responses with a constant level of interaction when threshold features are replaced by decision paths. Before proving a theorem for the general case, we need a few definitions.

Assume that all variables are bounded in $[t_{\min}, t_{\max}]$ and let $G : [t_{\min}, t_{\max}]^{|\mathcal{V}|} \to \mathbb{R}$ denote an arbitrary response function. The corresponding Gibbs distribution is defined as

$$q_G(x) = q_0(x) e^{G(\boldsymbol{v}(x))} / Z_G \ ,$$

where $\boldsymbol{v}$ denotes the vector of all variables and Z_G is the normalization constant. Again, we will write $\mathrm{L}_r(G)$ for

$$-\mathbf{E}_r \left[\ln \frac{q_G}{q_0} \right] \ .$$

We say that the *order of interaction* (or simply the *order*) of G is ℓ if G can be written as

$$G(\boldsymbol{t}) = \sum_{i \in \mathcal{I}} g_i(\boldsymbol{t}_{\boldsymbol{k}_i}) \tag{3.32}$$

where $g_i : [t_{\min}, t_{\max}]^\ell \to \mathbb{R}$, $\boldsymbol{k}_i \in \{1, \ldots, |\mathcal{V}|\}^\ell$, i comes from an index set $\mathcal{I}$, and $\boldsymbol{t}_{\boldsymbol{k}_i}$ denotes the vector $(t_{k_{i,1}}, t_{k_{i,2}}, \ldots, t_{k_{i,\ell}})$.

In the previous section, we used threshold features to express a function g in a single variable, beginning from the "left" extreme of g's range and gradually adding step functions to approximate g. In this section, we will use decision paths to express functions g_i and start from a "corner", where the "corner" will refer to one of the vertices of the hypercube $[t_{\min}, t_{\max}]^\ell$. For mathematical convenience, we will assume

that functions g_i in Eq. (3.32) are constant on the hypercube-bounding hyperplanes that are adjacent to the corner. More precisely, we say that $g : [t_{\min}, t_{\max}]^\ell \to \mathbb{R}$ has the *corner property* if there exists a vector $\boldsymbol{c} \in \{t_{\min}, t_{\max}\}^\ell$ such that $g(\boldsymbol{t}) = g(\boldsymbol{c})$ whenever $t_k = c_k$ for some $k \in \{1, \ldots, \ell\}$. The vector $\boldsymbol{c}$ is called a *corner* of g. We say that a response G of order ℓ has the corner property, if all the functions in the decomposition of Eq. (3.32) have the corner property.

Next, we define a multivariate version of total variation. Among several (non-equivalent!) versions, the Vitali variation (Vitali, 1908; Fréchet, 1910; Lebesgue, 1910) is the right choice for our purposes. For a sufficiently smooth function $g : [t_{\min}, t_{\max}]^\ell \to \mathbb{R}$, its Vitali variation $V(g)$ is defined as

$$V(g) = \int_{t_{\min}}^{t_{\max}} \cdots \int_{t_{\min}}^{t_{\max}} \left| \frac{\partial^\ell g(t_1, \ldots, t_\ell)}{\partial t_1 \cdots \partial t_\ell} \right| \mathrm{d}t_1 \cdots \mathrm{d}t_\ell \ .$$

If the function g is not differentiable then the partial derivatives need to be replaced by differences. Specifically, define the k-th-coordinate difference operator $\Delta_{k;\delta}$, parameterized by δ, as

$$\Delta_{k;\delta}(g; t_1, \ldots, t_\ell) = g(t_1, \ldots, t_\ell) - g(t_1, \ldots, t_{k-1}, t_k - \delta, t_{k+1}, \ldots, t_\ell) \ .$$

Let Π be an arbitrary subdivision of the box $[t_{\min}, t_{\max}]^\ell$ by axes-aligned hyperplanes. Thus Π is specified by ℓ sequences (one along every axis)

$$\begin{gathered} t_{\min} = t_{1,0} < t_{1,1} < \cdots < t_{1,n_1} = t_{\max} \\ \vdots \\ t_{\min} = t_{\ell,0} < t_{\ell,1} < \cdots < t_{\ell,n_\ell} = t_{\max} \ . \end{gathered}$$

For a subdivision Π, let $\delta_{k,i}$ denote the difference $t_{k,i} - t_{k,i-1}$. The *Vitali variation* of g is defined as the least upper bound on the sums of the form

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_\ell=1}^{n_\ell} \left| \Delta_{1;\delta_{1,i_1}} \Delta_{2;\delta_{2,i_2}} \cdots \Delta_{\ell;\delta_{\ell,i_\ell}}(g; t_{i_1}, t_{i_2}, \ldots, t_{i_\ell}) \right| \tag{3.33}$$

taken over all possible subdivisions Π.

Later in this section, we will see that G can be decomposed into a sum of decision paths weighted by coefficients corresponding to the values of the ℓ-th order difference in Eq. (3.33); note that this is a natural generalization of the single-variable case considered in the previous section. Next we state a few properties of the difference operator and Vitali variation (proofs are omitted).

Proposition 3.21 (Linearity of Difference)**.**

$$\Delta_{k;\delta}(a_1 g_1 + a_2 g_2; \boldsymbol{t}) = a_1 \Delta_{k;\delta}(g_1; \boldsymbol{t}) + a_2 \Delta_{k;\delta}(g_2; \boldsymbol{t})$$

where g_1 and g_2 are functions, a_1 and a_2 are scalars, and $a_1 g_1 + a_2 g_2$ refers to the function $\boldsymbol{t} \mapsto a_1 g_1(\boldsymbol{t}) + a_2 g_2(\boldsymbol{t})$.

Proposition 3.22. *$\Delta_{k;\delta}(g) = 0$ if and only if g is constant along the k-th coordinate.*

Proposition 3.23 (Distributive Law for Difference)**.**

$$\Delta_{1;\delta_1} \Delta_{2;\delta_2} \cdots \Delta_{\ell;\delta_\ell}(g; \boldsymbol{t}) = \sum_{\boldsymbol{b} \in \{0,1\}^\ell} (-1)^{(\sum_k b_k)} g(\boldsymbol{t} - \mathbf{D}\boldsymbol{b})$$

where $\mathbf{D}$ is a diagonal matrix with entries $D_{kk} = \delta_k$.

Proposition 3.24. *Vitali variation does not change under monotone transformations of coordinates. Specifically, let $g : [t_{\min}, t_{\max}]^\ell \to \mathbb{R}$ and $g' : [t'_{\min}, t'_{\max}]^\ell \to \mathbb{R}$ such that*

$$g'(t'_1, \ldots, t'_\ell) = g(h_1(t'_1), \ldots, h_\ell(t'_\ell))$$

where $h_1, \ldots, h_\ell$ are strictly monotone continuous functions mapping the endpoints $t'_{\min}$, $t'_{\max}$ to the endpoints $t_{\min}$, $t_{\max}$. Then $V(g') = V(g)$.

Now, we are ready to state and prove the theorem characterizing solutions of ℓ_1-regularized maxent with decision paths.

Theorem 3.25. *Let $\mathcal{F}$ be the set of decision paths of length ℓ. The ℓ_1-regularized maxent problem*

$$\inf_{\boldsymbol{\lambda}} \left(\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \beta \|\boldsymbol{\lambda}\|_1 \right) \tag{3.34}$$

is equivalent to the minimization

$$\inf_{G} \left(\mathrm{L}_{\tilde{\pi}}(G) + \beta \sum_{i \in \mathcal{J}} V(g_i) \right) \tag{3.35}$$

where G is taken from the set of responses of order ℓ with the corner property, decomposed as in Eq. (3.32)*. Moreover, if $\hat{\boldsymbol{\lambda}}$ solves the maxent problem for*

$$\beta = \sqrt{\frac{\ell \ln(2m^2|\mathcal{V}|) + \ln(1/\delta) + \ln(4e^8)}{2m}}$$

then with probability at least $1 - \delta$, for all responses $G^\star$ of order ℓ with the corner property, decomposed as in Eq. (3.32)*,*

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(G^\star) + 2\beta \sum_{i \in \mathcal{J}} V(g_i^\star) \ .$$

The theorem states that ℓ_1-regularized maxent with decision paths of length ℓ corresponds to regularized log loss optimization within a class of responses of order ℓ. If G is sufficiently smooth then the condition that its order of interaction be ℓ is equivalent to requiring that G vanish after taking $(\ell+1)$-st derivative with respect to any set of $\ell+1$ distinct coordinates. The corner property can always be achieved, for example, by slightly enlarging the range of G. The regularization penalty for using G can then be expressed without referring to the decomposition of G as

$$\frac{\beta}{(t_{\max}-t_{\min})^{|\mathcal{V}|-\ell}} \int_{t_{\min}}^{t_{\max}} \cdots \int_{t_{\min}}^{t_{\max}} \sum_{\boldsymbol{k}\in\{1,\ldots,|\mathcal{V}|\}^\ell} \left| \frac{\partial^\ell G(t_1,\ldots,t_{|\mathcal{V}|})}{\partial t_{k_1}\cdots\partial t_{k_\ell}} \right| \mathrm{d}t_1 \cdots \mathrm{d}t_{|\mathcal{V}|} \ .$$

Similar to bounds in the previous sections, this penalty can be decomposed into a factor corresponding to the Gibbs-distribution complexity and a factor corresponding to the feature-space complexity. Complexity of the Gibbs distribution q_G is measured by the integral above, whereas the feature-space complexity, i.e., the complexity of using decision paths, is captured by β, roughly proportional to $\ell \ln(m|\mathcal{V}|)$, as we have derived earlier by VC theory. (For a more refined analysis of generalization properties of decision paths in classification, see for example Golea et al., 1998.)

Proof of Theorem 3.25. We focus on a single term $g = g_i$ in the decomposition of G. Similar to threshold features, we will show that g can be written as a constant plus a weighted sum of decision paths such that the corresponding ℓ_1 penalty is $V(g)$. Summing contributions across all functions g_i then yields the results of the theorem.

Assume that $|\mathcal{X}| = N$. Thus any variable v can attain at most N distinct values on $\mathcal{X}$. By Proposition 3.24, we can transform variables v so that g has a corner at $\mathbf{0}$ and $v(x) \in \{0,1,\ldots,N-1\}$ for all $v \in \mathcal{V}$, $x \in \mathcal{X}$. This transformation has no impact on expressivity or regularization of decision paths, since we can appropriately transform their decision thresholds and possibly change their decision inequalities from $\leq$ to $>$ or vice versa. From now on, we restrict the domain of g to the lattice $\{0,1,\ldots,N-1\}^\ell$.

Because of the corner property, the function $g(\boldsymbol{t}) - g(\mathbf{0})$ can be nonzero only on the set $\mathcal{T} = \{1,2,\ldots,N-1\}^\ell$. Note that the decision paths $\mathbb{1}(\boldsymbol{t}; \boldsymbol{t} \geq \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathcal{T}$, form a basis for the vector space of functions on $\mathcal{T}$. To see this, consider the lexicographic ordering of $\mathcal{T}$. The decision path $\mathbb{1}(\boldsymbol{t}; \boldsymbol{t} \geq \boldsymbol{\theta})$ equals one if $\boldsymbol{t} = \boldsymbol{\theta}$, but equals zero if $\boldsymbol{t}$ precedes $\boldsymbol{\theta}$ (it can be both zero and one if $\boldsymbol{t}$ follows $\boldsymbol{\theta}$ in the lexicographic ordering). Thus, proceeding "left" to "right" in this ordering (this corresponds to proceeding "from the corner"), we can express an arbitrary function on $\mathcal{T}$. Since the space of functions on $\mathcal{T}$ is $|\mathcal{T}|$-dimensional, and we consider only $|\mathcal{T}|$ decision paths, they form a basis. Thus, there exist unique coefficients $a_{\boldsymbol{\theta}}$ such that the function g can be

written as

$$g(\boldsymbol{t}) = g(\mathbf{0}) + \sum_{\boldsymbol{\theta} \in \mathcal{T}} a_{\boldsymbol{\theta}} \mathbb{1}(\boldsymbol{t} \geq \boldsymbol{\theta}) \ . \tag{3.36}$$

Let Δ_k be a shorthand for $\Delta_{k;1}$. We saw that in the one-dimensional case, the coefficient a_θ would be equal to $g(\theta) - g(\theta - 1) = \Delta_1(g;\theta)$. We will show that in the multidimensional case, the coefficient $a_{\boldsymbol{\theta}}$ is simply a higher-order difference. Operators Δ_k are linear, hence their composition is linear as well. To derive $a_{\boldsymbol{\theta}}$, we apply $\Delta_1 \cdots \Delta_\ell$ to both sides of Eq. (3.36), and evaluate the result at a point $\boldsymbol{t}^\star \in \mathcal{T}$:

$$\begin{aligned}
\Delta_1 \cdots \Delta_\ell \big(g(\boldsymbol{t});\ \boldsymbol{t}^\star\big) &= \Delta_1 \cdots \Delta_\ell \left(g(\mathbf{0}) + \sum_{\boldsymbol{\theta} \in \mathcal{T}} a_{\boldsymbol{\theta}} \mathbb{1}(\boldsymbol{t} \geq \boldsymbol{\theta});\ \boldsymbol{t}^\star\right) \\
&= \sum_{\boldsymbol{\theta} \in \mathcal{T}} a_{\boldsymbol{\theta}} \Delta_1 \cdots \Delta_\ell \left(\mathbb{1}(\boldsymbol{t} \geq \boldsymbol{\theta});\ \boldsymbol{t}^\star\right) && (3.37) \\
&= \sum_{\boldsymbol{\theta} \in \mathcal{T}} a_{\boldsymbol{\theta}} \sum_{\boldsymbol{b} \in \{0,1\}^\ell} (-1)^{(\sum_k b_k)} \mathbb{1}(\boldsymbol{t}^\star - \boldsymbol{b} \geq \boldsymbol{\theta}) && (3.38) \\
&= \sum_{\boldsymbol{\theta} \in \mathcal{T}} a_{\boldsymbol{\theta}} \sum_{\boldsymbol{b} \in \{0,1\}^\ell} (-1)^{(\sum_k b_k)} \mathbb{1}(\boldsymbol{t}^\star \geq \boldsymbol{\theta} + \boldsymbol{b}) \\
&= \sum_{\boldsymbol{\theta} \in \mathcal{T}} a_{\boldsymbol{\theta}} \mathbb{1}(\boldsymbol{t}^\star = \boldsymbol{\theta}) && (3.39) \\
&= a_{\boldsymbol{t}^\star} && (3.40)
\end{aligned}$$

Eq. (3.37) follows by linearity and Proposition 3.22. Eq. (3.38) follows by Proposition 3.23. In Eq. (3.39) we used the principle of inclusion and exclusion

$$\begin{aligned}
\mathbb{1}(\boldsymbol{t}^\star = \boldsymbol{\theta}) = \mathbb{1}(\boldsymbol{t}^\star \geq \boldsymbol{\theta}) - \mathbb{1}\Big(&\big(\{t_1^\star \geq \theta_1 + 1\} \cap \{\boldsymbol{t}_{-1}^\star \geq \boldsymbol{\theta}_{-1}\}\big) \\
&\cup \big(\{t_2^\star \geq \theta_2 + 1\} \cap \{\boldsymbol{t}_{-2}^\star \geq \boldsymbol{\theta}_{-2}\}\big) \\
&\ \ \vdots \\
&\cup \big(\{t_\ell^\star \geq \theta_\ell + 1\} \cap \{\boldsymbol{t}_{-\ell}^\star \geq \boldsymbol{\theta}_{-\ell}\}\big)\Big) \\
= \sum_{\boldsymbol{b} \in \{0,1\}^\ell} (-1)^{(\sum_k b_k)} &\mathbb{1}(\boldsymbol{t}^\star \geq \boldsymbol{\theta} + \boldsymbol{b}) \ ,
\end{aligned}$$

where $\boldsymbol{t}_{-k}$ denotes the $(\ell-1)$-tuple obtained from $\boldsymbol{t}$ by deleting t_k, i.e., the $(\ell-1)$-tuple $(t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_\ell)$.

Eq. (3.40) implies that the weights $a_{\boldsymbol{\theta}}$ of the decision paths $\mathbb{1}(\boldsymbol{t};\boldsymbol{t} \geq \boldsymbol{\theta})$ in Eq. (3.36) are equal to $\Delta_1 \cdots \Delta_\ell(g;\boldsymbol{\theta})$. Therefore, the contribution of g to the regularization in problem Eq. (3.34) is at most

$$\beta \sum_{\boldsymbol{\theta} \in \mathcal{T}} \Big| \Delta_1 \cdots \Delta_\ell(g;\boldsymbol{\theta}) \Big| \leq \beta V(g) \ . \tag{3.41}$$

Summing across all the g_i's, we obtain

$$\inf_{\boldsymbol{\lambda}}\Big(\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\beta\|\boldsymbol{\lambda}\|_1\Big)\leq\inf_{G}\Big(\mathrm{L}_{\tilde{\pi}}(G)+\beta\sum_{i\in\mathcal{I}}V(g_i)\Big)\ .$$

Next note that decision paths have the corner property and their Vitali variation equals one. When multiplied by λ_j, their Vitali variation becomes $|\lambda_j|$. Thus, plugging the responses of Eq. (3.34) into Eq. (3.35), we obtain the same value of the regularized objective. Hence,

$$\inf_{G}\Big(\mathrm{L}_{\tilde{\pi}}(G)+\beta\sum_{i\in\mathcal{I}}V(g_i)\Big)\leq\inf_{\boldsymbol{\lambda}}\Big(\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda})+\beta\|\boldsymbol{\lambda}\|_1\Big)\ ,$$

which completes the proof of the first part of the theorem. The second part follows from Eq. (3.41) applied to $g_i^\star$. ■

The crucial step in the proof of Theorem 3.25 was the decomposition

$$g_i(\boldsymbol{t})=g_i(\mathbf{0})+\sum_{\boldsymbol{\theta}\in\mathcal{T}}\Big(\Delta_1\cdots\Delta_\ell(g_i;\boldsymbol{\theta})\Big)\mathbb{1}(\boldsymbol{t}\geq\boldsymbol{\theta})\ . \tag{3.42}$$

Next, we will use this decomposition to prove performance guarantees for infinite classes of real-valued features.

3.4.4 Infinite Classes of Real-valued Features

We consider classes of real-valued features derived from a finite set of variables $\mathcal{V}$. To relate variables to features, we use the concept of a *dictionary*. A dictionary $\mathcal{G}$ of order ℓ is a set of pairs $(\boldsymbol{k}_j,g_j)$, where $\boldsymbol{k}_j\in\{1,\ldots,|\mathcal{V}|\}^\ell$, $g_j:\mathbb{R}^\ell\to\mathbb{R}$, and j comes from an index set $\mathcal{J}$. The dictionary $\mathcal{G}$ specifies the feature set $\mathcal{F}(\mathcal{G})=\{f_j\}_{j\in\mathcal{J}}$ where

$$f_j(x)=g_j(\boldsymbol{v}_{k_j}(x))\ .$$

For example, threshold features are specified by the dictionary

$$\begin{aligned}&\big\{\big(k,\mathbb{1}(t;t\geq\theta)\big):k\in\{1,\ldots,|\mathcal{V}|\},\theta\in\mathbb{R}\big\}\\&\quad\cup\big\{\big(k,\mathbb{1}(t;t<\theta)\big):k\in\{1,\ldots,|\mathcal{V}|\},\theta\in\mathbb{R}\big\}\ .\end{aligned}$$

Hinge features are specified by the dictionary

$$\begin{aligned}&\big\{\big(k,\mathsf{lh}(t;\theta,v_{k;\min})\big):k\in\{1,\ldots,|\mathcal{V}|\},\theta\in[v_{k;\min},v_{k;\max}]\big\}\\&\quad\cup\big\{\big(k,\mathsf{lh}(t;\theta,v_{k;\max})\big):k\in\{1,\ldots,|\mathcal{V}|\},\theta\in[v_{k;\min},v_{k;\max}]\big\}\ .\end{aligned}$$

Next, we consider the generalization properties of ℓ_1-regularized maxent with the feature set $\mathcal{F}(\mathcal{G})$. A crucial problem is determining the set of error bounds β_j so that $|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_{\pi}[f_j]| \le \beta_j$ for all $j \in \mathcal{J}$, or, equivalently

$$\left|\mathbf{E}_{\tilde{\pi}}\left[g_j(\boldsymbol{v}_{\boldsymbol{k}_j})\right] - \mathbf{E}_{\pi}\left[g_j(\boldsymbol{v}_{\boldsymbol{k}_j})\right]\right| \le \beta_j \text{ for all } j.$$

To obtain bounds β_j, we will use the decomposition (3.42).

Consider a single index j and let $g = g_j$ and $\boldsymbol{k} = \boldsymbol{k}_j$. The empirical error of the average $\mathbf{E}_{\tilde{\pi}}[g(\boldsymbol{v}_{\boldsymbol{k}})]$ can be bounded as

$$\begin{aligned}
&\left|\mathbf{E}_{\tilde{\pi}}[g(\boldsymbol{v}_{\boldsymbol{k}})] - \mathbf{E}_{\pi}[g(\boldsymbol{v}_{\boldsymbol{k}})]\right| \\
&\quad = \left|\sum_{\boldsymbol{\theta}\in\mathcal{T}} \Big(\Delta_1 \cdots \Delta_\ell(g;\boldsymbol{\theta})\Big)\Big(\mathbf{E}_{\tilde{\pi}}[\mathbb{1}(\boldsymbol{v}_{\boldsymbol{k}} \ge \boldsymbol{\theta})] - \mathbf{E}_{\pi}[\mathbb{1}(\boldsymbol{v}_{\boldsymbol{k}} \ge \boldsymbol{\theta})]\Big)\right| \\
&\quad \le \left(\sum_{\boldsymbol{\theta}\in\mathcal{T}} \Big|\Delta_1 \cdots \Delta_\ell(g;\boldsymbol{\theta})\Big|\right) \sup_{\boldsymbol{\theta}\in\mathcal{T}} \left|\mathbf{E}_{\tilde{\pi}}[\mathbb{1}(\boldsymbol{v}_{\boldsymbol{k}} \ge \boldsymbol{\theta})] - \mathbf{E}_{\pi}[\mathbb{1}(\boldsymbol{v}_{\boldsymbol{k}} \ge \boldsymbol{\theta})]\right| \\
&\quad \le V(g) \sup_{\boldsymbol{\theta}\in\mathcal{T}} \left|\mathbf{E}_{\tilde{\pi}}[\mathbb{1}(\boldsymbol{v}_{\boldsymbol{k}} \ge \boldsymbol{\theta})] - \mathbf{E}_{\pi}[\mathbb{1}(\boldsymbol{v}_{\boldsymbol{k}} \ge \boldsymbol{\theta})]\right| \ . \qquad (3.43)
\end{aligned}$$

Note that the supremum over $\boldsymbol{\theta}$ on the right-hand side is simply a supremum over empirical errors of decision-path averages, which can be bounded using VC theory similar to Example 3.19. Hence, we obtain the following theorem.

Theorem 3.26. *Let $\mathcal{G}$ be a dictionary of order ℓ. Let $\hat{\boldsymbol{\lambda}}$ minimize $\mathrm{L}_{\tilde{\pi}}(\boldsymbol{\lambda}) + \sum_j \beta_j|\lambda_j|$ for the feature set $\mathcal{F}(\mathcal{G})$ and*

$$\beta_j = V(g_j)\sqrt{\frac{\ell \ln(2m^2|\mathcal{V}|) + \ln(1/\delta) + \ln(4e^8)}{2m}} \ .$$

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

$$\mathrm{L}_{\pi}(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_{\pi}(\boldsymbol{\lambda}^\star) + 2\sum_{j\in\mathcal{J}} \beta_j|\lambda_j^\star| \ .$$

Theorem 3.26 can be viewed as a generalization of performance guarantees for decision paths, derived in Example 3.19, to real-valued features of bounded order. In particular, if $\mathcal{G}$ represents decision paths of length ℓ then we obtain the bound of Example 3.19.

Theorem 3.26 outlines some qualitative properties of feature sets that lead to good performance. Specifically, the theorem states that two crucial properties are the order of interaction and Vitali variation. We will see below that both can be determined easily for all feature classes considered in Section 2.2. However, bounds β_j obtained from Theorem 3.26 should be considered with caution, because they are

based on the bound (3.43), which may be overly pessimistic.

For example, when features are scaled so that each of them has Vitali variation one then our theorem suggests regularization equal to that of decision paths of the same order of interaction. Thus, hinge features, which have order one and Vitali variation one, should be regularized the same as threshold features. Similarly, linear features scaled to $[0,1]$, which are just special cases of hinge features, should be regularized the same as threshold features. However, we will see in Chapter 5 that in practice, the best-performing regularization of hinge features is one half of the best-performing regularization of threshold features.

For linear features, the values β_j required by Theorem 3.26 can be compared to the values required by Theorem 3.3. Specifically, Theorem 3.26 requires

$$\beta_j = \sqrt{\frac{\ln(2|\mathcal{V}|/\delta) + \ln(4m^2e^8)}{2m}} \ ,$$

whereas Theorem 3.3 only

$$\beta_j = \sqrt{\frac{\ln(2|\mathcal{V}|/\delta)}{2m}} \ .$$

In practice, linear features can be regularized at even lower levels, as we will see in Chapter 5.

We finish this section by determining variations of the feature classes from Section 2.2.

Example 3.27. *Hinge features and monotone dictionaries of order one.* Consider dictionaries of order one, containing only monotone functions (non-decreasing or non-increasing). To determine the total variation of a monotone function, notice that all of the differences on the right-hand side of the definition Eq. (3.31) have the same sign. Therefore, if g is monotone then its variation equals the difference between its maximum and minimum, i.e., $V(g) = D(g)$. Thus, for example, hinge features have total variation one. Theorem 3.26, for monotone dictionaries of order one, yields the setting

$$\beta_j = D(g_j)\sqrt{\frac{\ln(2|\mathcal{V}|/\delta) + \ln(4m^2e^8)}{2m}} \ ,$$

which can be viewed as a generalization of Theorem 3.5(i), replacing size of the feature set by its VC dimension.

Example 3.28. *Decision paths, hinge paths, and product dictionaries.* Several natural feature classes, such as decision paths and hinge paths, can be defined as products of simpler features. If the component features are defined in terms of dictionaries, it is natural to define the composite features in terms of dictionaries as well.

Specifically, let $\mathcal{G}_1,\ldots,\mathcal{G}_\ell$ be dictionaries with index sets $\mathcal{J}_1,\ldots,\mathcal{J}_\ell$. The *product dictionary* $\mathcal{G} = \mathcal{G}_1\mathcal{G}_2\cdots\mathcal{G}_\ell$ has the index set $\mathcal{J} = \mathcal{J}_1 \times \mathcal{J}_2 \times \cdots \times \mathcal{J}_\ell$, and contains pairs $(\boldsymbol{k}_j, g_j)$, where $j = (j_1,\ldots,j_\ell) \in \mathcal{J}$, such that

$$\boldsymbol{k}_j = (\boldsymbol{k}_{j_1},\ldots,\boldsymbol{k}_{j_\ell})$$
$$g_j(\boldsymbol{t}_1,\ldots,\boldsymbol{t}_\ell) = g_{j_1}(\boldsymbol{t}_1)g_{j_2}(\boldsymbol{t}_2)\cdots g_{j_\ell}(\boldsymbol{t}_\ell) \ .$$

Using the distributive law in the definition of the Vitali variation for g_j, we obtain

$$V(g_j) = V(g_{j_1})V(g_{j_2})\cdots V(g_{j_\ell}) \ .$$

Since threshold features and hinge features have total variations equal to one, decision paths and hinge paths have Vitali variations equal to one as well.

Chapter 4

Algorithms

In the previous chapter, we have discussed performance bounds for maxent with various types of regularization. Now we turn our attention to algorithms for solving generalized maxent problems. We propose two algorithms for generalized maxent with complete proofs of convergence. Our algorithms cover a wide class of potentials including basic, box and ℓ_2^2 potentials. Polyhedral and ℓ_2-ball potentials do not fall in this class, but the corresponding maxent problems can be transformed into versions for which our algorithms can still be applied.

4.1 Selective-update Algorithm

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

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

As explained in Section 2.5, the goal of the algorithm is to produce a sequence

$\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ maximizing the objective Q. In this and the next section we assume that the number of features is finite and the potential U is *decomposable* as defined below:

Definition 4.1. A potential $\mathrm{U} : \mathbb{R}^{\mathcal{J}} \to (-\infty, \infty]$ is called *decomposable* if it can be written as a sum of coordinate potentials $\mathrm{U}(\boldsymbol{u}) = \sum_j \mathrm{U}_j(u_j)$, each of which is a closed proper convex function bounded from below.

As a consequence of this definition, the conjugate potential U^* equals the sum of conjugate coordinate potentials U_j^* (see Eq. (2.15)) and $\mathrm{U}_j^*(0) = \sup_{u_j}[-\mathrm{U}_j(u_j)]$ is finite for all j.

Throughout this section we assume that values of features f_j lie in the interval $[0,1]$ and that features and coordinate potentials are non-degenerate in the sense that ranges $f_j(\mathcal{X})$ and intersections $\operatorname{dom}\mathrm{U}_j \cap [0,1]$ differ from $\{0\}$ and $\{1\}$. In Section 4.3 we show that a generalized maxent problem with a decomposable potential can always be reduced to a non-degenerate form.

Our algorithm works by iteratively adjusting the single weight λ_j that maximizes (an approximation of) the change in Q. To be more precise, suppose we add δ to λ_j. Let $\boldsymbol{\lambda}'$ be the resulting vector of weights, identical to $\boldsymbol{\lambda}$ except that $\lambda'_j = \lambda_j + \delta$. Then the change in the objective is

$$\begin{aligned}
Q(\boldsymbol{\lambda}') - Q(\boldsymbol{\lambda}) &= -\ln Z_{\boldsymbol{\lambda}'} - \mathrm{U}^*(-\boldsymbol{\lambda}') + \ln Z_{\boldsymbol{\lambda}} + \mathrm{U}^*(-\boldsymbol{\lambda}) \\
&= -\ln\left(\mathbf{E}_{q_{\boldsymbol{\lambda}}}\left[e^{\delta f_j}\right]\right) - \sum_{j' \in \mathcal{J}} \left[\mathrm{U}_{j'}^*(-\lambda'_{j'}) - \mathrm{U}_{j'}^*(-\lambda_{j'})\right] && (4.1) \\
&\geq -\ln\left(\mathbf{E}_{q_{\boldsymbol{\lambda}}}\left[1 + (e^{\delta} - 1) f_j\right]\right) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) && (4.2) \\
&= -\ln\left(1 + (e^{\delta} - 1)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]\right) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \ . && (4.3)
\end{aligned}$$

Eq. (4.1) uses

$$Z_{\boldsymbol{\lambda}'} = \sum_{x \in \mathcal{X}} q_0(x) e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x) + \delta_j f_j(x)} = Z_{\boldsymbol{\lambda}} \sum_{x \in \mathcal{X}} q_{\boldsymbol{\lambda}}(x) e^{\delta_j f_j(x)} \ . \tag{4.4}$$

Eq. (4.2) is because $e^{\delta x} \leq 1 + (e^{\delta} - 1)x$ for $x \in [0,1]$ by convexity.

Let $F_j(\boldsymbol{\lambda}, \delta)$ denote the expression in (4.3):

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

Our algorithm, shown in Fig. 4.1, on each iteration, maximizes this lower bound over all choices of (j, δ), and for the maximizing j adds the corresponding δ to λ_j. We assume that for each j the maximizing δ is finite, which is always the case for non-degenerate potentials and features (see Section 4.3). Note that $F_j(\boldsymbol{\lambda}, \delta)$ is strictly concave in δ so we can use any of a number of search methods to find the optimal δ.

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

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

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

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

- let $(j^\star, \delta^\star) = \operatorname*{argmax}_{(j,\delta)} \left[-\ln\left(1 + (e^\delta - 1)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]\right) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \right]$
- $\lambda_{t+1,j} = \begin{cases} \lambda_{t,j^\star} + \delta^\star & \text{if } j = j^\star \\ \lambda_{t,j} & \text{otherwise} \end{cases}$

Figure 4.1. *Selective-update algorithm for maximum entropy (SUMMET).*

4.1.1 Solving ℓ_1-Regularized Maxent

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

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

since

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

The optimum δ can be obtained for each j via a simple case analysis on the sign of $\lambda_j + \delta$. In particular, using calculus, we see that we only need consider the possibility that $\delta = -\lambda_j$ or that δ is equal to

$$\ln\left(\frac{(\mathbf{E}_{\tilde{\pi}}[f_j] - \beta_j)(1 - \mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j])}{(1 - \mathbf{E}_{\tilde{\pi}}[f_j] + \beta_j)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]}\right) \quad \text{or} \quad \ln\left(\frac{(\mathbf{E}_{\tilde{\pi}}[f_j] + \beta_j)(1 - \mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j])}{(1 - \mathbf{E}_{\tilde{\pi}}[f_j] - \beta_j)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]}\right)$$

where the first and second of these can be valid only if $\lambda_j + \delta \geq 0$ and $\lambda_j + \delta \leq 0$, respectively. The complete algorithm, ℓ_1-SUMMET, is shown in Fig. 4.2.

4.1.2 Reductions from Non-decomposable Potentials

Polyhedral and ℓ_2-ball potentials are not decomposable. When a polyhedral potential is represented as an intersection of halfspaces $\boldsymbol{\eta}_k \cdot \boldsymbol{u} \geq a_k$, it suffices to use transformed features $f'_k(x) = \boldsymbol{\eta}_k \cdot \boldsymbol{f}(x)$ with coordinate potentials corresponding to the inequality constraints.

For the ℓ_2-ball potential, we replace the constraint $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_p[\boldsymbol{f}]\|_2 \leq \beta$ by an

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

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

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

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

- $(j^\star, \delta^\star) = \operatorname*{argmax}_{(j,\delta)} \left[-\ln\left(1 + (e^\delta - 1) q_t[f_j]\right) + \mathbf{E}_{\tilde{\pi}}[f_j]\delta - \beta_j(|\lambda_{t,j} + \delta| - |\lambda_{t,j}|) \right]$

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

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

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

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

Figure 4.2. *Selective-update algorithm for ℓ_1-regularized maxent (ℓ_1-SUMMET).*

equivalent constraint $\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_p[\boldsymbol{f}]\|_2^2 \leq \beta^2$, and obtain an equivalent primal

$$\min_{p \in \Delta} \mathrm{D}(p \parallel q_0) \text{ subject to } \|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_p[\boldsymbol{f}]\|_2^2 \leq \beta^2 \quad . \tag{4.5}$$

If $\beta > 0$ then, by Lagrange duality and Slater's conditions (Boyd and Vandenberghe, 2004, Chapter 5), the value of Eq. (4.5) is the same as the value of

$$\max_{\mu \geq 0} \min_{p \in \Delta} \left[\mathrm{D}(p \parallel q_0) + \mu(\|\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] - \mathbf{E}_p[\boldsymbol{f}]\|_2^2 - \beta^2) \right] \quad . \tag{4.6}$$

The max-min value of Eq. (4.6) is attained at the saddle-point of the objective, at $(\hat{\mu}, \hat{p})$, where $\hat{p}$ minimizes Eq. (4.5). Since the outer maximization of Eq. (4.6) is concave in μ, we can employ a range of search techniques to find the optimal μ, evaluating the inner minimum by ℓ_2^2-regularized SUMMET of this section (or PLUMMET of the next section).

4.1.3 Convergence

In order to prove convergence of SUMMET, we will measure its progress toward solving the primal and dual. One measure of progress is the difference between the

primal evaluated at $q_{\boldsymbol{\lambda}}$ and the dual evaluated at $\boldsymbol{\lambda}$:

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

By Theorem 2.3, this difference is non-negative and equals zero exactly when $q_{\boldsymbol{\lambda}}$ solves the primal and $\boldsymbol{\lambda}$ solves the dual.

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

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

for all j. When coordinate potentials express equality and inequality constraints, this characterization corresponds to the Karush-Kuhn-Tucker conditions (Rockafellar, 1970) .

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

Definition 4.2. A function $A : \mathbb{R}^{\mathcal{J}} \times \mathbb{R}^{\mathcal{J}} \to (-\infty, \infty]$ is called an *auxiliary function* if

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

where $\mathrm{B}(\cdot \parallel \cdot) : \mathbb{R}^{\mathcal{J}} \times \mathbb{R}^{\mathcal{J}} \to (-\infty, \infty]$ satisfies conditions (B1) and (B2) (see p. 29).

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

An auxiliary function is always non-negative since $\mathrm{U}(\boldsymbol{a}) + \mathrm{U}^*(-\boldsymbol{\lambda}) \geq -\boldsymbol{\lambda} \cdot \boldsymbol{a}$ by Fenchel's inequality and hence $A(\boldsymbol{\lambda}, \boldsymbol{a}) \geq \mathrm{B}(\boldsymbol{a} \parallel \mathbf{E}_{q_{\boldsymbol{\lambda}}}[\boldsymbol{f}]) \geq 0$. Moreover, if $A(\boldsymbol{\lambda}, \boldsymbol{a}) = 0$

then $\mathbf{E}_{q_\lambda}[\boldsymbol{f}] = \boldsymbol{a}$ and $A(\boldsymbol{\lambda}, \boldsymbol{a}) = P(q_\lambda) - Q(\boldsymbol{\lambda}) = 0$, i.e., by maxent duality, q_λ solves the primal and $\boldsymbol{\lambda}$ solves the dual.

It turns out, as we show in Lemma 4.4 below, that the optimality property generalizes to the case when $A(\boldsymbol{\lambda}_t, \boldsymbol{a}_t) \to 0$ provided that $Q(\boldsymbol{\lambda}_t)$ has a finite limit. In particular, it suffices to find a suitable sequence of $\boldsymbol{a}_t$'s for $\boldsymbol{\lambda}_t$'s produced by an algorithm to show its convergence. Note that the optimality in the limit trivially holds when $\boldsymbol{\lambda}_t$'s and $\boldsymbol{a}_t$'s come from a compact set, because $A(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{a}}) = 0$ at a cluster point of $\{(\boldsymbol{\lambda}_t, \boldsymbol{a}_t)\}$ by the lower semi-continuity of U and U^*.

In the general case, we follow the technique used by Della Pietra et al. (1997) for basic maxent: we consider a cluster point $\hat{q}$ of $\{q_{\lambda_t}\}$ and show that (i) $\hat{q}$ is primal feasible and (ii) the difference $P(\hat{q}) - Q(\boldsymbol{\lambda}_t)$ approaches zero. In basic maxent, $A(\boldsymbol{\lambda}, \boldsymbol{a}) = \mathrm{B}(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}] \parallel \mathbf{E}_{q_\lambda}[\boldsymbol{f}])$ whenever finite. Thus, (i) is obtained by (B2), and noting that $P(\hat{q}) - Q(\boldsymbol{\lambda}) = \mathrm{D}(\hat{q} \parallel q_\lambda)$ yields (ii). For a general potential, however, claims (i) and (ii) seem to require a novel approach. In both steps, we use decomposability and the technical Lemma 4.3 (proved in Section 4.1.3). Thus, for the non-compact case, decomposability seems crucial in the present approach.

Lemma 4.3. *Let* U_r *be a decomposable potential relative to a primal-feasible point* r. *Let* $S = \operatorname{dom} \mathrm{U}_r = \{\boldsymbol{u} \in \mathbb{R}^\mathcal{J} : \mathrm{U}_r(\boldsymbol{u}) < \infty\}$ *and* $T_c = \{\boldsymbol{\lambda} \in \mathbb{R}^\mathcal{J} : \mathrm{U}_r^*(\boldsymbol{\lambda}) \le c\}$. *Then there exists* $\alpha_c \ge 0$ *such that* $\boldsymbol{\lambda} \cdot \boldsymbol{u} \le \alpha_c \|\boldsymbol{u}\|_1$ *for all* $\boldsymbol{u} \in S, \boldsymbol{\lambda} \in T_c$.

Lemma 4.4. *Let* $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots \in \mathbb{R}^\mathcal{J}$, $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots \in \mathbb{R}^\mathcal{J}$ *be sequences such that* $Q(\boldsymbol{\lambda}_t)$ *has a finite limit and* $A(\boldsymbol{\lambda}_t, \boldsymbol{a}_t) \to 0$ *as* $t \to \infty$. *Then* $\lim_{t\to\infty} Q(\boldsymbol{\lambda}_t) = \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda})$.

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

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

$$A(\boldsymbol{\lambda}_\tau, \boldsymbol{a}_\tau) = \mathrm{U}_r(\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau) + \mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) - \boldsymbol{\lambda}_\tau \cdot (\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau) + \mathrm{B}(\boldsymbol{a}_\tau \parallel \mathbf{E}_{q_\tau}[\boldsymbol{f}]) \ . \tag{4.7}$$

Rearranging terms of Eq. (4.7), noting that $A(\boldsymbol{\lambda}_\tau, \boldsymbol{a}_\tau) \to 0$ and $\mathrm{B}(\boldsymbol{a}_\tau \parallel \mathbf{E}_{q_\tau}[\boldsymbol{f}]) \to 0$, and denoting the vanishing terms by $o(1)$, we get

$$\mathrm{U}_r(\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau) = -\mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) + \boldsymbol{\lambda}_\tau \cdot (\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau) + o(1) \ . \tag{4.8}$$

We use Eq. (4.8) to prove first the feasibility and then the optimality of $\hat{q}$ with respect to the primal objective.

Feasibility. We bound the right-hand side of Eq. (4.8) and take limits to show that $\mathrm{U}_r(\mathbf{E}_r[\boldsymbol{f}] - \mathbf{E}_{\hat{q}}[\boldsymbol{f}])$ is finite. The first term is bounded by Fenchel's inequality:

$$-\mathrm{U}_r^*(\boldsymbol{\lambda}_\tau) \leq -\boldsymbol{\lambda}_\tau \cdot \mathbf{0} + \mathrm{U}_r(\mathbf{0}) = \mathrm{U}_r(\mathbf{0}) \ , \tag{4.9}$$

which is finite by the feasibility of r. In order to bound $\boldsymbol{\lambda}_\tau \cdot (\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau)$, the second term of Eq. (4.8), we use Lemma 4.3. First note that $\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau$ is a feasible point of U_r for all τ by Eq. (4.7) and the finiteness of $A(\boldsymbol{\lambda}_\tau, \boldsymbol{a}_\tau)$. Next, we need to show that $\mathrm{U}_r^*(\boldsymbol{\lambda}_\tau)$ is bounded above by a constant. We rearrange Eq. (2.23),

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

and bound the right-hand side, term by term: $-Q(\boldsymbol{\lambda}_\tau)$ has a finite limit and is thus bounded above; $-\mathrm{D}(r \parallel q_\tau)$ is non-positive; and $\mathrm{D}(r \parallel q_0)$ is a finite constant. Hence we can apply Lemma 4.3, and obtain

$$\boldsymbol{\lambda}_\tau \cdot (\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau) \leq \alpha_r \|\mathbf{E}_r[\boldsymbol{f}] - \boldsymbol{a}_\tau\|_1 \tag{4.10}$$

for some constant α_r independent of τ. Plugging Eqs. (4.9) and (4.10) in Eq. (4.8) and taking limits, we obtain by lower semi-continuity of U_r

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

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

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

$$\mathrm{U}_{\hat{q}}(\mathbf{E}_{\hat{q}}[\boldsymbol{f}] - \boldsymbol{a}_\tau) = -\mathrm{U}_{\hat{q}}^*(\boldsymbol{\lambda}_\tau) + \boldsymbol{\lambda}_\tau \cdot (\mathbf{E}_{\hat{q}}[\boldsymbol{f}] - \boldsymbol{a}_\tau) + o(1) \tag{4.11}$$

$$\leq -\mathrm{U}_{\hat{q}}^*(\boldsymbol{\lambda}_\tau) + \alpha_{\hat{q}} \|\mathbf{E}_{\hat{q}}[\boldsymbol{f}] - \boldsymbol{a}_\tau\|_1 + o(1) \ . \tag{4.12}$$

Eq. (4.11) follows from Eq. (4.8). Eq. (4.12) follows from Eq. (4.10). Taking limits, we obtain

$$\mathrm{U}_{\hat{q}}(\mathbf{0}) \leq \lim_{\tau \to \infty} \left[-\mathrm{U}_{\hat{q}}^*(\boldsymbol{\lambda}_\tau) \right] \ . \tag{4.13}$$

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

$$\begin{aligned} P(\hat{q}) &= \mathrm{D}(\hat{q} \parallel q_0) + \mathrm{U}_{\hat{q}}(\mathbf{0}) \\ &\leq \mathrm{D}(\hat{q} \parallel q_0) + \lim_{\tau\to\infty} \left[-\mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau)\right] && (4.14) \\ &= \lim_{\tau\to\infty} \left[\mathrm{D}(\hat{q} \parallel q_0) - \mathrm{D}(\hat{q} \parallel q_\tau) - \mathrm{U}^*_{\hat{q}}(\boldsymbol{\lambda}_\tau)\right] && (4.15) \\ &= \lim_{\tau\to\infty} Q(\boldsymbol{\lambda}_\tau) && (4.16) \\ &\leq \sup_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}) \leq P(\hat{q}) \ . && (4.17) \end{aligned}$$

Eq. (4.14) follows from Eq. (4.13). Eq. (4.15) follows from the continuity of relative entropy since $q_\tau \to \hat{q}$. Eq. (4.16) follows from Eq. (2.23). Eq. (4.17) follows by maxent duality. Eqs. (4.14)–(4.17) show that

$$P(\hat{q}) = \lim_{\tau\to\infty} Q(\boldsymbol{\lambda}_\tau) \ .$$

Hence, by maxent duality, $\hat{q}$ minimizes the primal and $\boldsymbol{\lambda}_\tau$ maximizes the dual as $\tau \to \infty$. ■

Theorem 4.5. *SUMMET produces a sequence $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ for which*

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

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

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

In each step of the algorithm

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

Since Q has a finite limit, differences $Q(\boldsymbol{\lambda}_{t+1}) - Q(\boldsymbol{\lambda}_t)$ converge to zero and thus $F_{t,j} \to 0$. We use $F_{t,j}$ to define an auxiliary function. To begin, we rewrite $F_{t,j}$ us-

ing Fenchel's duality:

$$\begin{aligned}
F_{t,j} &= \max_{\delta}\Big[-\ln\Big(1+(e^{\delta}-1)\mathbf{E}_{q_t}[f_j]\Big) - \mathrm{U}_j^*(-\lambda_{t,j}-\delta) + \mathrm{U}_j^*(-\lambda_{t,j})\Big] \\
&= \max_{\delta}\Big[-\ln\Big\{\big(1-\mathbf{E}_{q_t}[f_j]\big)e^{0\cdot\delta} + \mathbf{E}_{q_t}[f_j]e^{1\cdot\delta}\Big\} - \mathrm{U}_j'^*(-\delta)\Big] + \mathrm{U}_j^*(-\lambda_{t,j}) \qquad (4.18)\\
&= \min_{\bar{a},a}\Big[\mathrm{D}\big((\bar{a},a)\,\big\|\,(1-\mathbf{E}_{q_t}[f_j],\mathbf{E}_{q_t}[f_j])\big) + \mathrm{U}_j'(0\cdot\bar{a}+1\cdot a)\Big] + \mathrm{U}_j^*(-\lambda_{t,j}) \qquad (4.19)\\
&= \min_{0\le a\le 1}\Big[\mathrm{D}(a\,\|\,\mathbf{E}_{q_t}[f_j]) + \mathrm{U}_j(a) + \lambda_{t,j}\cdot a\Big] + \mathrm{U}_j^*(-\lambda_{t,j})\ . \qquad (4.20)
\end{aligned}$$

In Eq. (4.18), we rearranged terms inside the logarithm so they would take the form of a partition function. We write $\mathrm{U}_j'^*(u)$ for $\mathrm{U}_j^*(u-\lambda_{t,j})$. In Eq. (4.19), we applied Theorem 2.3, noting that the conjugate of the log partition function is relative entropy (see Section 2.4). The value of the relative entropy $\mathrm{D}\big((\bar{a},a)\,\big\|\,(1-\mathbf{E}_{q_t}[f_j],\mathbf{E}_{q_t}[f_j])\big)$ is infinite whenever $(\bar{a},a)$ is not a probability distribution, so it suffices to consider pairs where $0\le a\le 1$ and $\bar{a}=1-a$. In Eq. (4.20), we use $\mathrm{D}(a\,\|\,\mathbf{E}_{q_t}[f_j])$ as a shorthand for $\mathrm{D}\big((1-a,a)\,\big\|\,(1-\mathbf{E}_{q_t}[f_j],\mathbf{E}_{q_t}[f_j])\big)$. We use Eq. (2.13) to convert U_j' into U_j:

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

The minimum in Eq. (4.20) is always attained because a comes from a compact set and the minimized expression is lower semi-continuous in a. We use $a_{t,j}$ to denote a value attaining this minimum. Thus

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

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

$$A(\boldsymbol{\lambda},\boldsymbol{a}) = \sum_{j\in\mathcal{J}}\Big[\mathrm{U}_j(a_j) + \mathrm{U}_j^*(-\lambda_j) + \lambda_j a_j + \mathrm{D}(a_j\,\|\,\mathbf{E}_{q_\lambda}[f_j])\Big]\ .$$

Now $A(\boldsymbol{\lambda}_t,\boldsymbol{a}_t) = \sum_j F_{t,j} \to 0$, and the result follows by Lemma 4.4. ■

Proof of Lemma 4.3

We will first prove a single coordinate version of Lemma 4.3 and then turn to the general case.

Lemma 4.6. *Let $\psi : \mathbb{R} \to (-\infty,\infty]$ be a proper closed convex function. Let $S = \operatorname{dom}\psi = \{u\in\mathbb{R} : \psi(u) < \infty\}$ and $T_c = \{\lambda\in\mathbb{R} : \psi^*(\lambda)\le c\}$. Then there exists $\alpha_c \ge 0$ such that $u\lambda \le \alpha_c|u|$ for all $u\in S, \lambda\in T_c$.*

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

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

$$u_+\lambda \le \psi(u_+) + \psi^*(\lambda) \le \psi(u_+) + c$$

and thus

$$\lambda \le \frac{\psi(u_+) + c}{u_+} \quad .$$

Now for any $u \in S_+$

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

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

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

To complete the proof we choose

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

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

Proof of Lemma 4.3. Assume that $\mathrm{U}_r(\boldsymbol{u}) < \infty$ and thus by decomposability $\mathrm{U}_{r,j}(u_j) < \infty$ for all j. Also assume that $\mathrm{U}^*_r(\boldsymbol{\lambda}) = \sum_j \mathrm{U}^*_{r,j}(\lambda_j) \le c$. By Fenchel's inequality $\mathrm{U}^*_{r,j}(\lambda_j) \ge -\mathrm{U}_{r,j}(0)$ which is finite by the feasibility of r. Since the sum of $\mathrm{U}^*_{r,j}(\lambda_j)$ is bounded above by c and individual functions are bounded below by constants, they must also be bounded above by some constants c_j. By Lemma 4.6 applied to coordinate potentials, we obtain that $u_j\lambda_j \le \alpha_j|u_j|$ for some constants α_j. The conclusion follows by taking $\alpha_c = \max_j \alpha_j$. ■

4.2 Parallel-update Algorithm

Much of this dissertation has tried to be relevant to the case in which we are faced with a very large number of features. However, when the number of features is relatively small, it may be reasonable to maximize Q using an algorithm that updates all features simultaneously on every iteration. In this section, we describe a variant of generalized iterative scaling (Darroch and Ratcliff, 1972) applicable to generalized maxent with an arbitrary decomposable potential, and prove its convergence. Note that gradient-based or Newton methods may be faster in practice similar to the unregularized case (Malouf, 2002).

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

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

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

- for each j, let

$$\delta_j = \operatorname*{argmax}_{\delta} \left[-\mathbf{E}_{q_t}[f_j](e^{\delta} - 1) - \mathrm{U}_j^*(-\lambda_{t,j} - \delta) + \mathrm{U}_j^*(-\lambda_{t,j}) \right]$$

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

Figure 4.3. *Parallel-update algorithm for maximum entropy (PLUMMET).*

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

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

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

- for each j, let

$$\delta_j = \operatorname*{argmax}_{\delta} \left[-\mathbf{E}_{q_t}[f_j](e^{\delta} - 1) + \delta \mathbf{E}_{\tilde{\pi}}[f_j] - \beta_j(|\lambda_{t,j} + \delta| - |\lambda_{t,j}|) \right]$$

it suffices to consider the following possibilities (whenever defined)

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

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

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

Figure 4.4. *Parallel-update algorithm for ℓ_1-regularized maxent (ℓ_1-PLUMMET).*

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

Similarly to SUMMET of the previous section, our *parallel-update algorithm for maximum entropy* (PLUMMET) is based on an approximation of the change in the

objective Q, in this case the following, where $\boldsymbol{\lambda}' = \boldsymbol{\lambda} + \boldsymbol{\delta}$:

$$\begin{aligned} Q(\boldsymbol{\lambda}') - Q(\boldsymbol{\lambda}) &= -\ln Z_{\boldsymbol{\lambda}'} - \mathrm{U}^*(-\boldsymbol{\lambda}') + \ln Z_{\boldsymbol{\lambda}} + \mathrm{U}^*(-\boldsymbol{\lambda}) \\ &= -\ln\bigl(\mathbf{E}_{q_{\boldsymbol{\lambda}}}\bigl[e^{\boldsymbol{\delta}\cdot\boldsymbol{f}}\bigr]\bigr) - \mathrm{U}^*(-\boldsymbol{\lambda} - \boldsymbol{\delta}) + \mathrm{U}^*(-\boldsymbol{\lambda}) \qquad (4.21) \\ &\geq \sum_{j \in \mathcal{J}} \Bigl[-(e^{\delta_j} - 1)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j] - \mathrm{U}_j^*(-\lambda_j - \delta_j) + \mathrm{U}_j^*(-\lambda_j) \Bigr] \ . \qquad (4.22) \end{aligned}$$

Eq. (4.21) uses Eq. (4.4). For Eq. (4.22), note first that if $x_j \in \mathbb{R}$ and $p_j \geq 0$ with $\sum_j p_j \leq 1$ then

$$\exp\Bigl(\sum_{j \in \mathcal{J}} x_j p_j\Bigr) - 1 \leq \sum_{j \in \mathcal{J}} (e^{x_j} - 1) p_j \ .$$

(See Collins et al. (2002) for a proof.) Thus,

$$\begin{aligned} \ln \mathbf{E}_{q_{\boldsymbol{\lambda}}}\bigl[\exp\bigl(\sum_{j \in \mathcal{J}} \delta_j f_j\bigr)\bigr] &\leq \ln \mathbf{E}_{q_{\boldsymbol{\lambda}}}\bigl[1 + \sum_{j \in \mathcal{J}} (e^{\delta_j} - 1) f_j\bigr] \\ &= \ln\bigl(1 + \sum_{j \in \mathcal{J}} (e^{\delta_j} - 1)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]\bigr) \\ &\leq \sum_{j \in \mathcal{J}} (e^{\delta_j} - 1)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j] \end{aligned}$$

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

PLUMMET, shown in Fig. 4.3, on each iteration, maximizes Eq. (4.22) over all choices of the δ_j's. For the basic potential $\mathrm{U}^{(0)}$, this algorithm reduces to the generalized iterative scaling of Darroch and Ratcliff (1972). For ℓ_1-style regularization, the maximizing $\boldsymbol{\delta}$ can be calculated explicitly (see algorithm ℓ_1-PLUMMET in Fig. 4.4). Again, it turns out that all the components of the maximizing $\boldsymbol{\delta}$ are finite as long as the features and potentials are non-degenerate (see Section 4.3). As before, we can prove the convergence of PLUMMET, and thus also of ℓ_1-PLUMMET.

Theorem 4.7. *PLUMMET produces a sequence $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots$ for which*

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

Proof. The proof mostly follows the same lines as the proof of Theorem 4.5. Here we sketch the main differences.

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

$$F_t = \sup_{\boldsymbol{\delta}} \sum_{j \in \mathcal{J}} \Bigl[-\mathbf{E}_{q_t}[f_j](e^{\delta_j} - 1) - \mathrm{U}_j^*(-\lambda_{t,j} - \delta_j) + \mathrm{U}_j^*(-\lambda_{t,j}) \Bigr] \ .$$

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

duality:

$$F_t = \sup_{\boldsymbol{\delta}} \sum_{j \in \mathcal{J}} \Big[-\mathbf{E}_{q_t}[f_j](e^{\delta_j} - 1) - \mathrm{U}_j'^*(-\delta_j) \Big] + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \tag{4.23}$$

$$= \inf_{\boldsymbol{a} \geq \mathbf{0}} \sum_{j \in \mathcal{J}} \Big[\widetilde{\mathrm{D}}(a_j \parallel \mathbf{E}_{q_t}[f_j]) + \mathrm{U}_j'(a_j) \Big] + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \tag{4.24}$$

$$= \inf_{\boldsymbol{a} \geq \mathbf{0}} \Big[\widetilde{\mathrm{D}}(\boldsymbol{a} \parallel \mathbf{E}_{q_t}[\boldsymbol{f}]) + \mathrm{U}(\boldsymbol{a}) + \boldsymbol{\lambda}_t \cdot \boldsymbol{a} + \mathrm{U}^*(-\boldsymbol{\lambda}_t) \Big] \ . \tag{4.25}$$

In Eq. (4.23) we write $\mathrm{U}_j'^*(u)$ for $\mathrm{U}_j^*(u - \lambda_{t,j})$. In Eq. (4.24) we use Theorem 2.3, noting that the conjugate of $u_0(e^u - 1)$ is the unnormalized relative entropy (see p. 29). In Eq. (4.25) we convert U_j' back into U_j and take the sum over j. Note that $\widetilde{\mathrm{D}}(\boldsymbol{a} \parallel \mathbf{E}_{q_t}[\boldsymbol{f}])$ increases without bound if $\|\boldsymbol{a}\|_\infty \to \infty$ and, by Fenchel's inequality,

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

so in Eq. (4.25) it suffices to take an infimum over the $\boldsymbol{a}$'s of bounded norm, i.e., over a compact set. By lower semi-continuity we thus obtain that the infimum is attained at some point $\boldsymbol{a}_t$ and

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

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

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

is an auxiliary function. Noting that $A(\boldsymbol{\lambda}_t, \boldsymbol{a}_t) = F_t \to 0$ and using Lemma 4.4 yields the result. ■

4.3 Ensuring Finite Updates

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

4.3.1 Non-degeneracy in SUMMET

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

ple space $\mathcal{X}$ and the feature set as follows:

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

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

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

Theorem 4.8. *Let $\boldsymbol{\lambda}$ and $Q(\boldsymbol{\lambda})$ be finite and f_j, U_j non-degenerate. Then $F_j(\boldsymbol{\lambda}, \delta)$ is maximized by a finite δ.*

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

$$\begin{aligned} F_j(\boldsymbol{\lambda}, \delta) &= -\ln\big(1 + (e^{\delta} - 1)\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]\big) - \mathrm{U}_j^*(-\lambda_j - \delta) + \mathrm{U}_j^*(-\lambda_j) \\ &= -\ln\left\{e^{\delta}\left[e^{-\delta}(1 - \mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]) + \mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]\right]\right\} + \delta\mathbf{E}_r[f_j] - \mathrm{U}_{r,j}^*(\lambda_j + \delta) + \mathrm{U}_{r,j}^*(\lambda_j) \\ &= -\ln\left[e^{-\delta}(1 - \mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]) + \mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j]\right] - \delta(1 - \mathbf{E}_r[f_j]) - \mathrm{U}_{r,j}^*(\lambda_j + \delta) + \mathrm{U}_{r,j}^*(\lambda_j). \end{aligned} \tag{4.26}$$

Suppose that $\mathbf{E}_r[f_j] < 1$. Then $F_j(\boldsymbol{\lambda}, \delta) \to -\infty$: the first term of (4.26) is bounded above by $-\ln(\mathbf{E}_{q_{\boldsymbol{\lambda}}}[f_j])$ which is finite by the non-degeneracy of f_j; the second term decreases without bound; the third term is bounded above by $\mathrm{U}_{r,j}(0)$ by Fenchel's inequality; and the fourth term is a finite constant because $Q(\boldsymbol{\lambda})$ is finite. In the case $\mathbf{E}_r[f_j] = 1$, the second term equals zero, but the third term decreases without bound because, by the non-degeneracy of U_j, there exists $\varepsilon > 0$ such that $\mathrm{U}_{r,j}(\varepsilon) = \mathrm{U}_j(1 - \varepsilon) < \infty$ and hence by Fenchel's inequality $-\mathrm{U}_{r,j}^*(\lambda_j + \delta) \le -(\lambda_j + \delta)\varepsilon + \mathrm{U}_{r,j}(\varepsilon)$.

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

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

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

Corollary 4.9. *Updates of SUMMET are always finite.*

Proof. We proceed by induction. In the first step, both $\boldsymbol{\lambda}_1$ and $Q(\boldsymbol{\lambda}_1)$ are finite (see proof of Theorem 4.5). Now suppose that in step t, $\boldsymbol{\lambda}_t$ and $Q(\boldsymbol{\lambda}_t)$ are finite. Then by Theorem 4.8, all considered coordinate updates will be finite, so $\boldsymbol{\lambda}_{t+1}$ will be finite too. Since $Q(\boldsymbol{\lambda}_{t+1}) \geq Q(\boldsymbol{\lambda}_t)$ and $Q(\boldsymbol{\lambda})$ is bounded above (see proof of Theorem 4.5), we obtain that $Q(\boldsymbol{\lambda}_{t+1})$ is finite. ■

4.3.2 Non-degeneracy in PLUMMET

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

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

Similarly to SUMMET, this preprocessing derives an equivalent form of the primal. Using analogous reasoning as in Theorem 4.8, we show below that non-degeneracy implies finite updates in PLUMMET.

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

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

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

First, we rewrite F_j as follows:

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

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

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

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

Chapter 5

Modeling Distributions of Species

In this chapter we put to use the theory and algorithms developed in the preceding chapters. We turn to the application of this dissertation: the problem of modeling geographic distributions of species.

Species-distribution modeling is a critical topic in ecology and conservation biology: to protect a threatened species, one first needs to know its environmental requirements, i.e., its *ecological niche* (Hutchinson, 1957). The ecological niche determines the *potential distribution* of the species (Anderson and Martínez-Meyer, 2004; Phillips et al., 2006), i.e., the set of locations where the species could persist or where conditions may become suitable under the future climate (Hannah et al., 2005). Further applications include predicting the spread of invasive species and infectious diseases (Welk et al., 2002; Peterson and Shaw, 2003), as well as understanding ecological processes such as speciation (Graham et al., 2006), or identifying areas of regional endemism (Raxworthy et al., 2003).

As mentioned earlier, the input for species-distribution modeling consists of a list of occurrences and data on a number of environmental variables. The most basic goal is to predict which areas within a region of interest are within the species' potential distribution. The potential distribution can then be used to estimate the species' *realized distribution*, for example by removing areas where the species is known to be absent because of deforestation or other habitat destruction. Although a species' realized distribution may exhibit some spatial correlation, the potential distribution does not, so considering spatial correlation is not necessarily desirable during species distribution modeling. In our approach, we therefore do not enforce or make use of spatial correlation, and derive maxent constraints based on the environmental variables only.

Quite a number of approaches have been suggested for species distribution modeling, including neural nets, nearest neighbors, genetic algorithms, generalized linear models, generalized additive models, bioclimatic envelopes, boosted regression

trees, and more; see Elith (2002) and the NCEAS comparison (Elith, Graham et al., 2006). The NCEAS comparison evaluates an implementation of ℓ_1-regularized maxent (partly developed in this dissertation) as one of a group of twelve methods in the task of modeling species distributions. We will use the data from this comparison and mention some of the results in more detail below. For now, we remark that in the NCEAS comparison, maxent is placed among the best methods alongside boosted regression trees (Leathwick et al., 2006), generalized dissimilarity models (Ferrier et al., 2002) and multivariate adaptive regression splines with the community level selection of basis functions (Moisen and Frescino, 2002; Leathwick et al., 2005). Among these, however, maxent is the only method designed for presence-only data. This is a typical scenario as most of the typical datasets have no information about the *failure* to observe the species at any given location.

In maxent, the distribution of a single species is modeled as an unknown density π over the set of pixels in the study area. This set of pixels is called the *background* and corresponds to the sample space $\mathcal{X}$. Recorded presence localities are samples $x_1, \ldots, x_m$, drawn from $\mathcal{X}$ according to π. The default distribution q_0 is uniform over $\mathcal{X}$.

To understand how π represents the realized distribution of the species, consider the following (idealized) sampling strategy. An observer first picks a random pixel x from the study area, and then records 1 if the species is present at the pixel x, and 0 if the species is absent. If we denote the response variable (presence or absence) as y, then $\pi(x)$ is the conditional probability distribution $p(x \mid y = 1)$, i.e., the probability of the observer being at x, given that the species is present. According to Bayes' rule, π is proportional to the species' probability of occurrence, $p(y = 1 \mid x)$. Indeed,

$$p(y = 1 \mid x) = \frac{p(x \mid y = 1)p(y = 1)}{p(x)}$$

where $p(x)$ equals $1/|\mathcal{X}|$ for all x, and $p(y = 1)$ is the prevalence of the species in the study area. However, if we have only presence-only data, we cannot determine the species' prevalence (Phillips et al., 2006). Therefore, instead of modeling $p(y = 1 \mid x)$, maxent models the distribution $p(x \mid y = 1)$. In this respect, maxent differs from other statistical approaches used in species distribution modeling, such as generalized additive models, boosted regression trees and multivariate adaptive regression splines mentioned above, which are based on logistic regression, and therefore require binary training data. Under the lack of absence data, these approaches use surrogate absences, for example drawn uniformly at random from $\mathcal{X}$, and instead of $p(y = 1 \mid x)$ they estimate its monotone approximation, $p(s = 1 \mid x)$, where s is a surrogate for the response y (Phillips, Dudík et al., 2007).

In this chapter, we focus exclusively on ℓ_1-regularized maxent. We begin with a set of preliminary experiments evaluating maxent in the task of estimating distributions of four bird species in North America. Specifically, we analyze how the choice of the feature set influences the predictive accuracy of maxent, depending on the number of occurrence records of the modeled species. We also explore the effects of regularization on predictive accuracy and interpretability of the maxent models.

Using the insights from the preliminary experiments, we then carry out a comprehensive evaluation on a large and diverse dataset, consisting of 226 species from six regions. This data was developed by a working group at the National Center for Ecological Analysis and Synthesis (NCEAS) as part of the previously mentioned large-scale comparison of species-distribution modeling methods (Elith, Graham et al., 2006). We refer to the data as "the NCEAS data," and the comparison of methods as "the NCEAS comparison." The NCEAS data consists of two independent datasets: the training dataset, with the data of low quality typical of many applications; and the evaluation dataset, obtained by rigorously planned surveys. We optimize the performance of maxent by tuning the feature-set and regularization settings on a small portion of the training data. The models are then constructed from all of the training data and evaluated on the evaluation data. We compare the maxent performance with the performance of several techniques included in the NCEAS comparison.

Careful tuning of the feature-set and regularization settings of maxent on the NCEAS data has an additional goal: determining well-performing "default settings." Default settings are desirable because parameter tuning may be prohibitively time-consuming to do separately for each species, or unreliable for small or biased datasets. Additionally, even with the abundance of good quality data, users interested in the application of species models may not have the statistical knowledge required for detailed tuning. To assess the quality of the settings determined from the NCEAS training data, we compare the performance of these settings with the optimal performance of the settings tuned on the evaluation data itself.

5.1 Maxent Implementation

As mentioned in Section 2.2, maxent uses features derived from environmental variables of two types: (i) *continuous* and (ii) *categorical*. Continuous variables take arbitrary real values corresponding to measured quantities such as altitude, annual precipitation, and maximum summer temperature. Categorical variables take only a limited number of values (typically 2–20) such as soil type or vegetation type. When a categorical variable quantifies the degree of some property (on a discrete scale), it can also be viewed as a continuous variable, for example, soil fertility. This type of

categorical variable is referred to as *discrete ordinal*. We will typically view discrete ordinal variables as continuous and point out whenever this is not the case.

In our experiments, we used the *Maxent*[1] software for species habitat modeling (Phillips, Dudík, and Schapire, 2007). The software implements ℓ_1-SUMMET, described in Chapter 4, with six feature classes: *linear* (L), *quadratic* (Q), *product* (P), *threshold* (T), *hinge* (H), and *categorical indicator* (C) features.

For a given set of environmental variables, it is possible to choose many combinations of feature classes. Combinations commonly used in *Maxent* are LC, LQC, (L)HC, (L)QHC, TC, and (L)QPHTC. Note that linear features, when scaled to take values in the interval $[0,1]$, are special cases of hinge features, so it is redundant to use L and H features simultaneously (that is why we have placed them in parentheses above).

5.2 Performance Measures

A natural measure of performance applicable to maxent is log loss, introduced in Section 2.1. Up to a constant, it equals the negative log likelihood of the test data, i.e., the sum of the negative log probabilities that the maxent model assigns to test localities. Smaller values correspond to better prediction (higher likelihood). The default distribution, in our case uniform, achieves a zero log loss. The true distribution π achieves the minimum log loss, equal to the negative of its relative entropy from the default: $-\mathrm{D}(\pi \parallel q_0)$.

Another performance measure, applicable to *any* species-distribution modeling method, is the *area under the ROC curve (AUC)* (Hanley and McNeil, 1982), which uses a binary-labeled test set to measure the quality of a ranking of map cells. Specifically, the AUC is the probability that a randomly chosen test positive will be ranked above a randomly chosen test negative. In a test set containing both presences and absences, the presences are positives and the absences are negatives. A random ranking (as well as a uniform ranking) has on average an AUC of 0.5, whereas a perfect ranking achieves the best possible AUC of 1.0. Models with AUC values above 0.75 are considered potentially useful (Elith, 2002).

It is also possible to use ROC curves with presence-only test data (Phillips et al., 2006). In that case we interpret as negatives all grid cells with no occurrence localities, even if they support good environmental conditions for the species. The maximum AUC is therefore less than one (Wiley et al., 2003), and is smaller for wider-ranging species.

[1] We italicize "Maxent" to distinguish the name of the software from the abbreviation "maxent," which we have used throughout this dissertation for "maximum-entropy density estimation."

5.3 Preliminary Experiments

5.3.1 Data and Experimental Design

In our first set of experiments we use *Maxent* to model distributions of bird species, based on occurrence records in the North American Breeding Bird Survey (Sauer et al., 2001), an extensive dataset consisting of thousands of occurrence localities for North American birds and used previously for species distribution modeling (Peterson, 2001). A preliminary version of these experiments and others was evaluated by Phillips, Dudík, and Schapire (2004).

We selected four species with a varying number of occurrence records: Hutton's Vireo (198 occurrences), Blue-headed Vireo (973 occurrences), Yellow-throated Vireo (1611 occurrences) and Loggerhead Shrike (1850 occurrences). The occurrence data of each species was divided into ten random partitions: in each partition, 50% of the occurrence localities were randomly selected for the training set, while the remaining 50% were set aside for testing. The environmental variables use a North American grid with 0.2 degree square cells. We used seven continuous environmental variables: elevation, aspect, slope, annual precipitation, number of wet days, average daily temperature and temperature range. The first three derive from a digital elevation model for North America USGS (2001), and the remaining four were interpolated from weather station readings (New et al., 1999). Each environmental variable is defined over a 386 × 286 grid, of which 58,065 points have data for all environmental variables. We used linear, quadratic, product, and threshold features. The remaining feature types available in *Maxent* (hinge features and categorical indicators) will be explored in the following sections.

Motivated by Theorem 3.4, we reduced the β_j's to a single regularization parameter β_0 by using $\beta_j = \beta_0\sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j]/m}$. According to the bounds of Sections 3.2.1 and 3.4, we expect that β_0 will depend on the number and complexity of features. Therefore, we expect that different values of β_0 will be optimal for different combinations of the feature types.

On each training set, we ran maxent with four different subsets of the feature types: L, LQ, LQP, and T. We ran two types of experiments. First, we ran maxent on increasing subsets of the training data and evaluated log loss on the test data. We took an average over ten partitions and plotted the log loss as a function of the number of training examples. These plots are referred to as learning curves, or m-curves, because they plot the performance as a function of the number of training examples m. Second, we also varied the regularization parameter β_0 and plotted the log loss for fixed numbers of training examples as functions of β_0. These curves are referred to as sensitivity curves, or β-curves, because they plot the performance as a

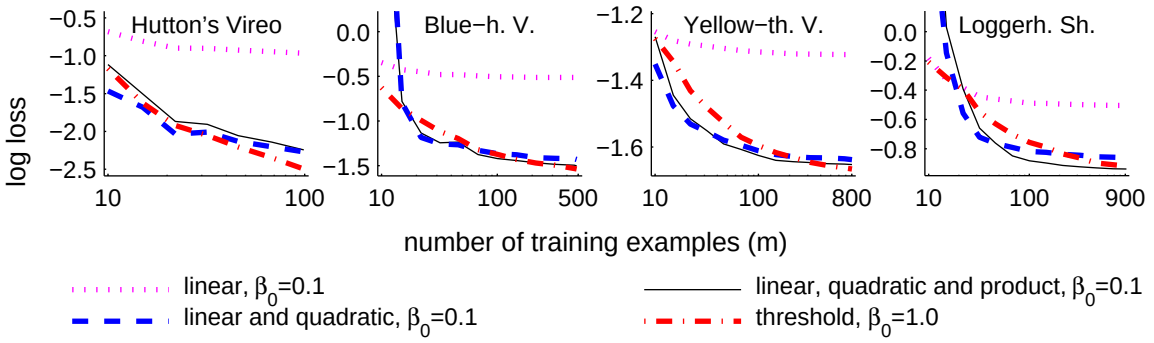


Figure 5.1. *Learning curves.* Log loss averaged over 10 partitions as a function of the number of training examples. Numbers of training examples are plotted on a logarithmic scale.

function of the regularization parameters β_j.

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

5.3.2 Results

Fig. 5.1 shows learning curves for the four studied species. We set $\beta_0 = 0.1$ in L, LQ and LQP runs and $\beta_0 = 1.0$ in T runs. This choice is justified by the sensitivity curve experiments described below. In all cases, the performance improves as more samples become available. This is especially striking in the case of threshold features. In the absence of regularization, maxent would exactly fit the training data with delta functions around sample values of the environmental variables which would result in severe overfitting even when the number of training examples is large. As the learning curves show, regularized maxent does not exhibit this behavior.

Note the heavy overfitting of LQ and LQP features on the smallest sample sizes of Blue-headed Vireo and Loggerhead Shrike. A more detailed analysis of the sensitivity curves suggests that this overfitting could be alleviated by using larger values of β_0, resulting in curves qualitatively similar to those of other species. Similarly, performance of linear features, especially for larger feature sizes, could be somewhat improved using smaller regularization values. Such fine-tuning will be explored in Sections 5.5 and 5.7.

Fig. 5.2 shows the sensitivity of maxent to the regularization value β_0. Note the remarkably consistent minimum at $\beta_0 \approx 1.0$ for threshold feature curves across different species, especially for larger sample sizes. It suggests that for the purposes of ℓ_1 regularization, $\sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j]/m}$ are good estimates of $|\mathbf{E}_{\tilde{\pi}}[f_j] - \mathbf{E}_{\pi}[f_j]|$ for threshold features. For L, LQ, and LQP runs, the minima are much less pronounced as the

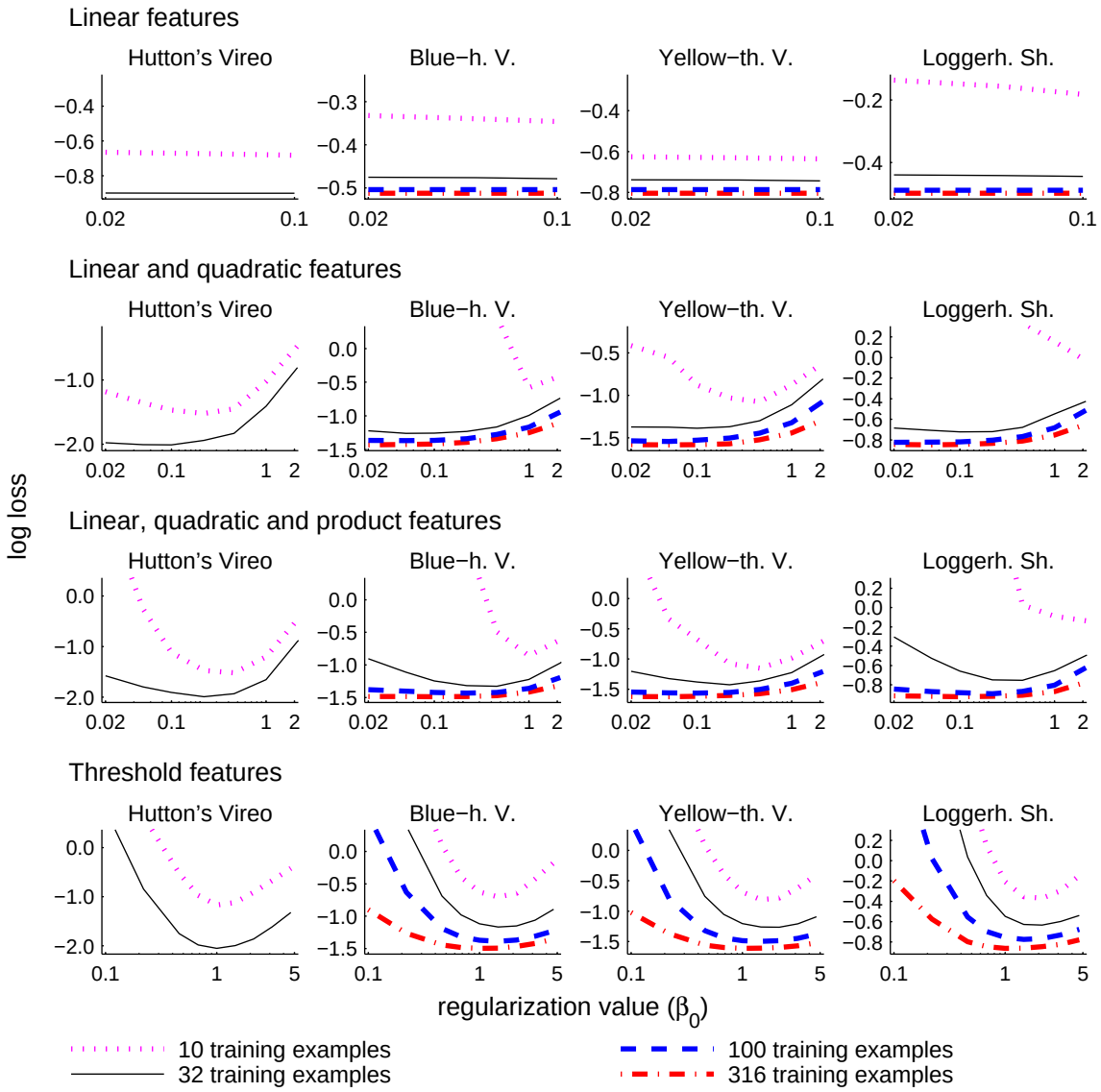


Figure 5.2. *Sensitivity curves.* Log loss averaged over 10 partitions as a function of β_0 for a varying number of training examples. For a fixed value of β_0, maxent finds better solutions (with smaller log loss) as the number of examples grows. Values of β_0 are plotted on a log scale.

number of samples increases and do not appear at the same value of β_0 across different species nor for different sizes of the same species. Benefits of regularization in L, LQ, and LQP runs diminish as the number of training examples increases. One possible explanation is that the relatively small number of features (compared with threshold features) prevents overfitting for large training sets.

To derive feature profiles, recall that maxent with a uniform default distribution returns the Gibbs distribution $q_{\boldsymbol{\lambda}}(x) \propto e^{\boldsymbol{\lambda} \cdot \boldsymbol{f}(x)}$ minimizing the regularized log loss. For L, LQ, and T runs, the exponent is additive in contributions of individual envi-

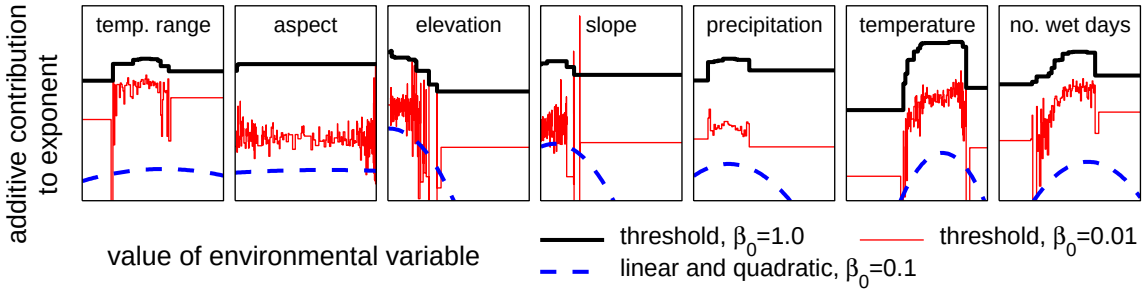


Figure 5.3. *Feature profiles learned on the first partition of the Yellow-throated Vireo.* For every environmental variable, its additive contribution to the exponent of the Gibbs distribution is given as a function of its value. Profiles have been shifted for clarity. This corresponds to adding a constant in the exponent, which has no effect on the resulting models since constants in the exponent cancel out with the normalization factor.

ronmental variables. Plotting this contribution as a function of the corresponding environmental variable we obtain feature profiles for the respective variables. Note that adding a constant to a profile has no impact on the resulting distribution as constants in the exponent cancel out with the normalization. For L models profiles are linear functions, for LQ models profiles are quadratic functions, and for T models profiles can be arbitrary piecewise constant functions. These profiles provide an easier to understand characterization of the distribution than the vector $\boldsymbol{\lambda}$.

Fig. 5.3 shows feature profiles for an LQ run on the first partition of the yellow-throated vireo and two T runs with different values of β_0. The value of $\beta_0 = 0.01$ only prevents components of $\boldsymbol{\lambda}$ from becoming extremely large, but it does little to prevent heavy overfitting with numerous peaks capturing single training examples. Raising β_0 to 1.0 completely eliminates these peaks. This is especially prominent for the aspect variable where the regularized T as well as the LQ model show no dependence, while the insufficiently regularized T model overfits heavily. Note the rough agreement between LQ profiles and regularized T profiles. Peaks in these profiles can be interpreted as intervals of environmental conditions favored by a species. However, such interpretations should be made with caution because the objective of maxent is based solely on the predictive performance. To see why this could be problematic, consider two strongly correlated environmental variables, only one of which has a causal effect on the species. Maxent has no knowledge which of the two variables is truly relevant, and may easily pick the wrong one, leaving the profile of the relevant one flat. Thus, interpretability is affected by correlations between variables.

Table 5.1. *Regions of the NCEAS dataset.*

Code	Region	Environmental variables	Species
AWT	Australian wet tropics	13 continuous	40
CAN	Canada	10 continuous, 1 categorical	20
NSW	North-east New South Wales	12 continuous, 1 categorical	54
NZ	New Zealand	13 continuous	52
SA	South America	11 continuous	30
SWI	Switzerland	13 continuous	30

5.4 The NCEAS Data

The NCEAS data consists of two independent datasets for 226 species from 6 regions of the world. The first dataset contains presence-only data, i.e., a set of geographic coordinates of recorded presence localities for each species together with a set of environmental variables for each of the 6 regions (see Table 5.1). The number of presence localities per species ranges from 2 to 5822, with a median of 57. These presence-only data constitute the training data used to make the models. The second data set is presence-absence evaluation data: for each species, the evaluation data contains a set of localities of confirmed presence and a set of localities of confirmed absence. The number of test sites (presence and absence combined) ranges from 102 to 19120. The presence-only localities are derived from museum and herbarium-type collections, while the presence-absence data are derived from rigorous surveys that sample across both environmental and geographic space. For more details, see Elith, Graham et al. (2006).

5.5 Tuning *Maxent* on the NCEAS Training Data

In the preliminary experiments of Section 5.3, we saw that the choice of the feature set and the regularization parameters influences the predictive performance of maxent. In this section, we describe tuning of regularization parameters and selection of the best-performing feature sets. The regularization parameters and feature sets will be determined using solely the presence-only training data. The resulting configuration (except for the hinge features, added later, but explored here) was used to construct the maxent models in the NCEAS comparison. The configuration (including hinge features) corresponds to the default settings of *Maxent* versions 1.8.3 through the time of writing (at least version 2.3.4).

Table 5.2. *Species used for tuning* Maxent *settings.* The fourth column gives the number of occurrence records in the presence-only dataset that was used for parameter tuning. The last column identifies lists the tuning experiments in which the species was included. Experiments of Sections 5.5.2, 5.5.3, 5.5.4, and 5.5.5 are referred to as *Reg*, *Cat*, *Ord*, and *Opt*.

Region/code	Species	Group	#PO	Experiments
AWT/*ausrob*	*Austrochaperina robusta*	frog	193	*Reg* L, LQ, LQP, T; *Opt*
AWT/*bhe*	*Lichenostomus frenatus*	bird	351	*Opt*
AWT/*coporn*	*Cophixalus ornatus*	frog	337	*Opt*
AWT/*cryliv*	*Cryptocarya lividula*	plant	44	*Opt*
AWT/*ghr*	*Heteromyias albispecularis*	bird	484	*Opt*
AWT/*guiacu*	*Guioa acutifolia*	plant	56	*Reg* L, LQ, LQP; *Opt*
AWT/*lamcog*	*Lampropholis coggeri*	reptile	165	*Opt*
AWT/*sapbas*	*Saproscincus basiliscus*	reptile	177	*Opt*
CAN/*amcr*	American Crow	bird	483	*Cat*; *Opt*
CAN/*cogr*	Common Grackle	bird	721	*Reg* L, LQ, LQP, C, T; *Cat*; *Opt*
CAN/*eato*	Eastern Towee	bird	119	*Cat*; *Opt*
CAN/*gcki*	Golden Crowned Kinglet	bird	18	*Cat*; *Opt*
CAN/*hosp*	House Sparrow	bird	615	*Cat*; *Opt*
CAN/*inbu*	Indigo Bunting	bird	138	*Reg* L, LQ, LQP, C; *Cat*; *Opt*
CAN/*modo*	Mourning Dove	bird	749	*Cat*; *Opt*
CAN/*wtsp*	White-throated Sparrow	bird	313	*Cat*; *Opt*
NSW/*basp2*	*Falsistrellus tasmaniensis*	mammal	28	*Cat*; *Opt*
NSW/*dbsp2*	*Calyptorhynchus lathami*	bird	426	*Reg* L, LQ, LQP, C, T; *Cat*; *Opt*
NSW/*dbsp7*	*Myzomela sanguinolenta*	bird	315	*Cat*; *Ord*; *Opt*
NSW/*nbsp2*	*Tyto tenebricosa*	bird	120	*Cat*; *Ord*; *Opt*
NSW/*otsp7*	*Eucalyptus campanulata*	plant	69	*Cat*; *Ord*; *Opt*
NSW/*rusp2*	*Cyathea leichhardtiana*	plant	42	*Cat*; *Opt*
NSW/*srsp5*	*Eulamprus murrayi*	reptile	186	*Cat*; *Opt*
NSW/*srsp7*	*Pseudechis porphyricaus*	reptile	118	*Reg* L, LQ, LQP, C; *Cat*; *Opt*
NZ/*clefor*	*Clematis forsteri*	plant	36	*Opt*
NZ/*coppro*	*Coprosma propinqua*	plant	40	*Opt*
NZ/*drauni*	*Dracophyllum uniflorum*	plant	174	*Reg* L, LQ, LQP, T; *Ord*; *Opt*
NZ/*libbid*	*Libocedrus bidwillii*	plant	105	*Opt*
NZ/*metper*	*Metrosideros perforata*	plant	87	*Opt*
NZ/*metrob*	*Metrosideros robusta*	plant	48	*Opt*
NZ/*phyalp*	*Phyllocladus alpinus*	plant	211	*Opt*
NZ/*prutax*	*Prumnopitys taxifolia*	plant	130	*Reg* L, LQ, LQP; *Ord*; *Opt*
SA/*amphpani*	*Amphilophium paniculatum*	plant	88	*Reg* L, LQ, LQP, T; *Opt*
SA/*arrabrac*	*Arrabidaea brachypoda*	plant	203	*Reg* L, LQ, LQP; *Opt*
SA/*arracinn*	*Arrabidaea cinnomomea*	plant	49	*Opt*
SA/*cydiaequ*	*Cydista aequinoctialis*	plant	138	*Opt*
SA/*distmagn*	*Distictella magnoliifolia*	plant	81	*Opt*
SA/*fridspec*	*Fridericia speciosa*	plant	57	*Opt*
SA/*lundvirg*	*Lundia virginalis*	plant	36	*Opt*
SA/*parapyra*	*Paragonia pyramidata*	plant	216	*Opt*
SWI/*abialb*	*Abies alba*	plant	3357	*Opt*
SWI/*acepse*	*Acer pseudoplatanus*	plant	2800	*Ord*; *Opt*

Continued on next page...

Table 5.2—Continued.

Region/code	Species	Group	#PO	Experiments
SWI/*betpen*	*Betula pendula*	plant	468	*Opt*
SWI/*fagsyl*	*Fagus sylvatica*	plant	5528	*Reg* L, LQ, LQP, T; *Ord*; *Opt*
SWI/*pincem*	*Pinus cembra*	plant	279	*Reg* L, LQ, LQP; *Ord*; *Opt*
SWI/*pinunc*	*Pinus uncinata*	plant	291	*Opt*
SWI/*poptre*	*Populus tremula*	plant	154	*Opt*
SWI/*pruavi*	*Prunus avium*	plant	613	*Opt*

5.5.1 Data

The environmental variables in the NCEAS dataset cover grids of about 10 million cells. In order to speed up the experiments of this section, we downscaled the resolution of all grids to yield 1.5 million cells. We selected a small subset of species in each region to use for tuning purposes. Our goal was to include a diverse set of biological groups and a wide range of sample sizes. Table 5.2 lists the selected species as well as experiments where they were used (experiments are described below).

In the tuning experiments, we measured the performance of *Maxent* for various parameter settings as follows. We randomly partitioned occurrence records of every species into a training set with 60% of the records and the test set with 40% of the records. We ran *Maxent* on the training set and evaluated its performance on the test set and took the average over 5–10 random partitions (see below). In all the tuning experiments, the performance is measured in terms of log loss and AUC. Since the tuning is done with presence-only data, the AUC values used in tuning are calculated with background data in place of absences.

5.5.2 Tuning Regularization Parameters (*Reg*)

In Section 5.3, we reduced the β_j's to a single regularization parameter β_0 by using $\beta_j = \beta_0\sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j]/m}$. Here, we allow more flexibility and allow β_0 to depend on the feature class. The resulting regularization parameters are called $\beta_{\mathrm{L}}, \beta_{\mathrm{Q}}, \beta_{\mathrm{P}}, \beta_{\mathrm{T}}, \beta_{\mathrm{H}}, \beta_{\mathrm{C}}$.

The goal of the first set of tuning experiments was to determine the values of regularization parameters yielding good performance in all six regions for varying numbers of occurrence records. We ran *Maxent* for different feature-set settings (L, LQ, LQP, T, H, C). For each setting, we assumed a single regularization parameter; in particular, for the LQ setting we kept $\beta_{\mathrm{L}} = \beta_{\mathrm{Q}}$, and for the LQP setting we kept $\beta_{\mathrm{L}} = \beta_{\mathrm{Q}} = \beta_{\mathrm{P}}$.

For each feature class setting, we varied the number of occurrences (by considering nested subsets of the full training set) and the value of the regularization parameter β. The number of occurrences was chosen from the geometrically increasing sequence {6, 10, 17, 30, 55, 100, 178, 316, 1000, 3162}, and the values of β from the

geometrically increasing sequence {0.02, 0.05, 0.10, 0.22, 0.46, 1.0, 2.2, 4.6} chosen to bracket the range of suitable values suggested by the preliminary experiments of Section 5.3. For each number of occurrences, we determined the average performance over five random partitions as a function of β, thus obtaining the β-curves. Peaks of β-curves (minima of log loss curves and maxima of AUC curves) correspond to optimal choices of β for each particular species. Results from the preliminary experiments were used to restrict the set of β's and numbers of occurrences to intervals where the peaks are likely to occur; in particular, L runs were not configured on large sample sizes and LQP runs were not configured on small sample sizes. Also note that C runs were only possible for regions *CAN* and *NSW*, with one categorical variable in each region.

For each feature class setting and number of occurrences, we selected the best β by visual inspection of β-curves. The goal was to choose β performing well in terms of both log loss and AUC on all of the evaluated species. We first excluded curves where *Maxent* was not performing well: log loss β-curves where *Maxent* never reached performance below zero (the log loss of the uniform distribution) and AUC curves that remained below 0.7 (based on the recommendations of Elith (2002) that values above 0.75 are considered potentially useful). On the remaining curves, we used visual inspection to employ two strategies that could be loosely termed as "mean criterion" and "median criterion". According to the former criterion, we chose the value β that was close to the peak of β-curves of as many species as possible. According to the latter criterion, we chose β at which about half of the β-curves were increasing and half were decreasing. The latter criterion was used whenever the peak was not identifiable in β-curves (they were monotone). The optimal values of β for numbers of occurrences not represented in the evaluated sequence were obtained by piecewise linear interpolation.

5.5.3 Combining Continuous and Categorical Variables (*Cat*)

In the initial block of tuning experiments (*Reg*, Section 5.5.2), the regularization parameter for binary indicators β_{C} was determined in *Maxent* runs with a single categorical variable. When categorical variables are used with additional continuous variables, the total number of features increases, so we expect that higher values of β_{C} will yield better performance (see the guarantees of Section 3.2.1). In this set of experiments, we explored a range of alternative settings of β_{C} in runs including L, Q and P features derived from continuous variables.

We carried out LC, LQC and LQPC runs with $\beta_{\mathrm{L}}, \beta_{\mathrm{Q}}, \beta_{\mathrm{P}}$ equal to the previously determined optimum and for three different settings of β_{C}. The first β_{C} setting, $\beta_{\mathrm{C}} = low$, corresponds to the value determined in the initial block of experiments;

the second setting $\beta_C = \beta_{LQP}$ corresponds to using the same regularization for binary indicators as for L, Q and P features; and the third setting $\beta_C = \beta_{LQP}^{1/2} \cdot low^{1/2}$, which equals the geometric average of the previous two settings, corresponds to an intermediate regularization choice.

While being a reasonable intermediate choice, the geometric average setting has a disadvantage in that it depends on whether L, LQ or LQP features are used. In addition, even though it lies between settings *low* and β_{LQP} which are piecewise linear, the geometric setting is not piecewise linear. For simplicity and consistency with the other feature classes, we prefer using a piecewise linear function independent of the feature set used. We therefore included a fourth setting for β_C, namely a piecewise linear setting which we chose to approximate the geometric average setting. We used some prior knowledge and approximated geometric averages for the feature set (one of L, LQ and LQP) which we expected to be the best at each number of occurrences.

For each setting of β_C, we plotted the average performance over 10 partitions as a function of an increasing number of samples from nested subsets of sizes {5, 10, 20, 40, 75, 150, 300, 750, 2000},[2] obtaining m-curves. The best setting of β_C was again chosen by visual inspection of graphs with the goal being to perform well on all evaluated species both in terms of AUC and log loss. In the current block of experiments, we marked all discrete ordinal variables as categorical to obtain a larger number of categorical variables and hence a more reliable tuning of β_C.

5.5.4 Using Discrete Ordinal Variables (*Ord*)

Next, we explored the effect of treating discrete ordinal variables as categorical or continuous. For the former case, we used the optimal β_C determined in the previous experiment. For the latter case, we consider two settings of β_C: the previously determined optimal setting and the baseline setting $\beta_C = \beta_{LQP}$ which uses a single regularization parameter for all features (this was the setting in versions of *Maxent* prior to 1.8.3). The optimal setting is determined by visual inspection of m-curves for LC, LQC and LQPC runs.

5.5.5 Choosing Optimal Feature Sets (*Opt*)

The final goal of the tuning experiments was to decide which sets of feature classes to use for what numbers of species occurrences. We used the previously determined regularization parameters for the LC, LQC, LQPC and LQPTC runs. The optimal

[2]Note that this sequence differs from the sequence used in Section 5.5.2. The rationale behind choosing a different sequence was to evaluate *Maxent* for the training set sizes where the interpolated values of β are used.

Table 5.3. *Regularization parameters determined by tuning* Maxent *on presence-only data.* Values in boldface were determined exactly, values in italics are linearly interpolated or extrapolated, with the exception that the values to the right of the listed ranges remain constant. For binary indicator features, the "low" settings were determined using a single (categorical) variable, while the piecewise linear settings (used for the current version of *Maxent*, version 2.3.4) were chosen to approximate the geometric average of the "low" setting and β_{L}.

	Number of occurrence records					
	0	6	10	17	30	100
Linear features: β_{L}	*1.0*	**1.0**	**1.0**	*0.72*	**0.2**	**0.05**
Linear and quadratic features: $\beta_{\mathrm{L}}, \beta_{\mathrm{Q}}$	*1.3*	**1.0**	**0.8**	**0.5**	**0.25**	**0.05**
Linear, quadratic and product features: $\beta_{\mathrm{L}}, \beta_{\mathrm{Q}}, \beta_{\mathrm{P}}$	*2.6*	**2.0**	**1.6**	**0.9**	**0.55**	**0.05**
Hinge features: β_{H}	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**
Threshold features: β_{T}	*2.0*	*1.94*	**1.9**	*1.83*	*1.7*	**1.0**
Binary indicators, a single categorical variable, "low": β_{C}	*0.2*	**0.2**	**0.2**	**0.1**	**0.05**	*0.05*
Binary indicators, "piecewise linear": β_{C}	**0.65**	**0.53**	**0.45**	**0.25**	**0.15**	**0.05**

ranges for different feature class settings were determined by visual inspection of m-curves. H features were not part of this block of experiments. Their optimal range was determined by an inspection of the β-curves from Section 5.5.2.

5.5.6 Results

By visual inspection of β-curves for L, LQ, LQP, T, H, and C runs, we propose the regularization parameter settings given in the top portion of Table 5.3. Values set in boldface were determined exactly, while the others were obtained from the boldface values by interpolation or extrapolation. Values for occurrence counts below 100 are linearly interpolated between the two closest counts. Values above 100 are kept the same as for 100.[3] Note that for each feature class, the value of β is monotonically non-increasing in the number of occurrence records. For L, Q, and P features, there appears to be a significant decrease in the values of the optimal settings. In other words, model performance is optimized if we use error bounds that decrease in width somewhat faster than the theory suggests (see Section 3.2.1).

To demonstrate how we determined these settings, consider the β-curves for the LQ run (Figure 5.4). First, we exclude log loss plots that are uniformly above zero (the log loss of the uniform distribution) and the AUC plots that are uniformly worse than 0.7 (our cutoff for informative AUC) since they indicate poor fits that cannot

[3]More precisely, values between the highest and lowest bold settings are linearly interpolated between the two closest bold settings. Values above the highest bold settings are kept the same as the highest bold settings, and the values below the lowest bold settings are linearly extrapolated from the two lowest bold settings.

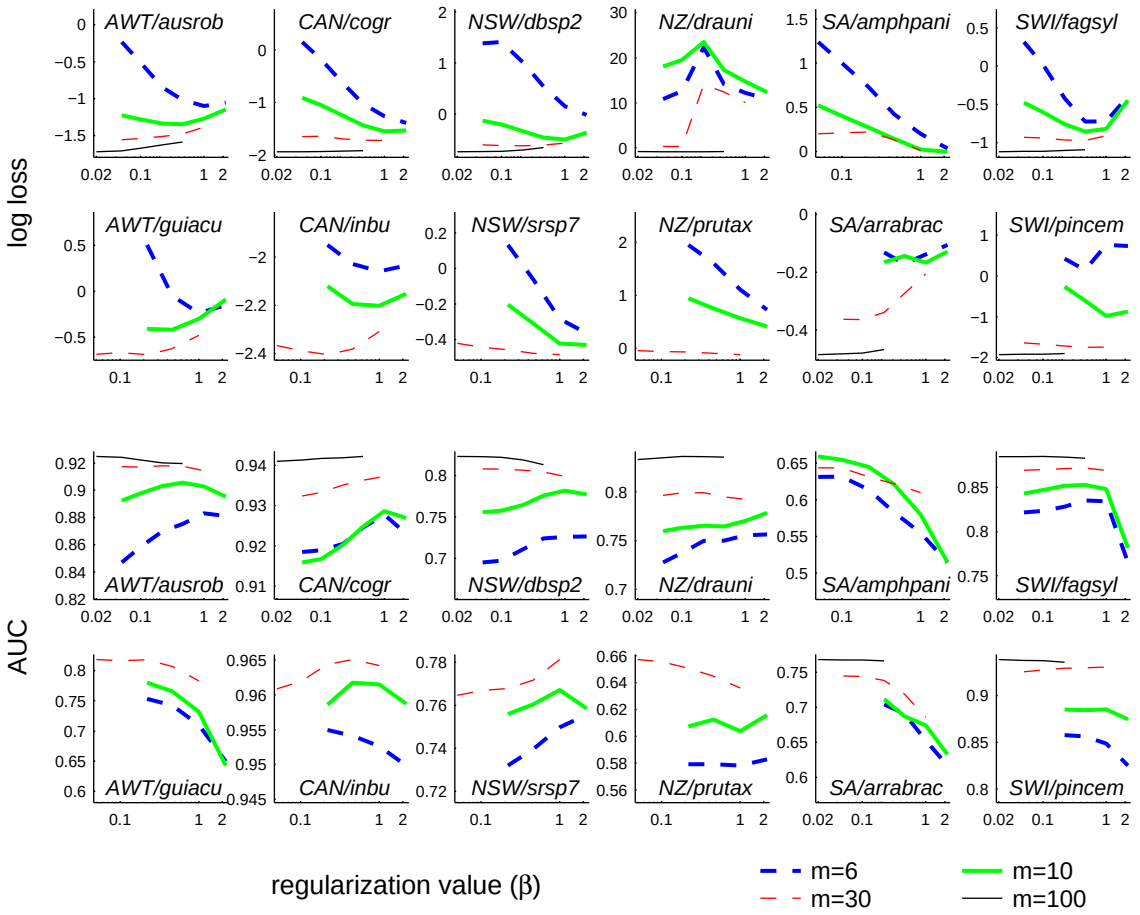


Figure 5.4. *A subset of the β-curves for LQ runs of* Maxent. Performance of *Maxent* is evaluated as a function of the regularization parameter β where $\beta_L = \beta_Q = \beta$. Different curves correspond to different numbers of occurrences m. Performance is evaluated in terms of log loss and AUC. Based on these curves, the regularization parameters β_L, β_Q listed in Table 5.3 were chosen to obtain satisfactory performance on the evaluated subset of species. The resulting values are default settings for *Maxent* versions 1.8.3 through at least 2.3.4.

be mitigated by regularization. Based on the remaining plots, we fill out the line LQ of Table 5.3. For example, for six samples, we determine the optimal setting as $\beta_L = \beta_Q = 1.0$. At this value, nine β-curves reach the peak performance, four β-curves have their peaks to the right of 1.0, and five β-curves have their peaks to the left of 1.0.[4]

When using both continuous and categorical variables, the geometric average set-

[4]The curves with the peak performance at $\beta = 1.0$ are the log loss curves of *AWT*/*ausrob*, *AWT*/*guiacu*, *CAN*/*inbu*, *SWI*/*fagsyl*, and the AUC curves of *AWT*/*ausrob*, *CAN*/*cogr*, *NSW*/*dbsp2*, *NZ*/*drauni*, *SWI*/*fagsyl*. The curves with peaks to the right of 1.0 are the log loss curves of *CAN*/*cogr*, *NSW*/*dbsp2*, *NSW*/*srsp7*, and the AUC curve of *NSW*/*srsp7*. The curves with peaks to the right of 1.0 are the log loss curve of *SA*/*arrabrac* and the AUC curves of *AWT*/*guiacu*, *CAN*/*inbu*, *SA*/*arrabrac*, *SWI*/*pincem*. The remaining curves were excluded.

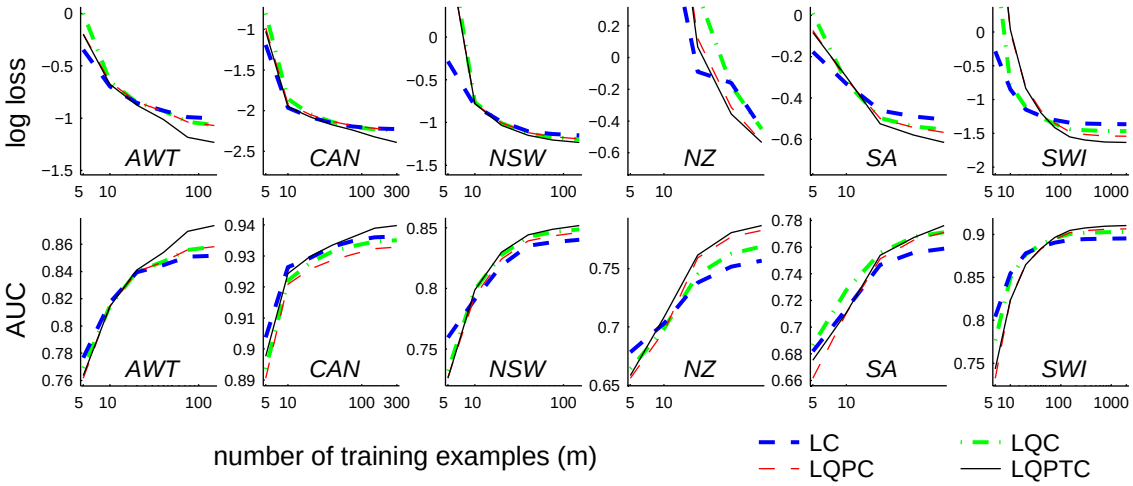


Figure 5.5. *Curves showing performance as a function of sample size (m-curves) averaged over all evaluated species (eight species per region).* Performance is evaluated in terms of log loss and AUC. By visual inspection, we determined the ranges of sample sizes in which to use the different sets of feature classes as: LC features for 2–9 samples, LQC features for 10–79 samples and LQPTC features for 80 and more samples.

ting for β_C gave the best performance in more than half of the m-curves and never gave the worst performance. Piecewise linear settings (Table 5.3) performed similarly to the geometric average settings (plots omitted), and they are therefore used as the settings for β_C in the current version of *Maxent* (versions 1.8.3 through at least version 2.3.4).[5] Discrete ordinal variables perform the best when viewed as continuous (details not shown).

Finally, we determined optimal combinations of feature classes. Figure 5.5 shows the performance of four feature class settings. From the figures, we determined ranges of individual feature classes as follows: LC features for 2–9 samples, LQC features for 10–79 samples, LQPTC features for 80 and more samples. These were the feature classes used in the NCEAS comparison. Hinge features were added afterwards. Their β-curves exhibited good performance already for low numbers of occurrences; therefore their optimal range was determined as 15 and more samples.

5.6 The NCEAS Comparison

In this section, we present a subset of results of the NCEAS comparison. The NCEAS comparison indicated that the twelve evaluated species-distribution modeling methods can be approximately divided according to their performance into three groups.

[5]The geometric average settings were used for the NCEAS comparison, and the piecewise linear settings were added afterwards as a simplification.

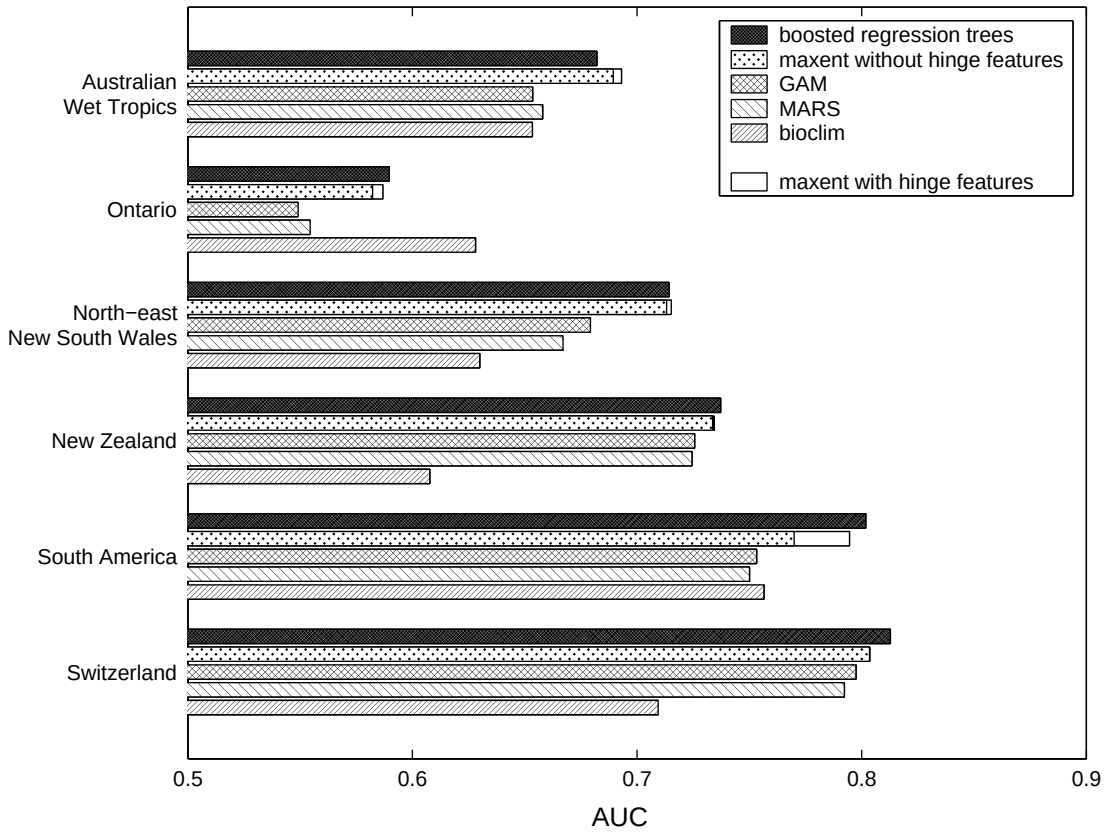


Figure 5.6. *Comparison of* Maxent *and other species distribution modeling techniques.* The results for all the methods except for *Maxent* with hinge features are taken from the NCEAS comparison.

The top performing group consists of methods that allow high expressivity of the models while controlling their complexity. The second group consisted largely of general purpose regression techniques. Finally, the worst-performing group consisted of methods that ignored the characteristics of the studied region and worked solely with the presences.

In Fig. 5.6, we report results for ℓ_1-regularized maxent and boosted regression trees (BRT) from the top group; generalized adaptive models (GAM) and multivariate adaptive regression splines (MARS) from the second group; and BIOCLIM from the third group. BRT, GAM, and MARS are based on logistic regression. BIOCLIM views presences as points in the environmental space and estimates the species distribution by fitting an envelope around the presences.

The results indicate that *Maxent* is comparable with BRT, and outperforms the remaining techniques, with the exception of the region Ontario. In the latter region, the poor performance of all the techniques, except BIOCLIM, can be explained by a large amount of sample selection bias. We will return to this example in Chapter 6.

5.7 Evaluating the *Maxent* Tuning

5.7.1 Experimental Design

We evaluate the performance of *Maxent* as tuned in Section 5.5 using the presence-absence (evaluation) portion of the NCEAS data. The goal of this evaluation is to compare the tuned settings of regularization parameters (based solely on presence-only data) with the best possible settings for the given evaluation data. To obtain the best possible settings, we tune regularization parameters to yield the best performance when models are trained on the presence-only dataset and evaluated on the presence-absence dataset. We call the latter parameter settings "pa-tuned", in contrast to the "po-tuned" values determined in Section 5.5.

In po-tuning of Section 5.5, a single set of regularization parameters is applied across all regions. However, it is conceivable that different sets of parameters may be appropriate for different regions, and better performance is obtained by tuning the parameters for each region separately. In pa-tuning, we therefore distinguished two cases: regional tuning and global tuning. In the former case, a separate set of regularization parameters is chosen to maximize the average AUC of the species in the relevant region only. In the latter case, a single set of regularization parameters is chosen to maximize the average AUC across all species.[6]

Sets of pa-tuned regularization parameters were obtained by a local search. We tried to optimize the value of the AUC on the presence-absence data by making incremental changes in parameters. We began with the po-tuned parameter values and then cyclically iterated through feature classes, trying to increase or decrease the corresponding regularization parameter. This was repeated until no changes in parameters yielded an improvement. We considered multiplicative changes in regularization parameters by a factor of $\sqrt{2}$ or $1/\sqrt{2}$. We allowed at most an 8-fold increase or decrease relative to the po-tuned parameter setting. Each feature class was applied in the same range of numbers of occurrences as determined by po-tuning in Section 5.5.5. To speed up tuning, we replaced the background (consisting of 1.5 million cells even after downsampling), by a random sample of 10,000 cells. Since the running time of *Maxent* depends linearly on the number of background points, the subsampling results in a significant speed-up while exhibiting almost no deterioration of the predictive performance (Phillips and Dudík, 2007).

Table 5.4. *The AUC performance of* Maxent *with globally and regionally optimized parameters.*

	default settings	improvement from default: globally optimized settings	improvement from default: regionally optimized settings
AWT	0.693	0.004	0.015
CAN	0.594	0.008	0.023
NSW	0.711	0.005	0.022
NZ	0.733	0.008	0.009
SA	0.796	0.007	0.014
SWI	0.803	0.003	0.001
all species	0.726	0.006	0.014

5.7.2 Results

In Table 5.4, we compare performance of po-tuned *Maxent* parameters and pa-tuned parameters. Global pa-tuning results in an improvement of average AUC by 0.006 whereas regional pa-tuning improves the AUC by 0.014.

In Table 5.5, we report parameters obtained by global pa-tuning and medians of parameters obtained by regional pa-tuning. For regions *AWT* and *NSW*, the regional tuning was performed separately for each taxonomic group (birds and plants in *AWT*,[7] and small mammals, reptiles, birds and plants in *NSW*), resulting in a total of 10 regionally optimized parameter sets. Each median is thus taken over a set of 10 values.

To compare pa-tuned regularization parameters with po-tuned regularization parameters, we determined the median training set size in each range and report the corresponding po-tuned values. Note that the pa-tuned values are almost always larger than the po-tuned values. Larger regularization represents increased uncertainty in feature-expectation estimates as a result of differences between training and test distributions. This is in line with the intuition that if training and test distributions differ, it is preferable to predict distributions that are more spread-out.

According to Table 5.4 the benefits of pa-tuning seem fairly small. They are of similar magnitude as the within-group differences in the NCEAS comparison. It is

[6]Note that global tuning puts a somewhat larger weight on regions with more species.

[7]An early version of the NCEAS data contained additional taxonomic groups in region *AWT*, see Table 5.2, which were excluded in pa-tuning because of low data quality. Since the quality of po-data for these groups was similar to that of the remaining groups, they were included in po-tuning to obtain a more diverse dataset.

Table 5.5. *Overview of pa-tuned parameters: globally optimized parameters and medians of 10 regionally optimized parameters.* The global settings optimize the average performance across all species. The regional settings optimize the performance separately for each of 10 taxonomic groups in the 6 regions.

		number of occurrences			
		2–9	10–14	15–79	≥ 80
β_L:	global optimum	1.00	1.41		
	regional median	1.00	1.00		
	default*	1.00	.71		
β_Q:	global optimum		1.41	.50	.35†
	regional median		1.00	.85	.05
	default*		.71	.23	.05
β_P:	global optimum				.35†
	regional median				.04
	default*				.05
β_T:	global optimum				2.00
	regional median				1.21
	default*				1.00
β_H:	global optimum			.35	.50
	regional median			.85	.50
	default*			.50	.50
β_C:	global optimum	1.41	.50	.03**	.03
	regional median	.71	.50	.18	.04
	default*	.53	.39	.14	.05

* po-tuned values for the median training-set size in each range: 6, 12, 36, and 221
† the largest possible value in local search
** the smallest possible value in local search

tempting to use the pa-tuned settings as default settings in *Maxent*, since they give marginally better performance on the evaluation data. However, doing so may result in overfitting to this particular evaluation dataset, since the pa-tuned settings are being evaluated here on the same data on which they were tuned. Therefore, in the following chapters we use the po-tuned settings, which have been validated on independent test data.

Chapter 6

Biased Density Estimation

In this chapter, we study maxent density estimation under sample selection bias. In density estimation it is very common to assume access to independent samples from the distribution being estimated. In practice, this assumption is violated for various reasons. For example, in species distribution modeling most sampling is done in locations that are easier to access, such as areas close to towns, roads, airports or waterways (Reddy and Dávalos, 2003). This presents a significant sample-selection bias since roads and waterways are often correlated with topography and vegetation which also influence species distributions. New unbiased sampling may be expensive or even impossible, if original landcover has been cleared, so much can be gained by using the extensive existing biased data.

Although the available data may have been collected in a biased manner, we usually have some information about the nature of the bias. In species distribution modeling, some factors influencing the sampling distribution are well known, such as distance from roads, towns, etc. In addition, a list of visited sites may be available and viewed as a sample of the sampling distribution itself. If such a list is not available, the set of sites where any species from a large group has been observed may be a reasonable approximation of all visited locations.

We will assume that the sampling distribution (or an approximation) is known during *training*, but we require that models not use any knowledge of sample selection bias during *testing*. This requirement is vital for species distribution modeling where models are often applied to a different region or under different climatic conditions.

We propose two approaches that incorporate sample selection bias in maximum entropy density estimation. The first approach uses a bias correction technique similar to that of Zadrozny (2004) and Zadrozny et al. (2003) to obtain unbiased confidence regions from biased samples. We prove that, as in the unbiased case, this produces models whose log loss approaches that of the best possible Gibbs distribu-

tion (with increasing sample size).

In contrast, our second approach estimates the biased distribution and then factors the bias out. When the target distribution is a Gibbs distribution, the solution again approaches the log loss of the target distribution. When the target distribution is not Gibbs, we demonstrate that the second approach need not produce the optimal Gibbs distribution (with respect to log loss) even in the limit of infinitely many samples. However, we observe good empirical performance for moderate sample sizes. In addition, the second approach can be easily extended to situations, in which we only have access to samples from the sampling distribution (such as the list of all visited sites in species distribution modeling) instead of the distribution itself.

One of the challenges in studying methods for correcting sample selection bias is that unbiased data sets, though not required during training, are needed as test sets to evaluate performance. Unbiased data sets are difficult to obtain—this is the very reason why we study this problem! Thus, it is almost inevitable that synthetic data must be used. In Section 6.4, we describe synthetic experiments evaluating performance of our two approaches.

In Section 6.5, we consider sample selection bias in the context of species distribution estimation. We evaluate our debiasing approaches on the NCEAS dataset. The NCEAS training data is biased, whereas the NCEAS evaluation data has been collected independently in a reasonably unbiased manner.

A somewhat surprising result of our real-data experiments is that the empirical version of our second approach (factor-bias-out) outperforms other approaches, although its performance guarantees are the weakest. A similar approach has been previously used with regression-based techniques (Ferrier et al., 2002; Zaniewski et al., 2002). We compare the performance of maxent and other methods from the NCEAS comparison using this debiasing approach. Similarly to Section 5.7, we evaluate the effect of the regularization tuning on maxent with bias correction.

Related Work

A traditional field where sample selection bias arises is econometrics. In econometrics, the data from surveys is affected by factors such as attrition, nonresponse and self selection (Heckman, 1979; Groves, 1989; Little and Rubin, 2002). An approach to coping with sample selection bias has been suggested by Heckman (1979) in linear regression. Here the bias is first estimated and then a transform of the estimate is used as an additional regressor.

In the machine learning community, sample selection bias has been recently considered for classification problems by Zadrozny (2004). Here the goal is to learn a decision rule from a biased sample. The problem is closely related to cost-sensitive

learning (Elkan, 2001; Zadrozny et al., 2003) and the same techniques such as resampling or differential weighting of samples apply.

However, the methods of the previous two approaches do not apply directly to density estimation where the setup is "unconditional," i.e., there is no dependent variable, or, in the classification terminology, we only have access to positive examples, and the cost function (log loss) is unbounded. In addition, in the case of modeling species distributions, we face the challenge of sample sizes that are very small (2–100) by machine learning standards.

6.1 Setup for Biased Density Estimation

In biased density estimation, our goal is to estimate the target distribution π, but samples do not come directly from π. For nonnegative functions p_1, p_2 defined on $\mathcal{X}$, let $p_1 p_2$ denote the distribution obtained by multiplying weights $p_1(x)$ and $p_2(x)$ at every point and renormalizing:

$$p_1 p_2(x) = \frac{p_1(x) p_2(x)}{\sum_{x'} p_1(x') p_2(x')} \ .$$

We assume that samples $x_1, \ldots, x_m$ come from the *biased distribution* $\pi\sigma$ where σ is the *sampling distribution*. This setup corresponds to the situation when an event occurs at the point x with probability proportional to $\pi(x)$ while we perform an independent observation with probability $\sigma(x)$. Let y denote a binary response, equal to one if the event (for example, the species presence) is observed and zero if the event is not observed. If we denote the probability under the described sampling scenario as $\mathbf{P}_\sigma$, then $\mathbf{P}_\sigma(x) = \sigma(x)$ and $\mathbf{P}_\sigma(y = 1 \mid x) \propto \pi(x)$. Thus, samples come from the distribution

$$\mathbf{P}_\sigma(x \mid y = 1) \propto \mathbf{P}_\sigma(x) \mathbf{P}_\sigma(y = 1 \mid x) \propto \pi\sigma(x) \ .$$

The empirical distribution of m samples drawn from $\pi\sigma$ will be denoted by $\widetilde{\pi\sigma}$.

We assume that σ is known and strictly positive on $\mathcal{X}$. The smallest and largest sampling density are denoted as $\sigma_{\min} = \min_x \sigma(x)$ and $\sigma_{\max} = \max_x \sigma(x)$. Note that $\sigma_{\min} > 0$ because $\mathcal{X}$ is finite.

6.2 Approach I: Debiasing Averages

Our first approach is based on maxent with an indicator potential corresponding to the confidence region for unbiased averages. Since we do not have direct access to samples from π, we use a version of the Bias Correction Theorem of Zadrozny (2004) to convert expectations with respect to $\pi\sigma$ to expectations with respect to π.

Theorem 6.1 (Bias Correction Theorem, Zadrozny, 2004; first in Zadrozny et al., 2003, as Translation Theorem)**.**

$$\frac{\mathbf{E}_{\pi\sigma}[\boldsymbol{f}/\sigma]}{\mathbf{E}_{\pi\sigma}[1/\sigma]} = \mathbf{E}_{\pi}[\boldsymbol{f}] \ .$$

Proof. Calculate

$$\mathbf{E}_{\pi\sigma}[\boldsymbol{f}/\sigma] = \sum_{x\in\mathcal{X}} \pi\sigma(x)\frac{\boldsymbol{f}(x)}{\sigma(x)} = \sum_{x\in\mathcal{X}} \frac{\pi(x)\sigma(x)}{\mathbf{E}_{\pi}[\sigma]}\frac{\boldsymbol{f}(x)}{\sigma(x)} = \frac{1}{\mathbf{E}_{\pi}[\sigma]}\sum_{x\in\mathcal{X}} \pi(x)\boldsymbol{f}(x) = \frac{\mathbf{E}_{\pi}[\boldsymbol{f}]}{\mathbf{E}_{\pi}[\sigma]} \ .$$

Similarly,

$$\mathbf{E}_{\pi\sigma}[1/\sigma] = \frac{\mathbf{E}_{\pi}[1]}{\mathbf{E}_{\pi}[\sigma]} = \frac{1}{\mathbf{E}_{\pi}[\sigma]} \ .$$

Dividing the two expressions, we obtain the result. ■

Hence, in order to obtain a confidence region for $\mathbf{E}_{\pi}[\boldsymbol{f}]$, it suffices to obtain confidence intervals for $\mathbf{E}_{\pi\sigma}[f_j/\sigma]$, $j \in \mathcal{J}$ and $\mathbf{E}_{\pi\sigma}[1/\sigma]$. Such confidence intervals can be derived from biased empirical averages for example by the Hoeffding or Bernstein inequalities as we saw in Section 3.2.1.

Assume that such intervals are given. Let $[c_j, d_j]$, $j \in \mathcal{J}$ denote confidence intervals for $\mathbf{E}_{\pi\sigma}[f_j/\sigma]$, $j \in \mathcal{J}$, and $[c', d']$ denote a confidence interval for $\mathbf{E}_{\pi\sigma}[1/\sigma]$ such that $c' > 0$ (this is always possible since $\sigma(x) \geq \sigma_{\min}$). If the expectations $\mathbf{E}_{\pi\sigma}[f_j/\sigma]$, $j \in \mathcal{J}$, and $\mathbf{E}_{\pi\sigma}[1/\sigma]$ lie in their corresponding confidence intervals then by Theorem 6.1

$$\frac{c_j}{d'} \leq \mathbf{E}_{\pi}[f_j] \leq \frac{d_j}{c'} \quad \text{for all } j \in \mathcal{J}.$$

This set of constraints defines the box-shaped confidence region used in our first debiasing approach

$$B = \left\{ \boldsymbol{u} \in \mathbb{R}^{\mathcal{J}} : \frac{c_j}{d'} \leq u_j \leq \frac{d_j}{c'} \text{ for all } j \right\} \ . \tag{6.1}$$

The corresponding potential and regularization are defined by

$$\begin{aligned} \mathrm{U}(\boldsymbol{u}) &= \mathrm{I}_B(\boldsymbol{u}) \\ \mathrm{U}^*(-\boldsymbol{\lambda}) &= \mathrm{I}_B^*(-\boldsymbol{\lambda}) = \sum_{j\in\mathcal{J}} \left[-\frac{1}{2}\left(\frac{c_j}{d'} + \frac{d_j}{c'}\right)\lambda_j + \frac{1}{2}\left(\frac{d_j}{c'} - \frac{c_j}{d'}\right)|\lambda_j| \right] \ , \end{aligned} \tag{6.2}$$

where the regularization is derived by Eq. (2.13) from the observation that B is a box centered at $\langle (c_j/d' + d_j/c')/2 \rangle_{j\in\mathcal{J}}$ with width $d_j/c' - c_j/d'$ along the j-th coordinate.

When the intervals $[c_j, d_j]$, $j \in \mathcal{J}$, and $[c', d']$ are derived by Hoeffding's inequality, we obtain the following theorem (similar to Theorem 3.3 for unbiased maxent).

Theorem 6.2. *Assume that features f_j, $j \in \mathcal{J}$ are bounded in $[0,1]$. Let the bias σ be bounded in $[\sigma_{\min}, \sigma_{\max}]$ where $\sigma_{\min} > 0$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ minimize $\ln Z_{\boldsymbol{\lambda}} + \mathrm{I}_B^*(-\boldsymbol{\lambda})$ where $\mathrm{I}_B^*(-\boldsymbol{\lambda})$ is defined by Eq. (6.2) with*

$$c_j = \mathbf{E}_{\widetilde{\pi\sigma}}[f_j/\sigma] - \beta \qquad d_j = \mathbf{E}_{\widetilde{\pi\sigma}}[f_j/\sigma] + \beta$$
$$c' = \max\{1/\sigma_{\max}, \mathbf{E}_{\widetilde{\pi\sigma}}[1/\sigma] - \beta\} \qquad d' = \mathbf{E}_{\widetilde{\pi\sigma}}[1/\sigma] + \beta$$

where

$$\beta = \frac{1}{\sigma_{\min}} \sqrt{\frac{\ln(2(n+1)/\delta)}{2m}} \ .$$

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

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\boldsymbol{\lambda}^\star\|_1 + \sum_{j\in\mathcal{J}} \mathbf{E}_\pi[f_j]|\lambda_j^\star|}{\sqrt{m}} \cdot \frac{\alpha\sigma_{\max}}{\sigma_{\min}} \tag{6.3}$$

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

Moreover, if $m \ge \left(\alpha \mathbf{E}_\pi[\sigma]/\sigma_{\min}\right)^2$ then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\boldsymbol{\lambda}^\star\|_1 + \sum_{j\in\mathcal{J}} \mathbf{E}_\pi[f_j]|\lambda_j^\star|}{\sqrt{m}} \cdot \frac{\alpha\mathbf{E}_\pi[\sigma]}{\sigma_{\min}} \left(1 - \frac{1}{\sqrt{m}} \frac{\alpha\mathbf{E}_\pi[\sigma]}{\sigma_{\min}}\right)^{-1} \ . \tag{6.4}$$

Before proving Theorem 6.2, we discuss how this theorem expresses the "price of bias," i.e., the amount by which the guarantees for the debiasing potential I_B lag behind the box potential $\mathrm{U}^{(1)}$ for unbiased estimation.

Comparing Theorem 6.2 with Theorem 3.3, we note two differences. The first is the additive difference: the ℓ_1-norm of Theorem 3.3 is replaced by the ℓ_1-norm plus the term $\sum_{j\in\mathcal{J}} \mathbf{E}_\pi[f_j]|\lambda_j^\star|$. This suggests that the particular debiasing potential I_B has more problems fitting target distributions that put more probability on high feature values. This seems somewhat artificial and is due to the fact that the width of confidence intervals estimated by I_B for $\mathbf{E}_\pi[f_j]$ and $\mathbf{E}_\pi[f_j']$, where $f_j'(x) = 1 - f_j(x)$, is in general different.[1] A simple approach, short of replacing the potential I_B, is doubling the feature set, using f_j' alongside f_j for every j. This modification will guarantee that the additive difference is at most $\|\boldsymbol{\lambda}^\star\|_1/2$, i.e., the worst-case multiplicative increase (due to the additive term) is by a factor of 1.5.

The second difference is multiplicative. The multiplicative constant of Theorem 3.3 is roughly equal to α, whereas the multiplicative constant of Theorem 6.2 is equal to $\alpha\sigma_{\max}/\sigma_{\min}$ for small m, and approximately equal to $\alpha\mathbf{E}_\pi[\sigma]/\sigma_{\min}$ for

[1] The reason is that $\left|\frac{c_j}{d'} - \frac{d_j}{c'}\right| \ne \left|\frac{e-d_j}{d'} - \frac{e-c_j}{c'}\right|$ for general e; in our case, $e = \mathbf{E}_{\widetilde{\pi\sigma}}[1/\sigma]$.

larger m. These multiplicative constants, always greater than α, reflect the dependence on the bias. Intuitively, this dependence should not be surprising. For small values of m, we cannot distinguish whether undersampling of certain areas of sample space is due to π or due to the sample selection bias; this difficulty is quantified by the ratio $\sigma_{\max}/\sigma_{\min}$. For larger values of m, we get a slightly better ratio of $\mathbf{E}_\pi[\sigma]/\sigma_{\min}$, quantifying the correlation between the target distribution and the bias. This ratio reflects the intuition that the effects of π and σ on the sampling process are more difficult to disambiguate if π puts large weight on points with a large bias.

Proof of Theorem 6.2. For the settings of c_j, d_j, c', d' assumed by the theorem, we obtain by Hoeffding's inequality and the union bound that with probability at least $1-\delta$, $c_j \le \mathbf{E}_{\pi\sigma}[f_j/s] \le d_j$ for all j and $c' \le \mathbf{E}_{\pi\sigma}[1/\sigma] \le d'$ (note that c' is set to the minimum possible value of $1/\sigma(x)$ whenever the lower bound obtained by Hoeffding's inequality would be smaller). We further analyze the case when all of these inequalities hold. Then $\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f}]$ lies in B as defined in Eq. (6.1). Hence by Eq. (3.6)

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_B(\boldsymbol{\lambda}^\star) + \mathrm{I}^*_B(-\boldsymbol{\lambda}^\star) \ . \tag{6.5}$$

Using Eq. (6.2) we obtain

$$\begin{aligned}
\mathrm{I}^*_B(-\boldsymbol{\lambda}^\star) + \mathrm{I}^*_B(\boldsymbol{\lambda}^\star) &= \sum_{j\in\mathcal{J}} \left(\frac{d_j}{c'} - \frac{c_j}{d'}\right)|\lambda^\star_j| \\
&= \sum_{j\in\mathcal{J}} \left(\frac{(d_j - c_j)|\lambda^\star_j|}{c'} + \frac{d'-c'}{c'}\cdot\frac{c_j|\lambda^\star_j|}{d'}\right) \\
&\le \sum_{j\in\mathcal{J}} \frac{2\beta}{c'}\Big(|\lambda^\star_j| + \mathbf{E}_\pi[f_j]|\lambda^\star_j|\Big) \qquad (6.6) \\
&= \frac{\|\boldsymbol{\lambda}^\star\|_1 + \sum_{j\in\mathcal{J}} \mathbf{E}_\pi[f_j]|\lambda^\star_j|}{\sqrt{m}} \cdot \frac{\alpha}{\sigma_{\min} c'} \qquad (6.7)
\end{aligned}$$

Eq. (6.6) follows since $d_j - c_j = 2\beta$, $d' - c' \le 2\beta$, and $c_j/d' \le \mathbf{E}_\pi[f_j]$. The last inequality holds because $c_j \le \mathbf{E}_{\pi\sigma}[f_j/\sigma]$, $d' \ge \mathbf{E}_{\pi\sigma}[1/\sigma]$, and hence by Theorem 6.1

$$\frac{c_j}{d'} \le \frac{\mathbf{E}_{\pi\sigma}[f_j/\sigma]}{\mathbf{E}_{\pi\sigma}[1/\sigma]} = \mathbf{E}_\pi[f_j] \ .$$

Eq. (6.7) follows since $2\beta = \alpha/(\sigma_{\min}\sqrt{m})$. To obtain the bounds (6.3) and (6.4), it now suffices to bound $1/c'$ and combine Eq. (6.7) with Eq. (6.5). Using the bound $1/c' \le \sigma_{\max}$, we obtain Eq. (6.3). To prove Eq. (6.4), note that

$$\mathbf{E}_{\pi\sigma}[1/\sigma] \le d' \le c' + 2\beta \ .$$

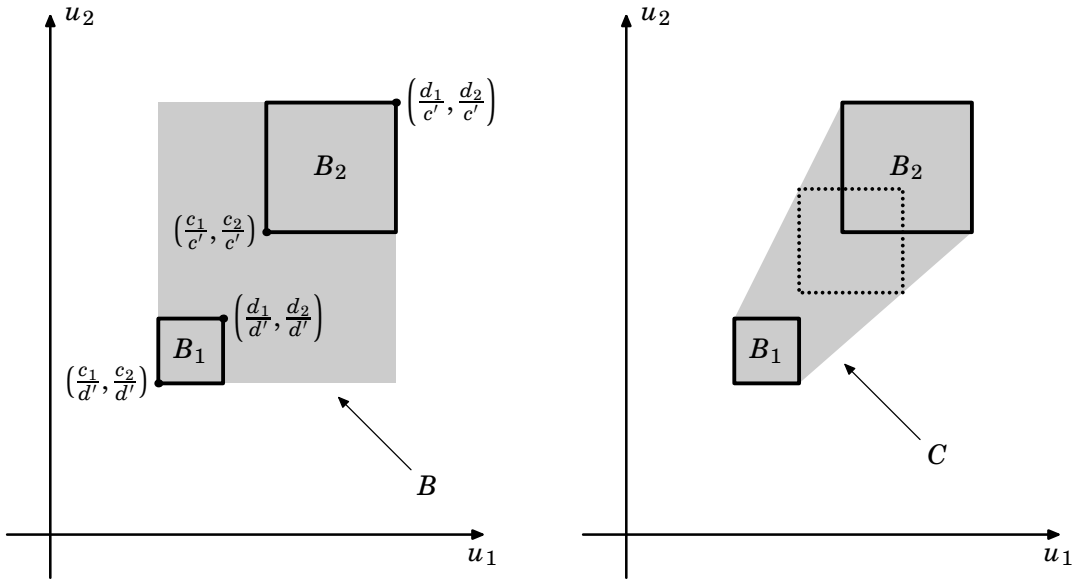


Figure 6.1. *Comparison of the box and polyhedral debiasing potentials.* Box B is the smallest box covering the extreme boxes B_1 and B_2 corresponding to the extreme values of the denominator in Theorem 6.1. Polytope C is the union of boxes across all values of the denominator (one of these boxes is shown dotted).

Hence, if $\mathbf{E}_{\pi\sigma}[1/\sigma] \geq 2\beta$ then

$$\begin{aligned}
\frac{1}{c'} &\leq \frac{1}{\mathbf{E}_{\pi\sigma}[1/\sigma] - 2\beta} \\
&= \left(\frac{1}{\mathbf{E}_{\pi}[\sigma]} - \frac{\alpha}{\sigma_{\min}\sqrt{m}}\right)^{-1} \qquad (6.8) \\
&= \mathbf{E}_{\pi}[\sigma]\left(1 - \frac{1}{\sqrt{m}}\frac{\alpha\mathbf{E}_{\pi}[\sigma]}{\sigma_{\min}}\right)^{-1} . \qquad (6.9)
\end{aligned}$$

Eq. (6.8) follows by Theorem 6.1 and the definitions of α and β. Note that the condition $\mathbf{E}_{\pi\sigma}[1/\sigma] \geq 2\beta$ is equivalent to the assumption $m \geq \left(\alpha\mathbf{E}_{\pi}[\sigma]/\sigma_{\min}\right)^2$. Combining Eqs. (6.9), (6.7), and (6.5), we obtain Eq. (6.4). ■

In the previous discussion, we have shown how Theorem 6.1 can be used to construct a box-shaped confidence region for $\mathbf{E}_{\pi}[\boldsymbol{f}]$ based on confidence intervals $[c_j, d_j]$ for $\mathbf{E}_{\pi\sigma}[f_j/\sigma]$ and $[c', d']$ for $\mathbf{E}_{\pi\sigma}[1/\sigma]$. However, Theorem 6.1 and the same confidence intervals can also be used to obtain a tighter (polyhedral) confidence region for $\mathbf{E}_{\pi}[\boldsymbol{f}]$. The tighter confidence region, denoted C, is derived from the observation that the value $\mathbf{E}_{\pi\sigma}[1/\sigma]$ required in the denominator of Theorem 6.1 is the same across all

features f_j:

$$C = \bigcup_{c' \le t \le d'} \left\{ \boldsymbol{u} \in \mathbb{R}^{\mathcal{J}} : \frac{c_j}{t} \le u_j \le \frac{d_j}{t} \text{ for all } j \right\} = \text{convex hull}(B_1 \cup B_2) \tag{6.10}$$

where B_1, B_2 are boxes corresponding to the extreme values of t

$$B_1 = \left\{ \boldsymbol{u} : \frac{c_j}{c'} \le u_j \le \frac{d_j}{c'} \text{ for all } j \right\} , \qquad B_2 = \left\{ \boldsymbol{u} : \frac{c_j}{d'} \le u_j \le \frac{d_j}{d'} \text{ for all } j \right\} .$$

In Fig. 6.1 we show the boxes B_1, B_2, and B (see Eq. 6.1) as well as the polytope C. Note that the polytope C is smaller than the box B, hence it should yield better performance guarantees. The potential and regularization corresponding to the polytope C are

$$\begin{aligned} \mathrm{U}(\boldsymbol{u}) &= \mathrm{I}_C(\boldsymbol{u}) \\ \mathrm{U}^*(-\boldsymbol{\lambda}) = \mathrm{I}_C^*(-\boldsymbol{\lambda}) &= \max_{t \in \{c', d'\}} \left[\sum_{j \in \mathcal{J}} \frac{-(c_j + d_j)\lambda_j + (d_j - c_j)|\lambda_j|}{2t} \right] , \end{aligned} \tag{6.11}$$

where the regularization is derived by Eq. (2.11).

Using the same settings for $[c_j, d_j]$ and $[c', d']$ as in Theorem 6.2, we obtain the following (tighter) guarantee.

Theorem 6.3. *Assume that features $f_j, j \in \mathcal{J}$ are bounded in $[0,1]$ and the bias σ is bounded in $[\sigma_{\min}, \sigma_{\max}], \sigma_{\min} > 0$. Let $\delta > 0$ and set $c_j, d_j, c', d', \beta, \alpha$ as in Theorem 6.2. Let $\hat{\boldsymbol{\lambda}}$ minimize $\ln Z_{\boldsymbol{\lambda}} + \mathrm{I}_C^*(-\boldsymbol{\lambda})$ where*

$$\mathrm{I}_C^*(-\boldsymbol{\lambda}) = \max_{t \in \{c', d'\}} \left[\frac{-\boldsymbol{\lambda} \cdot \mathbf{E}_{\widetilde{\pi\sigma}}[\boldsymbol{f}/\sigma] + \beta \|\boldsymbol{\lambda}\|_1}{t} \right] . \tag{6.12}$$

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

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\boldsymbol{\lambda}^\star\|_1 + |\boldsymbol{\lambda}^\star \cdot \mathbf{E}_\pi[\boldsymbol{f}]|}{\sqrt{m}} \cdot \frac{\alpha \sigma_{\max}}{\sigma_{\min}} . \tag{6.13}$$

Moreover, if $m \ge \left(\alpha \mathbf{E}_\pi[\sigma]/\sigma_{\min}\right)^2$ then with probability at least $1-\delta$, for every Gibbs distribution $q_{\boldsymbol{\lambda}^\star}$,

$$\mathrm{L}_\pi(\hat{\boldsymbol{\lambda}}) \le \mathrm{L}_\pi(\boldsymbol{\lambda}^\star) + \frac{\|\boldsymbol{\lambda}^\star\|_1 + |\boldsymbol{\lambda}^\star \cdot \mathbf{E}_\pi[\boldsymbol{f}]|}{\sqrt{m}} \cdot \frac{\alpha \mathbf{E}_\pi[\sigma]}{\sigma_{\min}} \left(1 - \frac{1}{\sqrt{m}} \frac{\alpha \mathbf{E}_\pi[\sigma]}{\sigma_{\min}}\right)^{-1} . \tag{6.14}$$

The bounds (6.13) and (6.14) improve over the bounds of Theorem 6.2 because they replace $\sum_{j \in \mathcal{J}} \mathbf{E}_\pi[f_j]|\lambda_j^\star|$ by the smaller term $|\boldsymbol{\lambda}^\star \cdot \mathbf{E}_\pi[\boldsymbol{f}]|$. Improvement in the

guarantee due to the polyhedral regularization will be most significant when $\boldsymbol{\lambda}^\star$ and $\mathbf{E}_\pi[\boldsymbol{f}]$ are close to orthogonal.

Proof of Theorem 6.3. Similarly to Theorem 6.2, we analyze the case when all the instances of Hoeffding's inequality used to set the confidence intervals $[c_j, d_j]$ and $[c', d']$ hold. Note that for our settings of c_j, d_j, c', d', the regularization $\mathrm{I}_C^*(-\boldsymbol{\lambda}^\star)$, defined in Eq. (6.11), indeed equals the expression given in Eq. (6.12). As in Theorem 6.2, it suffices to bound $\mathrm{I}_C^*(\boldsymbol{\lambda}^\star) + \mathrm{I}_C^*(-\boldsymbol{\lambda}^\star)$ and apply Eq. (3.6).

Let $\tilde{\boldsymbol{u}}$ denote $\mathbf{E}_{\widetilde{\pi\sigma}}[\boldsymbol{f}/\sigma]$. Rewrite Eq. (6.12) using the identity $\max\{a,b\} = (a+b)/2 + |a-b|/2$:

$$\mathrm{I}_C^*(-\boldsymbol{\lambda}) = \left(\frac{1}{c'} + \frac{1}{d'}\right)\frac{-\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}} + \beta\|\boldsymbol{\lambda}\|_1}{2} + \left(\frac{1}{c'} - \frac{1}{d'}\right)\frac{\left|-\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}} + \beta\|\boldsymbol{\lambda}\|_1\right|}{2} \quad .$$

Hence,

$$\begin{aligned}
\mathrm{I}_C^*(-\boldsymbol{\lambda}) + \mathrm{I}_C^*(\boldsymbol{\lambda}) &= \left(\frac{1}{c'} + \frac{1}{d'}\right)\beta\|\boldsymbol{\lambda}\|_1 + \left(\frac{1}{c'} - \frac{1}{d'}\right)\frac{\left|-\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}} + \beta\|\boldsymbol{\lambda}\|_1\right| + \left|\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}} + \beta\|\boldsymbol{\lambda}\|_1\right|}{2} \\
&= \left(\frac{1}{c'} + \frac{1}{d'}\right)\beta\|\boldsymbol{\lambda}\|_1 + \left(\frac{1}{c'} - \frac{1}{d'}\right)\max\{|\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}}|, \beta\|\boldsymbol{\lambda}\|_1\} \qquad (6.15)
\end{aligned}$$

where Eq. (6.15) follows because $\max\{|a|,|b|\} = |a+b|/2 + |a-b|/2$. Next, we bound $|\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}}|$ from above using our assumption that $|\mathbf{E}_{\widetilde{\pi\sigma}}[f_j/\sigma] - \mathbf{E}_{\pi\sigma}[f_j/\sigma]| \le \beta$ for all j:

$$|\boldsymbol{\lambda}\cdot\tilde{\boldsymbol{u}}| = |\boldsymbol{\lambda}\cdot\mathbf{E}_{\widetilde{\pi\sigma}}[\boldsymbol{f}/\sigma]| \le |\boldsymbol{\lambda}\cdot\mathbf{E}_{\pi\sigma}[\boldsymbol{f}/\sigma]| + \beta\|\boldsymbol{\lambda}\|_1 \quad . \qquad (6.16)$$

Furthermore, by Theorem 6.1 and the assumption $\mathbf{E}_{\pi\sigma}[1/\sigma] \le d'$, we obtain

$$|\boldsymbol{\lambda}\cdot\mathbf{E}_{\pi\sigma}[\boldsymbol{f}/\sigma]| = \mathbf{E}_{\pi\sigma}[1/\sigma]|\boldsymbol{\lambda}\cdot\mathbf{E}_\pi[\boldsymbol{f}]| \le d'|\boldsymbol{\lambda}\cdot\mathbf{E}_\pi[\boldsymbol{f}]| \quad . \qquad (6.17)$$

Combining Eqs. (6.15), (6.16), and (6.17) yields

$$\begin{aligned}
\mathrm{I}_C^*(-\boldsymbol{\lambda}) + \mathrm{I}_C^*(\boldsymbol{\lambda}) &\le \left(\frac{1}{c'} + \frac{1}{d'}\right)\beta\|\boldsymbol{\lambda}\|_1 + \left(\frac{1}{c'} - \frac{1}{d'}\right)\left(d'|\boldsymbol{\lambda}\cdot\mathbf{E}_\pi[\boldsymbol{f}]| + \beta\|\boldsymbol{\lambda}\|_1\right) \\
&= \frac{2\beta\|\boldsymbol{\lambda}\|_1}{c'} + \frac{d'-c'}{c'}|\boldsymbol{\lambda}\cdot\mathbf{E}_\pi[\boldsymbol{f}]| \\
&\le \frac{2\beta}{c'}(\|\boldsymbol{\lambda}\|_1 + |\boldsymbol{\lambda}\cdot\mathbf{E}_\pi[\boldsymbol{f}]|)
\end{aligned}$$

where the last inequality follows because $d' \le c' + 2\beta$. The bounds (6.13) and (6.14) can now be derived by expressing 2β as $\alpha/(\sigma_{\min}\sqrt{m})$ and bounding $1/c'$ as in Theorem 6.2. ■

In practice, confidence intervals $[c_j, d_j]$ and $[c', d']$ may be determined from sam-

Input: finite domain $\mathcal{X}$
default estimate q_0
strictly positive sample selection bias σ
features $f_j : \mathcal{X} \to [0,1]$
samples $x_1, \ldots, x_m \in \mathcal{X}$
regularization parameter β_0

Output: $q_{\hat{\boldsymbol{\lambda}}}$ approximating the target distribution

Let $\gamma' = \min_{x \in \mathcal{X}} 1/\sigma(x)$, $\delta' = \max_{x \in \mathcal{X}} 1/\sigma(x)$

$$\beta' = \frac{\beta_0}{\sqrt{m}} \cdot \min\left\{ \mathbf{V}'_{\widetilde{\pi\sigma}}[1/\sigma], \frac{\delta' - \gamma'}{2} \right\}$$

$$[c', d'] = \left[\mathbf{E}_{\widetilde{\pi\sigma}}[1/\sigma] - \beta', \mathbf{E}_{\widetilde{\pi\sigma}}[1/\sigma] + \beta' \right] \cap \left[\gamma', \delta' \right]$$

For $j \in \mathcal{J}$:

$$\gamma_j = \min_{x \in \mathcal{X}} f_j(x)/\sigma(x) \ , \quad \delta_j = \max_{x \in \mathcal{X}} f_j(x)/\sigma(x)$$

$$\beta_j = \frac{\beta_0}{\sqrt{m}} \cdot \min\left\{ \mathbf{V}'_{\widetilde{\pi\sigma}}[f_j/\sigma], \frac{\delta_j - \gamma_j}{2} \right\}$$

$$[c_j, d_j] = \left[\mathbf{E}_{\widetilde{\pi\sigma}}[f_j/\sigma] - \beta_j, \mathbf{E}_{\widetilde{\pi\sigma}}[f_j/\sigma] + \beta_j \right] \cap \left[\gamma_j, \delta_j \right]$$

Solve maxent with potential I_B (or I_C) where B (or C) is defined in Eq. (6.1) (or Eq. 6.10)

Figure 6.2. *Maxent with debiasing of averages.*

ple variances similar to confidence intervals used in unbiased ℓ_1-regularized maxent. In the experiments of Sections 6.4 and 6.5, we will use the approximations

$$|\mathbf{E}_{\pi\sigma}[1/\sigma] - \mathbf{E}_{\widetilde{\pi\sigma}}[1/\sigma]| \approx \beta_0 \sqrt{\mathbf{V}'_{\widetilde{\pi\sigma}}[1/\sigma]/m} \ , \quad |\mathbf{E}_{\pi\sigma}[f_j/\sigma] - \mathbf{E}_{\widetilde{\pi\sigma}}[f_j/\sigma]| \approx \beta_0 \sqrt{\mathbf{V}'_{\widetilde{\pi\sigma}}[f_j/\sigma]/m}$$

where β_0 is a single tuning constant. After restricting the confidence intervals in a natural fashion, we obtain the algorithm in Fig. 6.2. Alternatively, we could use bootstrap or other types of estimates for the confidence intervals.

6.2.1 Solving Maxent with the Polyhedral Potential I_C

We discuss two approaches to solving maxent with potential I_C. The first approach explicitly enumerates the bounding hyperplanes of C and then applies one of the algorithms of Chapter 4. However, as we will see below, the number of bounding hyperplanes of C is quadratic in the number of features $|\mathcal{J}|$, so the resulting algorithm will be too slow when the number of features is large. Therefore, for a large number of features we suggest an alternative approach. The alternative approach finds the appropriate t in Eq. (6.10) by a line search while calling maxent with box constraints in each iteration. The box constraints are over the original set of features of size $|\mathcal{J}|$, and thus the resulting algorithm may be significantly faster.

Explicit Optimization

Following the intuition of Fig. 6.1, it is possible to derive C explicitly as an intersection of a box and a cone

$$C = \left\{ \boldsymbol{u} : \frac{c_j}{d'} \le u_j \le \frac{d_j}{c'} \text{ for all } j \in \mathcal{J} \text{ and } \right. \tag{6.18}$$

$$\left. d_j u_k - u_j c_k \ge 0 \text{ for all } j,k \in \mathcal{J},\, j \ne k \right\} \ . \tag{6.19}$$

To see how the box and conic constraints of Eqs. (6.18) and (6.19) define C, first consider the prototype box

$$B_0 = \{\boldsymbol{u} : c_j \le u_j \le d_j\}$$

and note that

$$C = \bigcup_{1/d' \le t_0 \le 1/c'} t_0 B_0 \ , \tag{6.20}$$

where t_0 corresponds to $1/t$ of Eq. (6.10). Note that the constraints of Eq. (6.19) are satisfied if and only if $\boldsymbol{u} = \alpha \boldsymbol{u}_0$ for some $\alpha \ge 0$ and $\boldsymbol{u}_0 \in B_0$. The constraints of Eq. (6.18) guarantee that $\boldsymbol{u}$ lies inside the box B (enclosing C), which restricts the range of α to $[1/d', 1/c']$. Thus, altogether these constraints specify C. The total number of constraints grows quadratically with the size of the feature set $|\mathcal{J}|$.

Line-search Optimization

Instead of using the explicit form of C, we will show how the decomposition in Eq. (6.20) can be used to solve maxent more efficiently.

Specifically, consider a "slice" of the primal objective

$$P_0(p, t_0) = \mathrm{D}(p \parallel q_0) + \mathrm{I}_{t_0 B_0}(\mathbf{E}_p[\boldsymbol{f}]) \ .$$

The primal objective

$$P(p) = \mathrm{D}(p \parallel q_0) + \mathrm{I}_C(\mathbf{E}_p[\boldsymbol{f}])$$

can be expressed in terms of the sliced primal as

$$P(p) = \min_{1/d' \le t_0 \le 1/c'} P_0(p, t_0) \ .$$

Thus, the maxent primal can be re-formulated as

$$\min_{p \in \Delta} \min_{1/d' \le t_0 \le 1/c'} P_0(p, t_0) = \min_{1/d' \le t_0 \le 1/c'} \min_{p \in \Delta} P_0(p, t_0) \ .$$

We solve the outer minimization on the right-hand side by a line search, evaluating

the inner minimization in each iteration. Note that the inner minimization is a maxent problem with the box potential $\mathrm{I}_{t_0 B_0}$, which can be solved using ℓ_1-SUMMET or ℓ_1-PLUMMET. It turns out that the outer minimization is convex, so it can be solved efficiently using standard techniques. To see that the outer minimization is convex, consider the function

$$\hat{P}_0(t_0) = \min_{p \in \Delta} P_0(p, t_0) \ ,$$

which is being minimized over $t_0 \in [1/d', 1/c']$. Note that $\mathrm{I}_{t_0 B_0}(\boldsymbol{u})$ is jointly convex over $t_0 \geq 0$ and $\boldsymbol{u} \in \mathbb{R}^{\mathcal{J}}$. Thus, $P_0(p, t_0)$, defined above, is jointly convex in p and t_0, and hence $\hat{P}_0(t_0)$ is convex in t_0 (note that the epigraph of $\hat{P}_0$ can be viewed as a "shadow" of P_0's epigraph). Since the number of features of the inner optimization is only $|\mathcal{J}|$, as opposed to $O(|\mathcal{J}|^2)$ obtained by a reduction from the explicit representation of I_C, the algorithm based on the optimization of t_0 may significantly outperform the explicit optimization if the number of features is large.

6.3 Approach II: Factoring Bias Out

Our second approach does not approximate π directly, but first uses maxent to estimate the distribution $\pi\sigma$ and then converts this estimate into an approximation of π. If the default estimate of π is q_0 then the default estimate of $\pi\sigma$ should be $q_0\sigma$. In the first step of our approach, we apply unbiased maxent with the appropriate regularization to the empirical distribution $\widetilde{\pi\sigma}$ with the default estimate $q_0\sigma$. This yields a distribution

$$q_0\sigma(x) e^{\hat{\boldsymbol{\lambda}} \cdot \boldsymbol{f}(x)} / Z$$

approximating $\pi\sigma$. To obtain an estimate of π, we simply factor out σ, and obtain the distribution

$$q_0(x) e^{\hat{\boldsymbol{\lambda}} \cdot \boldsymbol{f}(x)} / Z' \ .$$

To distinguish among Gibbs distributions derived from various default distributions, we will prefix "Gibbs" by the corresponding default distribution. Thus the previous two examples are instances of $q_0\sigma$-Gibbs and q_0-Gibbs distributions respectively.

If we use ℓ_1-regularized maxent to estimate $\pi\sigma$ then the resulting approach corresponds to ℓ_1-regularized maximum likelihood estimation of π by q_0-Gibbs distributions. When π itself is q_0-Gibbs then the distribution $\pi\sigma$ is $q_0\sigma$-Gibbs. Performance guarantees for unbiased maxent imply that estimates of $\pi\sigma$ converge to $\pi\sigma$ as the number of samples increases. Since $\sigma_{\min} > 0$, the estimates of π obtained by factoring out σ converge to π as well.

Table 6.1. *Comparison of distributions $q^\star$ and $q^{\star\star}$ minimizing* $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}})$ *and* $\mathrm{D}(\pi\sigma \parallel q_{\boldsymbol{\lambda}}\sigma)$ *in Example 6.4.*

x	$\boldsymbol{f}(x)$	$\pi(x)$	$\sigma(x)$	$\pi\sigma(x)$	$q^\star(x)$	$q^{\star\star}\sigma(x)$	$q^{\star\star}(x)$
1	(0,0)	0.4	0.4	0.64	0.25	0.544	0.34
2	(0,1)	0.1	0.4	0.16	0.25	0.256	0.16
3	(1,0)	0.1	0.1	0.04	0.25	0.136	0.34
4	(1,1)	0.4	0.1	0.16	0.25	0.064	0.16

When π is not q_0-Gibbs then $\pi\sigma$ is not $q_0\sigma$-Gibbs either. The present approach approximates π by a q_0-Gibbs distribution $\hat{q} = q_{\hat{\boldsymbol{\lambda}}}$ which, with an increasing number of samples, minimizes $\mathrm{D}(\pi\sigma \parallel q_{\boldsymbol{\lambda}}\sigma)$ rather than $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}})$. Our next example shows that these two minimizers may be different. Thus, unlike the solution of maxent with a debiasing potential, the solution of maxent with factor-bias-out does not necessarily approach the best Gibbs distribution even as the number of samples grows to infinity.

Example 6.4. Consider the space $\mathcal{X} = \{1,2,3,4\}$ with two features f_1, f_2. Features f_1, f_2, target distribution π, sampling distribution σ and the biased distribution $\pi\sigma$ are given in Table 6.1. We use the uniform distribution as a default estimate. The minimizer of $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}})$ is the unique uniform-Gibbs distribution $q^\star$ such that $\mathbf{E}_{q^\star}[\boldsymbol{f}] = \mathbf{E}_\pi[\boldsymbol{f}]$. Similarly, the minimizer $q^{\star\star}\sigma$ of $\mathrm{D}(\pi\sigma \parallel q_{\boldsymbol{\lambda}}\sigma)$ is the unique σ-Gibbs distribution for which $\mathbf{E}_{q^{\star\star}\sigma}[\boldsymbol{f}] = \mathbf{E}_{\pi\sigma}[\boldsymbol{f}]$. Solving for these exactly, we find that $q^\star$ and $q^{\star\star}$ are as given in Table 6.1, and that these two distributions differ.

6.3.1 Using the Empirical Sampling Distribution

As mentioned in the introduction, knowing the sampling distribution σ exactly is unrealistic. However, we often have access to samples from σ. Within the factoring-bias-out framework, it is possible to use an empirical estimate $\tilde{\sigma}$ instead of σ.

Specifically, assume that σ is unknown but that, in addition to samples $x_1, \ldots, x_m$ from $\pi\sigma$, we are also given a separate set of samples $x_{(1)}, x_{(2)}, \ldots, x_{(M)}$ from σ. We use the factor-bias-out approach with the sampling distribution σ replaced by

$$\tilde{\sigma}(x) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(x_{(i)} = x) \ .$$

To simplify the algorithm, we note that instead of using $q_0\tilde{\sigma}$ as a default estimate for $\pi\sigma$, it suffices to replace the sample space $\mathcal{X}$ by $\mathcal{X}_{\tilde{\sigma}} = \{x_{(1)}, x_{(2)}, \ldots, x_{(M)}\}$ and use q_0 restricted to $\mathcal{X}_{\tilde{\sigma}}$ as a default. The last step of factoring out $\tilde{\sigma}$ is equivalent to using $\hat{\boldsymbol{\lambda}}$, returned for space $\mathcal{X}_{\tilde{\sigma}}$, on the entire sample space $\mathcal{X}$.

When the sampling distribution σ is correlated with feature values, $\mathcal{X}_{\tilde{\sigma}}$ might not cover all feature ranges. In that case, re-projecting on $\mathcal{X}$ may yield poor estimates outside of these ranges. We therefore do "clamping", i.e., we restrict values $f_j(x)$ to their ranges over $\mathcal{X}_{\tilde{\sigma}}$ and cap values of the exponent $\hat{\boldsymbol{\lambda}} \cdot \boldsymbol{f}(x)$ at its maximum over $\mathcal{X}_{\tilde{\sigma}}$.

6.4 Synthetic Experiments

Conducting real-data experiments to evaluate bias correction techniques is difficult, because bias is typically unknown and samples from unbiased distributions are not available. Therefore, synthetic experiments are often a necessity for precise evaluation. In this section we describe a set of synthetic experiments; in the next section we turn to experiments on real data.

Experimental Design

We generated three target uniform-Gibbs distributions π_1, π_2, π_3 over a domain $\mathcal{X}$ of size 10,000. These distributions were derived from linear, quadratic, and product features derived from 10 variables $v_1, \ldots, v_{10}$. The values $v_k(x)$ were chosen independently and uniformly in $[0, 1]$. Fixing the variables, we generated parameter vectors $\boldsymbol{\lambda}$ capturing a range of different behaviors.

Let U_S denote a random variable uniform over the set S. Each instance of U_S corresponds to a new independent variable. We set

$$\lambda_{v_k^2} = U_{\{-1,0,1\}} U_{[1,5]}$$

$$\lambda_{v_k} = \begin{cases} \lambda_{v_k^2} U_{[-3,1]} & \text{if } \lambda_{v_k^2} \neq 0, \\ U_{\{-1,1\}} U_{[2,10]} & \text{otherwise.} \end{cases}$$

Weights $\lambda_{v_k v_{k'}}$, $k < k'$ were chosen to create correlations between variables that would be observable, but not strong enough to dominate over the effects of linear and quadratic features. We set $\lambda_{v_k v_{k'}} = -0.5$ or 0 or 0.5 with respective probabilities 0.05, 0.9 and 0.05.

In maxent algorithms, we used linear and quadratic features derived from variables including $v_1, \ldots, v_6$ (relevant variables) as well as additional (irrelevant) variables $v_{11}, \ldots, v_{14}$, generated similarly to the previous variables. Once generated, we used the same set of variables in all experiments.

We generated a sampling distribution σ correlated with target distributions. More specifically, σ was a Gibbs distribution generated from linear and quadratic features derived from variables $v_5, \ldots, v_8$ and additional variables v_{15}, v_{16}. The parameters of

Table 6.2. *Summary of variables and feature types used in synthetic experiments.* Target distributions π_1, π_2, π_3, the sampling distribution σ, and the maxent distribution $\hat{q}$ are all uniform-Gibbs, derived from the shown features and variables.

	Feature types	Variables						
		v_1, v_2, v_3, v_4	v_5, v_6	v_7, v_8	v_9, v_{10}	$v_{11}, v_{12}, v_{13}, v_{14}$		v_{15}, v_{16}
target π_1, π_2, π_3	LQP	√	√	√	√			
maxent model $\hat{q}$	LQ	√	√			√		
bias σ	LQ		√	√				√

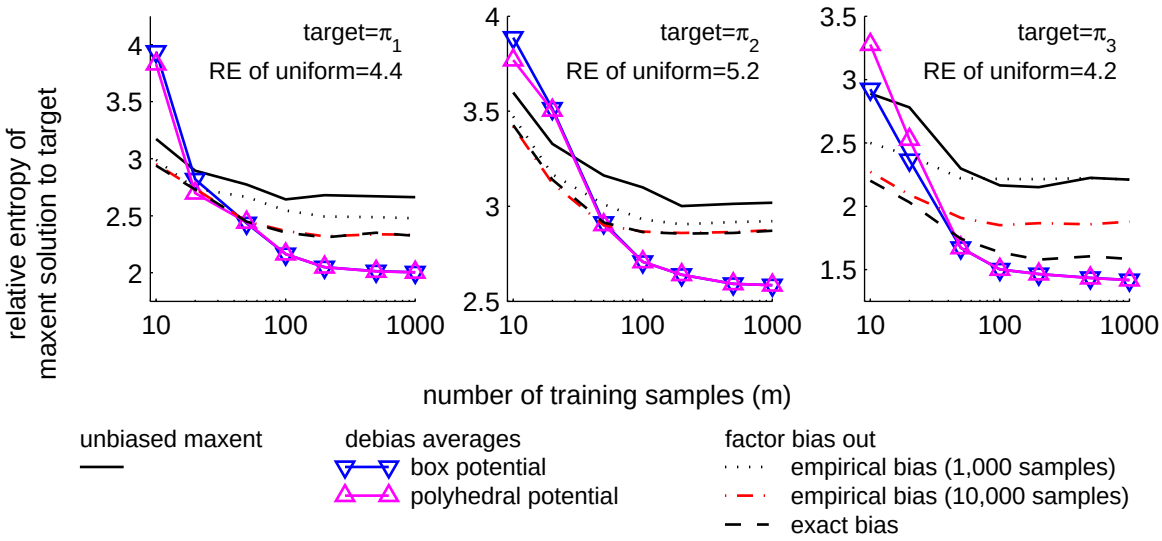


Figure 6.3. *Learning curves for synthetic experiments.* Performance is measured in terms of relative entropy to the target distribution as a function of an increasing number of training samples. The number of samples is plotted on a log scale.

the sampling distribution were set to $\lambda_{v_k} = 0$ and $\lambda_{v_k^2} = -1$. The choice of variables and feature types is summarized in Table 6.2.

For every target distribution, we evaluated the performance of unbiased maxent, maxent with debiasing potential I_B and I_C, as well as the factor-bias-out approach using exact knowledge of σ and an empirical distribution $\tilde{\sigma}$ consisting of 1,000 or 10,000 samples. Performance was evaluated in terms of relative entropy to the target distribution. We used training sets of sizes 10 to 1,000. We considered three randomly generated training sets and took the average performance for settings of β_0 from the range $[0.01, 100.00]$. We report results for the best β_0, chosen separately for each average. The rationale behind this approach is that we want to explore the potential performance of each method.

Table 6.3. *Results of real-data experiments.* Average performance of unbiased maxent and bias correction approaches over all species in six regions. Results of bias correction approaches are set in boldface if they are significantly better than those of unbiased maxent according to a paired t-test at the 5% significance level.

	Log loss				AUC			
		difference from unbiased				difference from unbiased		
	unbiased maxent	debias avgs (box)	factor bias out	factor bias out (empir.)	unbiased maxent	debias avgs (box)	factor bias out	factor bias out (empir.)
AWT	-0.44	0.14	0.12	0.11	0.69	-0.02	0.02	**0.03**
CAN	-1.33	0.21	0.24	0.50	0.58	0.06	**0.11**	**0.15**
NSW	-0.82	0.47	0.65	0.78	0.71	-0.06	0.00	0.01
NZ	-0.47	0.55	0.44	0.31	0.72	-0.05	0.00	0.01
SA	-1.09	0.96	0.53	0.48	0.78	-0.10	-0.00	-0.00
SWI	-1.41	0.78	0.65	0.60	0.81	-0.02	**0.03**	**0.03**

Results

Figure 6.3 shows the results at the optimal β as a function of an increasing number of samples. The factor-bias-out approach with the exact bias or 10,000 samples is always better than unbiased maxent. Maxent with debiasing potentials is worse than the unbiased maxent for small sample sizes, but as the number of training samples increases, it soon outperforms unbiased maxent and eventually also outperforms factor-bias-out. The two debiasing potentials lead to different performance only for small training set sizes. Factor-bias-out with empirical bias improves as the number of samples increases, approaching the performance of factor-bias-out with exact knowledge of the bias.

6.5 Real-data Experiments

We compared the performance of our bias correction approaches with that of unbiased maxent as well as other species distribution modeling techniques using the NCEAS dataset (see Section 5.4). We used the training presence-only portion of the data to construct models, assuming that the occurrence records are biased. The evaluation presence-absence portion of the data was viewed as the unbiased test set.

6.5.1 Evaluation of the Bias Removal Approaches

Experimental Design

We evaluated the performance of unbiased maxent, maxent with debiasing box potential, factor-bias-out, and empirical factor-bias-out. In all cases, we used thresh-

old features and categorical indicators, regularized according to the default settings given in Table 5.3.

We treated training occurrence locations for all species in each region as the samples from the sampling distribution. These were used directly in empirical factor-bias-out. To apply the other debiasing approaches, we estimated the sampling distribution using unbiased maxent with threshold features and categorical indicators with the default regularization. The resulting debiased distributions were evaluated on test presences according to log loss and on test presences and absences according to AUC (see Section 5.2). Sampling distribution estimation is also the first step in the work of Zadrozny (2004). In contrast with that work, however, our evaluation measures do not use the sampling-distribution estimate and hence do not depend on its quality.

Results

In Table 6.3 we show performance of the three bias correction approaches compared with unbiased maxent. All three algorithms yield on average a worse log loss than unbiased maxent. In contrast, when the performance is measured in terms of AUC, factor-bias-out (both with the estimated bias and the empirical bias) yields on average the same or better AUC as unbiased maxent in five out of six regions. Improvements in regions *AWT*, *CAN*, and *SWI* are dramatic enough that both of these methods perform better (according to AUC) than any method evaluated in the original NCEAS comparison (see Section 5.6).

6.5.2 The NCEAS Comparison Incorporating the Bias Removal

According to the previous experiments the best performance on the NCEAS data (measured by AUC) is obtained by empirical factor-bias-out. This approach replaces the sample set $\mathcal{X}$ by a set of occurrences of all of the species in the same region, viewed as samples from the biased sampling distribution. We expect that a better empirical estimate of the sampling distribution is obtained if we restrict the set of species to those that are likely to be collected in a similar manner, species within the same *target group*, since these are likely to share similar sample selection bias. We refer to this particular choice of the empirical bias distribution as *target-group background*. In this and the next section, we will explore the use of maxent with target-group background.

The use of target-group background is not restricted to maxent. In particular, target-group background can be readily incorporated in regression based techniques, which require surrogates for the absence of the species. Instead of choosing these sur-

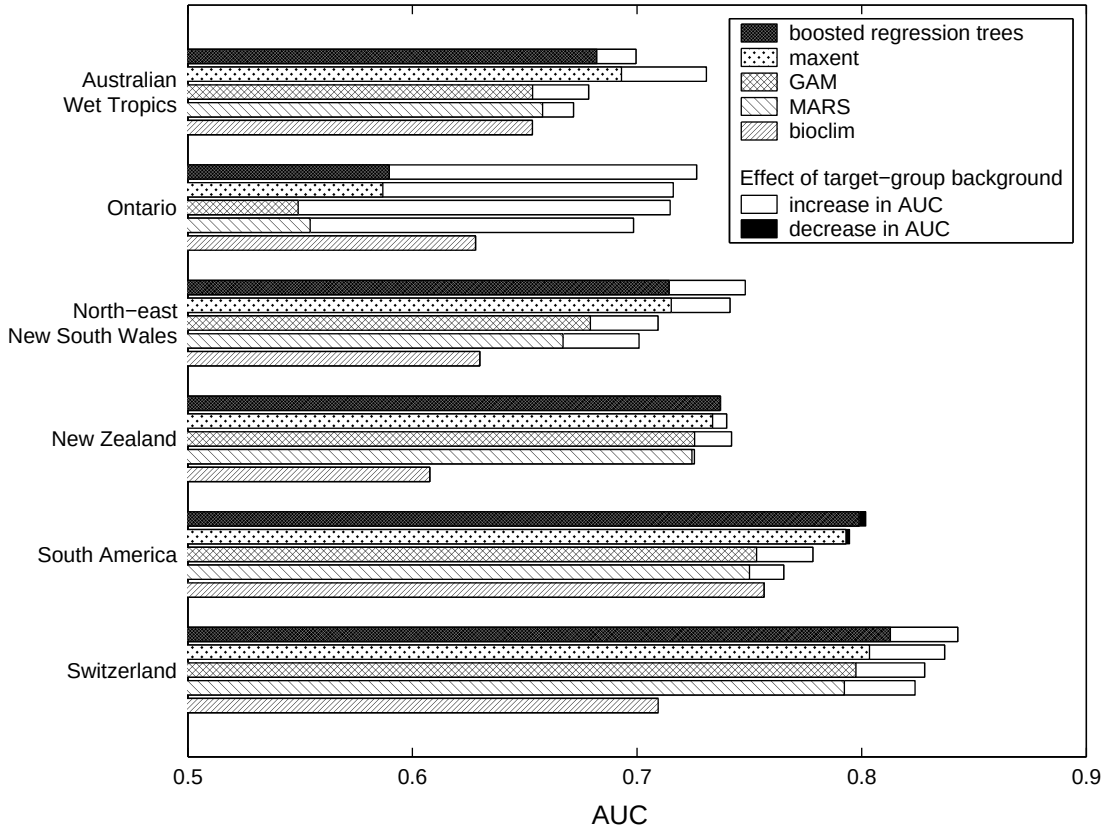


Figure 6.4. *Comparison of* Maxent *and other species distribution modeling techniques using target-group background to remove the bias.* The results are taken from the work of Phillips, Dudík et al. (2007).

rogate absences from the entire region at random (similar to using the entire region or its subsample as the sample space in maxent), we can use the target-group background. The approach was suggested earlier by Ferrier et al. (2002) and Zaniewski et al. (2002) and comprehensively investigated by Phillips, Dudík et al. (2007). In Fig. 6.4, we summarize some of the findings of Phillips, Dudík et al. (2007) comparing the performance of maxent with that of the other species distribution modeling methods.

We consider the same set of techniques as in Section 5.6: BRT, maxent, MARS, GAM, and BIOCLIM. All of the methods, except for BIOCLIM, allow incorporating target-group background. There are two target groups in *AWT* (birds and plants), four target groups in *NSW* (birds, plants, mammals, and reptiles), and only one target group in each of the remaining regions. To obtain maxent models, we used the *Maxent* software with the default settings determined in Section 5.5 including hinge features.

Fig. 6.4 shows that the performance of all four methods that allow incorporating

Table 6.4. *Maxent performance in terms of AUC for globally and regionally optimized parameters, using either random or target-group background.*

	Random background			Target-group background		
		improvement from default			improvement from default	
	default settings	globally optimized settings	regionally optimized settings	default settings	globally optimized settings	regionally optimized settings
AWT	0.693	0.004	0.015	0.729	0.000	0.009
CAN	0.594	0.008	0.023	0.719	0.011	0.020
NSW	0.711	0.005	0.022	0.742	0.009	0.020
NZ	0.733	0.008	0.009	0.741	0.009	0.011
SA	0.796	0.007	0.014	0.793	0.003	0.005
SWI	0.803	0.003	0.001	0.837	0.006	0.006
all species	0.726	0.006	0.014	0.757	0.006	0.012

the target-group background significantly improves. Note that the Ontario anomaly, on which we commented in Section 5.6, disappears after the bias removal. Maxent remains among the top performing approaches alongside BRT.

6.5.3 Evaluating the *Maxent* Tuning

In the previous section we have seen that the default settings of *Maxent* lead to good performance using the target-group background. However, the tuning of Section 5.5 ignored sample-selection bias, so it is conceivable that a different set of regularization parameters will yield better performance. In this section, we evaluate the extent of such improvement, using the same experimental design as in Section 5.7.

Specifically, we perform a local search in the space of possible regularization parameters to optimize the performance of maxent models. Maxent models are constructed from the training portion of the NCEAS data and evaluated on the evaluation portion of the NCEAS data. The resulting parameters are referred to as pa-tuned.

In Table 6.4, we compare the performance of the default settings with the performance of pa-tuned settings. For comparison, we also list the results for unbiased maxent. Both global and regional pa-tuning result in a similar improvement of AUC as for the unbiased maxent. Note that the improvement due to the use of target-group background is significantly larger than the improvement due to the global or regional pa-tuning.

In Table 6.5, we report parameters obtained by global pa-tuning and medians of parameters obtained by regional pa-tuning. Again, for comparison we include the

Table 6.5. *Overview of pa-tuned parameters: globally optimized parameters and medians of 10 regionally optimized parameters.* The global settings optimize the average performance across all species. The regional settings are optimized separately for each of 10 taxonomic groups in the 6 regions.

		Random background				Target-group background			
		number of occurrences				number of occurrences			
		2–9	10–14	15–79	≥ 80	2–9	10–14	15–79	≥ 80
β_L:	global optimum	1.00	1.41			2.00	1.41		
	regional median	1.00	1.00			1.00	1.00		
	default*	1.00	.71			1.00	.71		
β_Q:	global optimum		1.41	.50	.35†		2.00	1.00	.35†
	regional median		1.00	.85	.05		1.00	.71	.07
	default*		.71	.23	.05		.71	.23	.05
β_P:	global optimum				.35†				.35†
	regional median				.04				.11
	default*				.05				.05
β_T:	global optimum				2.00				8.00†
	regional median				1.21				1.71
	default*				1.00				1.00
β_H:	global optimum			.35	.50			.71	1.41
	regional median			.85	.50			1.21	.50
	default*			.50	.50			.50	.50
β_C:	global optimum	1.41	.50	.03**	.03	.50	1.00	.13	.25
	regional median	.71	.50	.18	.04	.71	.50	1.00	.04
	default*	.53	.39	.14	.05	.53	.39	.14	.05

* po-tuned values for the median training-set size in each range: 6, 12, 36, and 221
† the largest possible value in local search
** the smallest possible value in local search

results for unbiased maxent. Similar to the unbiased case, the pa-tuned values are larger than the default values, suggesting additional uncertainty in feature-mean estimates, possibly due to the imperfect knowledge of the sampling distribution. Since the improvement due to pa-tuning is only marginal, we suggest using the default settings to prevent possible overfitting of our evaluation data.

6.6 Discussion

We have proposed two approaches that incorporate information about sample selection bias in maxent and demonstrated their utility in synthetic and real-data experi-

ments. Experiments also raise several questions that merit further research.

Maxent with debiasing potentials has the strongest performance guarantees, but it performs the worst in real-data experiments and catches up with other methods only for moderate sample sizes in synthetic experiments. This may be due to poor estimates of unbiased confidence intervals and could be possibly improved using a different estimation method.

Maxent with factor-bias-out improves over unbiased maxent in terms of AUC over real data, but is worse in terms of log loss. This disagreement suggests that methods which aim to optimize AUC directly could be more successful in species modeling, possibly implementing a version of the factor-bias-out approach.

Maxent with empirical factor-bias-out performs the best on real world data, possibly due to the direct use of samples from the sampling distribution rather than a sampling distribution estimate. However, this method comes without performance guarantees and does not exploit the knowledge of the full sample space.

Some of the paradoxes mentioned above could be explained by presence of identical or similar sample-selection bias in both the training and the evaluation data. Let σ be the sampling distribution used to collect the training data. Recall that unbiased maxent returns the distribution $\hat{q}$ which, with an increasing number of *biased* samples, optimizes $\mathrm{D}(\pi\sigma \parallel q_{\boldsymbol{\lambda}})$; maxent with a debiasing potential optimizes $\mathrm{D}(\pi \parallel q_{\boldsymbol{\lambda}})$; maxent with factor-bias-out optimizes $\mathrm{D}(\pi\sigma \parallel q_{\boldsymbol{\lambda}}\sigma)$. If another sampling distribution σ' is used to collect the evaluation data then the test AUC will be a proxy for $\mathrm{D}(\pi\sigma' \parallel \hat{q}\sigma')$ (since absences share the sample-selection bias with presences, so $\hat{q}$ is effectively "normalized" only against points under the same bias), whereas the test log loss corresponds to $\mathrm{D}(\pi\sigma' \parallel \hat{q})$ (since $\hat{q}$ is normalized uniformly over the entire region). If σ and σ' are similar, then, indeed, unbiased maxent will optimize the test log loss and maxent with factor-bias-out will optimize the test AUC, as we have observed. The possibility of shared sample-selection bias in the NCEAS data is partly explored by Phillips, Dudík et al. (2007) with a somewhat indefinite conclusion: a shared sampling distribution seems to play a role, but its effects are difficult to distinguish from the effects of factoring the bias out. Better understanding of this interplay may lead to further improvements of bias correction techniques.

Chapter 7

Multiple-density Estimation

Many real-world applications, including species-distribution modeling, require solving multiple related learning problems. In this chapter, we use the insights of the generalization analysis of Chapter 3 to develop a maximum-entropy approach to multiple-density estimation.

Specifically, we study the problem of simultaneously estimating several probability distributions on the same space, where the datasets for each are organized into overlapping groups such as a hierarchy. In problems of multiple estimation, we can typically either pool our data or treat each estimation problem individually. In pooling data, we obtain a confident estimate from a large sample but ignore the important differences between datasets. On the other hand, individual estimates address the separate nature of each dataset but may lead to poor generalization because of small sample sizes.

In maxent, pooling of the data corresponds to choosing constraints based on the feature averages across samples in all the datasets, whereas individual estimation is based on averages within each dataset separately. Here, we develop *hierarchical maximum entropy density estimation* (HME), a procedure that lies in the powerful middle-ground between these choices. The datasets are grouped, and the constraints are imposed simultaneously using averages in each dataset as well as averages within groups of datasets. Using the general theory developed in Chapter 3, we show that for an appropriate choice of regularization parameters it is possible to share information within groups and also account for differences between the datasets. The density estimates from small sample sizes are influenced by the estimates for which we have more confidence; estimates from large sample sizes are less influenced by others. In statistics, this is known as *hierarchical/multi-level modeling* (Gelman and Hill, 2007) or *shrinkage* (originally introduced by Stein, 1956; and James and Stein, 1961). In machine learning, hierarchical models have been used, for example, by McCallum et al. (1998) and Teh et al. (2005). These methods are also related to

multitask or *transfer learning* (Caruana, 1993; Baxter, 2000; Raina et al., 2006).

In the first part of this chapter, we develop the theory of HME including duality results and generalization guarantees similar to maxent duality (Theorem 2.4) and the Generalization Lemma (Lemma 3.1). As a specific instance of the developed theory, we study a hierarchical version of ℓ_1-regularized maxent, ℓ_1-regularized HME. We show that ℓ_1-regularized HME is closely related to maximum *a posteriori* estimation with a hierarchical prior, or maximum likelihood estimation with hierarchical regularization (shrinkage). We prove strong generalization guarantees. The guarantees depend favorably both on the number of features and the number of groups. They provide guidance to choosing hyperparameters.

In the second part of the chapter, we explore the utility of ℓ_1-regularized HME on synthetic data and two large-scale real-world datasets from species distribution modeling. Specifically, we evaluate HME on the data from regions *AWT* and *NSW* of the NCEAS dataset (see Section 5.4), where the taxonomy of species provides a natural hierarchy. For example, *AWT* contains bird species such as the golden bowerbird or the tooth-billed catbird and plant species such as the black treefern or the black tulip oak. In recent solutions to species distribution modeling, including all methods of the NCEAS comparison (see Section 5.6), each species distribution is modeled separately, even though some methods use combined data in the preprocessing stages to transform the environmental space (Ferrier et al., 2002) or to construct a set of possibly relevant features (Leathwick et al., 2005). When modeling distributions of rare or endangered species, the number of occurrence records of a species is typically fewer than ten, and, as expected, the resulting estimates of its distribution are poor. With our approach, the information from several species is combined to produce better estimates for each individual species. A crucial insight is that a bird's distribution is likely to be more similar to other bird distributions than it is to plant distributions. Our results in Section 7.7 show significant improvements in predictive performance in both *AWT* and *NSW*.

7.1 Hierarchical Maximum Entropy Setup

Our goal is to model multiple densities over the same sample space.[1] Density estimation problems are referred to as *classes*, which are organized into *groups*; note that we are not performing classification. The set of classes will be denoted $\mathcal{Y}$, the set of groups will be denoted $\mathcal{G}$. Individual groups $g \in \mathcal{G}$ specify subsets of $\mathcal{Y}$, thus $\mathcal{G}$ is a subset of the power set $2^{\mathcal{Y}}$.

[1] The restriction that the densities are over the same sample space simplifies the exposition, but it could be omitted.

In *AWT*, the set $\mathcal{Y}$ contains 10 plant and 10 bird species. The set $\mathcal{G}$ contains three groups: *plants* with 10 elements, *birds* with 10 elements, and *all species* with 20 elements (see Fig. 7.1). Note that we make no requirements on the composition of groups. In particular, groups can arbitrarily overlap. For example, in *NSW* (Fig. 7.2), we consider the groups *trees* and *rainforest plants*, which intersect in the set of rainforest trees.

The sample space shared by the estimation problems is denoted $\mathcal{X}$. As in single-density estimation, the space $\mathcal{X}$ is described by features f_j, $j \in \mathcal{J}$, which may also depend on the class y, i.e., $f_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Input consists of pairs (x_1, y_1), (x_2, y_2), ..., $(x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$, representing a pooled sample across all classes. In *AWT*, y_1 may be the *golden bowerbird*, and x_1 geographic coordinates where it was observed. We assume that samples (x_i, y_i) come from an unknown joint distribution π and use the maximum entropy principle to approximate π. Our interest lies in approximating conditional distributions of location given species, $\pi_y(x) = \pi(x \mid y)$. This is in contrast to logistic regression, where the goal is to approximate $\pi(y \mid x)$ for classification.

Using maxent separately for each class would mean imposing constraints on the class distributions based on the feature averages within each class, such as requiring the model of the golden bowerbird to match the average altitude and the average squared altitude in which the golden bowerbird was observed. In HME, we use the group information to leverage information across species. In addition to requiring that feature expectations of each individual class be close to their empirical averages, we also require that feature expectations for each group be close to the group empirical averages. Thus, in *AWT*, we require that expectation of altitude across all birds is not too far from the average altitude across all samples from the group *birds*. Since the total number of samples in the group *birds* is larger than, for example, the number of samples of the class *golden bowerbird*, we can be more confident about our estimates of the means. This amounts to sharing information across all bird species.

We express both class and group constraints in terms of conditional expectations on the joint distribution. Before presenting a general form of HME in Section 7.2, we focus on a specific example based on ℓ_1-regularized maxent:

$$\begin{aligned} &\min_{p \in \Delta} \mathrm{D}(p \parallel q_0) \\ &\text{s.t. } p(y) = \tilde{\pi}(y) \text{ for all } y \in \mathcal{Y} \\ &\quad \left|\mathbf{E}_{\tilde{\pi}}[f_j \mid y] - \mathbf{E}_p[f_j \mid y]\right| \le \beta_{y,j} \text{ for all } y \in \mathcal{Y},\, j \in \mathcal{J} \\ &\quad \left|\mathbf{E}_{\tilde{\pi}}[f_j \mid y \in g] - \mathbf{E}_p[f_j \mid y \in g]\right| \le \beta_{g,j} \text{ for all } g \in \mathcal{G},\, j \in \mathcal{J}. \end{aligned} \tag{7.1}$$

Here, Δ is the simplex of probability distributions over $\mathcal{X} \times \mathcal{Y}$, q_0 is a default estimate, $p(y)$ is the marginal probability of class y, and $\beta_{y,f}$, $\beta_{g,f}$ are widths of box constraints.

Note that if $\mathcal{G} = \emptyset$, ℓ_1-regularized HME reduces to a series of ℓ_1-regularized maxent problems for each class: given fixed class probabilities, the joint relative entropy is minimized when class relative entropies are minimized. When $\mathcal{G}$ is non-empty, the set of constraints in ℓ_1-regularized HME is more restrictive than a series of single-class maxent problems, so the resulting solutions differ.

Similar to single-class maxent, we will see that HME is equivalent to a regularized maximum likelihood problem. When class probabilities are fixed as above, the relative entropy is minimized by a distribution which takes the form $p(x,y) = \tilde{\pi}(y) q_{\boldsymbol{\lambda}_y;y}(x)$, where $q_{\boldsymbol{\lambda}_y;y}$ is a Gibbs distribution on $\mathcal{X}$, parameterized by $\boldsymbol{\lambda}_y \in \mathbb{R}^{\mathcal{J}}$ given the default estimate $q_{0;y} := q_0(\cdot \mid y)$ and features $f_j(\cdot, y)$. Eq. (7.1) is equivalent to the following regularized maximum likelihood problem:

$$\max_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}} \\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}} \left\{ \frac{1}{m}\sum_{i=1}^{m}\left(\ln q_{\boldsymbol{\lambda}_{y_i};y_i}(x_i)\right) - \sum_{y\in\mathcal{Y},j\in\mathcal{J}}\left(\tilde{\pi}(y)\beta_{y,j}\left|\lambda_{y,j} - \sum_{g:y\in g}\eta_{g,j}\right|\right) - \sum_{g\in\mathcal{G},j\in\mathcal{J}}\left(\tilde{\pi}(g)\beta_{g,j}|\eta_{g,j}|\right)\right\} . \tag{7.2}$$

Here, we used $\tilde{\pi}(g)$ for the probability that $y \in g$ under the distribution $\tilde{\pi}$. The objective of Eq. (7.2) is a function of vectors $\boldsymbol{\lambda}_y$, which describe class distributions $q_{\boldsymbol{\lambda}_y}$, and vectors $\boldsymbol{\eta}_g$, which account for effects of membership in different groups. The goal is to optimize log likelihood of the data (the first term) under an ℓ_1-style penalty for deviating from group effects (the second term), which are themselves regularized by an ℓ_1-style penalty (the third term).

In the following sections, we prove a general duality result and a generalization lemma for HME. The duality for ℓ_1-regularized HME and the performance guarantees will follow as special cases.

7.2 HME with General Regularization

Similar to generalized single-class maxent, constraints in HME are represented by arbitrary convex functions. We introduce two types of constraints: (i) constraints on class and group probabilities, i.e., on $p(y)$ and $p(g)$, and (ii) constraints on conditional feature expectations, i.e., on $\mathbf{E}_p[\boldsymbol{f} \mid y]$ and $\mathbf{E}_p[\boldsymbol{f} \mid g]$, where $\mathbf{E}_p[\boldsymbol{f} \mid g]$ is a shorthand for $\mathbf{E}_p[\boldsymbol{f} \mid y \in g]$. We consider the following general formulation of HME:

$$\min_{p\in\Delta}\left[\mathrm{D}(p \parallel q_0) + \mathrm{V}(p_{\mathcal{Y}}, p_{\mathcal{G}}) + \sum_{y\in\mathcal{Y}} p(y)\mathrm{U}_y(\mathbf{E}_p[\boldsymbol{f} \mid y]) + \sum_{g\in\mathcal{G}} p(g)\mathrm{U}_g(\mathbf{E}_p[\boldsymbol{f} \mid g])\right] \tag{7.3}$$

where $p_{\mathcal{Y}}$ is the marginal distribution over classes, $p_{\mathcal{G}}$ is the vector of group probabilities $\langle p(g)\rangle_{g\in\mathcal{G}}$, and $\mathrm{V} : \mathbb{R}^{\mathcal{Y}} \times \mathbb{R}^{\mathcal{G}} \to (-\infty, \infty]$ as well as $\mathrm{U}_y : \mathbb{R}^{\mathcal{J}} \to (-\infty, \infty]$ and

$\mathrm{U}_g : \mathbb{R}^{\mathcal{J}} \to (-\infty, \infty]$ are closed proper convex functions. Functions V, U_y, and U_g specify a potential for the joint estimation problem. We assume that $\operatorname{dom} \mathrm{V}$ is a subset of $(0,\infty)^{\mathcal{G}}$, and Eq. (7.3) is feasible.

For example, the potential of ℓ_1-regularized HME, Eq. (7.1), is defined by

$$\begin{aligned}
\mathrm{V}(\boldsymbol{t}, \boldsymbol{t}^{(\text{group})}) &= \mathrm{I}(\boldsymbol{t} = \tilde{\pi}_{\mathcal{Y}}) \\
\mathrm{U}_y(\boldsymbol{u}) &= \mathrm{I}\Big(\big|\mathbf{E}_{\tilde{\pi}}[f_j \mid y] - u_j\big| \le \beta_{y,j} \text{ for all } j \in \mathcal{J}\Big) \\
\mathrm{U}_g(\boldsymbol{u}) &= \mathrm{I}\Big(\big|\mathbf{E}_{\tilde{\pi}}[f_j \mid g] - u_j\big| \le \beta_{g,j} \text{ for all } j \in \mathcal{J}\Big) \ .
\end{aligned}$$

To prove HME versions of maxent duality and the Generalization Lemma, we will reduce the HME primal, Eq. (7.3), to the generalized maxent primal, Eq. (2.17), and use the theory of Chapters 2 and 3.

7.3 Reduction to Generalized Maxent

In generalized maxent, constraints are expressed as a potential on unconditional expectations, but HME defines a potential on marginal probabilities and conditional feature expectations. We begin the reduction by introducing a new set of features derived from f_j, and then show that HME potential is a closed proper convex function of unconditional expectations of the new features.

Specifically, we introduce new features indexed by $y \in \mathcal{Y}$ and $(y, j) \in \mathcal{Y} \times \mathcal{J}$:

$$\begin{aligned}
\mathbb{1}_y(x', y') &= \mathbb{1}(y' = y) \\
h_{y,j}(x', y') &= \mathbb{1}(y' = y) f_j(x', y') \ .
\end{aligned}$$

Features $\mathbb{1}_y$ are class indicators, features $h_{y,j}$ are the original features f_j restricted to a single class y. The class probabilities, group probabilities, and the corresponding conditional expectations can be expressed using $\mathbb{1}_y$ and $h_{y,j}$ as

$$p(y) = \textstyle\sum_{x' \in \mathcal{X}, y' \in \mathcal{Y}} p(x', y') \mathbb{1}(y' = y) = \mathbf{E}_p[\mathbb{1}_y] \tag{7.4}$$

$$p(g) = \textstyle\sum_{y \in g} p(y) = \sum_{y \in g} \mathbf{E}_p[\mathbb{1}_y] \tag{7.5}$$

$$\mathbf{E}_p[\boldsymbol{f} \mid y] = \frac{\sum_{x' \in \mathcal{X}, y' \in \mathcal{Y}} \boldsymbol{f}(x', y') p(x', y') \mathbb{1}(y' = y)}{p(y)} = \frac{\mathbf{E}_p[\boldsymbol{h}_y]}{\mathbf{E}_p[\mathbb{1}_y]} \tag{7.6}$$

$$\mathbf{E}_p[\boldsymbol{f} \mid g] = \frac{\sum_{y \in g} p(y) \mathbf{E}_p[\boldsymbol{f} \mid y]}{p(g)} = \frac{\sum_{y \in g} \mathbf{E}_p[\boldsymbol{h}_y]}{\sum_{y \in g} \mathbf{E}_p[\mathbb{1}_y]} \tag{7.7}$$

where $\boldsymbol{h}_y$ denotes the vector of features $\langle h_{y,j} \rangle_{j \in \mathcal{J}}$.

Next we will rewrite Eqs. (7.4)–(7.7) and the HME potential into a more compact

form, so that we can argue that the HME potential is a closed proper convex function of the feature expectations, and, in the next section, derive an explicit conjugate.

We begin by rewriting expressions for $p(g)$ and $\mathbf{E}_p[\boldsymbol{f} \mid g]$, using the group-membership matrix $\mathbf{M} \in \{0,1\}^{\mathcal{G}\times\mathcal{Y}}$ with entries $M_{gy} = \mathbb{1}(y \in g)$,

$$p(g) = \mathbf{M}_g \mathbf{E}_p[\mathbb{1}_\mathcal{Y}] \tag{7.8}$$

$$\mathbf{E}_p[\boldsymbol{f} \mid g] = \frac{(\mathbf{M}_g \otimes \mathbf{I}_\mathcal{J})\mathbf{E}_p[\boldsymbol{h}]}{\mathbf{M}_g \mathbf{E}_p[\mathbb{1}_\mathcal{Y}]} \tag{7.9}$$

where $\mathbb{1}_\mathcal{Y}$ denotes the vector of features $\mathbb{1}_y$, $\boldsymbol{h}$ is the concatenation of $\boldsymbol{h}_y$'s, and $\mathbf{M}_g$ is the row of $\mathbf{M}$ indexed by g. The notation $\otimes$ is used for the *tensor* (or *Kronecker*) *product* and $\mathbf{I}_\mathcal{J}$ for an identity matrix of size $|\mathcal{J}|$. For a pair of matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times \ell}$, their tensor product $\mathbf{A}\otimes\mathbf{B} \in \mathbb{R}^{mk\times n\ell}$ is defined as

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} A_{11}\mathbf{B} & A_{12}\mathbf{B} & \dots & A_{1n}\mathbf{B} \\ A_{21}\mathbf{B} & A_{22}\mathbf{B} & \dots & A_{2n}\mathbf{B} \\ \dots & \dots & \dots & \dots \\ A_{m1}\mathbf{B} & A_{m2}\mathbf{B} & \dots & A_{mn}\mathbf{B} \end{pmatrix}$$

where the A_{ij}'s are the entries of $\mathbf{A}$. The tensor product in Eq. (7.9) is needed to ensure that only the components $h_{y,j}$ with the identical indices j are matched in the summations $\sum_{y\in g} \mathbf{E}_p[h_{y,j}]$, represented as the evaluation of $(\mathbf{M}_g \otimes \mathbf{I}_\mathcal{J})\mathbf{E}_p[\boldsymbol{h}]$.

Using Eqs. (7.8) and (7.9), the HME potential can be written as a function of feature expectations $\boldsymbol{t} = \mathbf{E}_p[\mathbb{1}_\mathcal{Y}]$ and $\boldsymbol{u} = \mathbf{E}_p[\boldsymbol{h}]$:

$$\mathrm{U}(\boldsymbol{t},\boldsymbol{u}) = \mathrm{V}(\boldsymbol{t},\mathbf{M}\boldsymbol{t}) + \sum_{y\in\mathcal{Y}} t_y \mathrm{U}_y\left(\frac{\boldsymbol{u}_y}{t_y}\right) + \sum_{g\in\mathcal{G}} (\mathbf{M}_g\boldsymbol{t})\mathrm{U}_g\left(\frac{(\mathbf{M}_g\otimes\mathbf{I}_\mathcal{J})\boldsymbol{u}}{\mathbf{M}_g\boldsymbol{t}}\right) \tag{7.10}$$

where $\boldsymbol{u}_y$ is the slice of vector $\boldsymbol{u}$ corresponding to $\mathbf{E}_p[\boldsymbol{h}_y]$. To simplify Eq. (7.10), we consider an extended group set $\bar{\mathcal{G}} = \mathcal{Y} \cup \mathcal{G}$, containing both groups and classes.[2] We extend the matrix $\mathbf{M}$ into a matrix $\bar{\mathbf{M}} \in \{0,1\}^{\bar{\mathcal{G}}\times\mathcal{Y}}$ by adding the elements $\bar{M}_{yy'} = \mathbb{1}(y = y')$. Thus

$$\bar{\mathbf{M}} = \begin{pmatrix} \mathbf{I}_\mathcal{Y} \\ \mathbf{M} \end{pmatrix} \ . \tag{7.11}$$

Eq. (7.10) can now be simplified to

$$\mathrm{U}(\boldsymbol{t},\boldsymbol{u}) = \mathrm{V}(\bar{\mathbf{M}}\boldsymbol{t}) + \sum_{\bar{g}\in\bar{\mathcal{G}}} (\bar{\mathbf{M}}_{\bar{g}}\boldsymbol{t})\mathrm{U}_{\bar{g}}\left(\frac{(\bar{\mathbf{M}}_{\bar{g}}\otimes\mathbf{I}_\mathcal{J})\boldsymbol{u}}{\bar{\mathbf{M}}_{\bar{g}}\boldsymbol{t}}\right) \ . \tag{7.12}$$

[2]Note that $\mathcal{Y} \cap \mathcal{G} = \emptyset$ since $\mathcal{G}$ consists of subsets of $\mathcal{Y}$. Even if $\mathcal{G}$ contains singletons $\{y\}$, these are formally different from classes y. Thus $|\bar{\mathcal{G}}| = |\mathcal{Y} \cup \mathcal{G}| = |\mathcal{Y}| + |\mathcal{G}|$.

To complete the reduction to the generalized maxent, it remains to argue that U is a closed proper convex function.

First, note that U is proper by the feasibility assumption. To prove convexity and closedness, we use the fact that both of these properties are preserved under linear transformations of arguments (Rockafellar, 1970, Theorems 5.7 and 9.5). Thus, it suffices to rewrite U as

$$\mathrm{U}(\boldsymbol{t},\boldsymbol{u}) = \mathrm{U}'\big(\bar{\mathbf{M}}\boldsymbol{t}, (\bar{\mathbf{M}} \otimes \mathbf{I}_{\mathcal{J}})\boldsymbol{u}\big) \tag{7.13}$$

and prove the convexity and closedness of

$$\mathrm{U}'(\boldsymbol{t}',\boldsymbol{u}') = \mathrm{V}(\boldsymbol{t}') + \sum_{\bar{g}\in\bar{\mathcal{G}}} t'_{\bar{g}}\mathrm{U}_{\bar{g}}(\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}}) \tag{7.14}$$

where $\boldsymbol{t}' \in \mathbb{R}^{\bar{\mathcal{G}}}$, $\boldsymbol{u}' \in \mathbb{R}^{\bar{\mathcal{G}}\times\mathcal{J}}$ (note the difference from $\boldsymbol{t} \in \mathbb{R}^{\mathcal{Y}}$, $\boldsymbol{u} \in \mathbb{R}^{\mathcal{Y}\times\mathcal{J}}$).

Convexity of U' follows by convexity of V and $\mathrm{U}_{\bar{g}}$, because convexity is preserved under perspective transformation, i.e., the operation $(t'_{\bar{g}}, \boldsymbol{u}'_{\bar{g}}) \mapsto t'_{\bar{g}}\mathrm{U}_{\bar{g}}(\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}})$ defining terms of the sum above (see Rockafellar, 1970, page 35; or Boyd and Vandenberghe, 2004, Section 3.2.6). To show that U' is closed, it suffices to argue that it is a sum of closed functions. In our case, these are the function V and *closures* of the terms $t'_{\bar{g}}\mathrm{U}_{\bar{g}}(\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}})$, where closures are the largest closed functions bounding the terms from below (Rockafellar, 1970, Section 7). It can be shown that the closure of the perspective transformation equals the perspective transformation if $t'_{\bar{g}} > 0$ (Rockafellar, 1970, page 67). Otherwise, i.e., for $t'_{\bar{g}} \leq 0$, we have $\mathrm{V}(\boldsymbol{t}') = \infty$ by assumption and hence $\mathrm{U}'(\boldsymbol{u}', \boldsymbol{t}') = \infty$ regardless of the value of the terms in the sum, so we can substitute the closure values for the actual terms without loss of generality. Thus, U' is indeed equal to a sum of closed functions, which completes the reduction to generalized maxent.

In the following sections we place mild assumptions on the potential functions V, U_y, and U_g to obtain specific duality results and generalization guarantees.

7.4 Polyhedral HME

We saw that U is a closed proper convex function of the expectations of $\mathbb{1}_y$, $h_{y,j}$, so the duality results and performance guarantees for generalized maxent apply. However, we would like the duality and the performance guarantees to be interpretable from the point of view of the HME potential functions V, U_y, U_g. In this section, we will express U^* in terms of V^*, U^*_y, and U^*_g. To accomplish this, we will need an additional assumption that V, U_y, and U_g are polyhedral as in the example of ℓ_1-regularized HME. Other choices of hierarchical polyhedral potentials are obtained similar to Sections 3.2.2 and 6.2.

To derive the conjugate of U, we begin by deriving the conjugate of U' defined in Eq. (7.14):

$$\begin{aligned}
\mathrm{U}'^*(\boldsymbol{s}',\boldsymbol{\lambda}') &= \sup_{\boldsymbol{t}',\boldsymbol{u}'}\Big(\boldsymbol{s}'\cdot\boldsymbol{t}'+\boldsymbol{\lambda}'\cdot\boldsymbol{u}'-\mathrm{U}'(\boldsymbol{t}',\boldsymbol{u}')\Big) \\
&= \sup_{\boldsymbol{t}',\boldsymbol{u}'}\Big(\boldsymbol{s}'\cdot\boldsymbol{t}'+\boldsymbol{\lambda}'\cdot\boldsymbol{u}'-\mathrm{V}(\boldsymbol{t}')-\sum_{\bar{g}\in\bar{\mathcal{G}}} t'_{\bar{g}}\cdot\mathrm{U}_{\bar{g}}(\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}})\Big) \qquad (7.15)\\
&= \sup_{\boldsymbol{t}',\boldsymbol{u}'}\Big(\sum_{\bar{g}\in\bar{\mathcal{G}}} t'_{\bar{g}}\cdot\Big[\boldsymbol{\lambda}'_{\bar{g}}\cdot\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}}+s'_{\bar{g}}-\mathrm{U}_{\bar{g}}(\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}})\Big]-\mathrm{V}(\boldsymbol{t}')\Big) \\
&= \sup_{\boldsymbol{t}'}\Big(\sum_{\bar{g}\in\bar{\mathcal{G}}} t'_{\bar{g}}\cdot\Big[s'_{\bar{g}}+\sup_{\boldsymbol{u}''_{\bar{g}}:=\boldsymbol{u}'_{\bar{g}}/t'_{\bar{g}}}\Big(\boldsymbol{\lambda}'_{\bar{g}}\cdot\boldsymbol{u}''_{\bar{g}}-\mathrm{U}_{\bar{g}}(\boldsymbol{u}''_{\bar{g}})\Big)\Big]-\mathrm{V}(\boldsymbol{t}')\Big) \\
&= \sup_{\boldsymbol{t}'}\Big(\sum_{\bar{g}\in\bar{\mathcal{G}}} t'_{\bar{g}}\cdot\Big[s'_{\bar{g}}+\mathrm{U}^*_{\bar{g}}(\boldsymbol{\lambda}'_{\bar{g}})\Big]-\mathrm{V}(\boldsymbol{t}')\Big) \qquad (7.16)\\
&= \mathrm{V}^*\Big(\big\langle s'_{\bar{g}}+\mathrm{U}^*_{\bar{g}}(\boldsymbol{\lambda}'_{\bar{g}})\big\rangle_{\bar{g}\in\bar{\mathcal{G}}}\Big)\ . \qquad (7.17)
\end{aligned}$$

Eq. (7.15) follows by Eq. (7.14), Eqs. (7.16) and (7.17) from the definition of conjugacy.

To derive U^*, we will combine Eqs. (7.13) and (7.17) using the identity for conjugacy under linear transformations, Eq. (2.14). However, we cannot apply Eq. (2.14) immediately, because the matrices in Eq. (7.13) are not square. We will extend the matrices into invertible square matrices, and appropriately extend the vectors $\boldsymbol{t}$ and $\boldsymbol{u}$. Specifically, let

$$\bar{\bar{\mathbf{M}}} = \begin{pmatrix} \mathbf{I}_{\mathcal{Y}} & \mathbf{0}_{\mathcal{Y}}\mathbf{0}_{\mathcal{G}}^\top \\ \mathbf{M} & \mathbf{I}_{\mathcal{G}} \end{pmatrix} \qquad (7.18)$$

where $\mathbf{0}_{\mathcal{Y}}$ and $\mathbf{0}_{\mathcal{G}}$ are all-zero vectors of sizes $|\mathcal{Y}|$ and $|\mathcal{G}|$. Note that $\bar{\bar{\mathbf{M}}}$ is a square matrix in a lower triangular form without zeros on the diagonal, so it is invertible. Comparing with Eq. (7.11), we find that the first argument on the right-hand side of Eq. (7.13), $\bar{\mathbf{M}}\boldsymbol{t}$, can be rewritten as

$$\bar{\mathbf{M}}\boldsymbol{t} = \begin{pmatrix}\mathbf{I}_{\mathcal{Y}}\\ \mathbf{M}\end{pmatrix}\boldsymbol{t} = \begin{pmatrix} \mathbf{I}_{\mathcal{Y}} & \mathbf{0}_{\mathcal{Y}}\mathbf{0}_{\mathcal{G}}^\top \\ \mathbf{M} & \mathbf{I}_{\mathcal{G}} \end{pmatrix}\begin{pmatrix}\boldsymbol{t}\\ \mathbf{0}_{\mathcal{G}}\end{pmatrix} = \bar{\bar{\mathbf{M}}}\begin{pmatrix}\boldsymbol{t}\\ \mathbf{0}_{\mathcal{G}}\end{pmatrix}\ .$$

Similarly,

$$(\bar{\mathbf{M}}\otimes\mathbf{I}_{\mathcal{J}})\boldsymbol{u} = (\bar{\bar{\mathbf{M}}}\otimes\mathbf{I}_{\mathcal{J}})\begin{pmatrix}\boldsymbol{u}\\ \mathbf{0}_{\mathcal{G}\times\mathcal{J}}\end{pmatrix}\ .$$

Thus,

$$\mathrm{U}(\boldsymbol{t},\boldsymbol{u}) = \mathrm{U}''(\boldsymbol{t},\mathbf{0}_{\mathcal{G}},\boldsymbol{u},\mathbf{0}_{\mathcal{G}\times\mathcal{J}}) \qquad (7.19)$$

where, for $\boldsymbol{t}''\in\mathbb{R}^{\bar{\mathcal{G}}}$, $\boldsymbol{u}''\in\mathbb{R}^{\bar{\mathcal{G}}\times\mathcal{J}}$, the function U'' is defined by

$$\mathrm{U}''(\boldsymbol{t}'',\boldsymbol{u}'') = \mathrm{U}'(\bar{\bar{\mathbf{M}}}\boldsymbol{t}'',(\bar{\bar{\mathbf{M}}}\otimes\mathbf{I}_{\mathcal{J}})\boldsymbol{u}'')\ . \qquad (7.20)$$

Next, we will express U''^* in terms of $\mathrm{U}'*$ (using Eq. 2.14) and then U^* in terms of U''^* using the following technical proposition characterizing the conjugates of "slices" of polyhedral functions.

Proposition 7.1. *Let* $\varphi : \mathbb{R}^n \to (-\infty, \infty]$, $\psi : \mathbb{R}^{n+m} \to (-\infty, \infty]$ *be proper polyhedral functions such that* $\varphi(\boldsymbol{u}) = \psi(\boldsymbol{u}, \boldsymbol{0})$. *Then* $\varphi^*(\boldsymbol{\lambda}) = \inf_{\boldsymbol{\eta}'} \psi^*(\boldsymbol{\lambda}, \boldsymbol{\eta}')$.

Proof. Define $\chi : \mathbb{R}^{n+m} \to (0, \infty]$ by

$$\chi(\boldsymbol{u}, \boldsymbol{v}) = \psi(\boldsymbol{u}, \boldsymbol{v}) + \mathrm{I}(\boldsymbol{v} = \boldsymbol{0}) \tag{7.21}$$

where $\boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{v} \in \mathbb{R}^m$. Thus, $\varphi(\boldsymbol{u}) = \inf_{\boldsymbol{v}} \chi(\boldsymbol{u}, \boldsymbol{v})$. Therefore,

$$\begin{aligned} \varphi^*(\boldsymbol{\lambda}) &= \sup_{\boldsymbol{u}} \left[\boldsymbol{\lambda} \cdot \boldsymbol{u} - \varphi(\boldsymbol{u}) \right] \\ &= \sup_{\boldsymbol{u}, \boldsymbol{v}} \left[\boldsymbol{\lambda} \cdot \boldsymbol{u} + \boldsymbol{0} \cdot \boldsymbol{v} - \chi(\boldsymbol{u}, \boldsymbol{v}) \right] = \chi^*(\boldsymbol{\lambda}, \boldsymbol{0}) \ . \end{aligned} \tag{7.22}$$

In order to derive φ^*, it suffices to derive χ^*. First rewrite Eq. (7.21) as

$$\chi(\boldsymbol{u}, \boldsymbol{v}) = \psi(\boldsymbol{u}, \boldsymbol{v}) + \mathrm{I}_{\{\boldsymbol{0}\}}(\boldsymbol{v}) + \mathrm{I}^*_{\{\boldsymbol{0}\}}(\boldsymbol{u}) \ ,$$

noting that $\mathrm{I}^*_{\{\boldsymbol{0}\}}(\boldsymbol{u}) = \boldsymbol{u} \cdot \boldsymbol{0} = 0$. Now, by Eq. (2.16)

$$\begin{aligned} \chi^*(\boldsymbol{\lambda}, \boldsymbol{\eta}) &= \inf_{\boldsymbol{\lambda}', \boldsymbol{\eta}'} \left[\psi^*(\boldsymbol{\lambda}', \boldsymbol{\eta}') + \mathrm{I}^*_{\{\boldsymbol{0}\}}(\boldsymbol{\eta} - \boldsymbol{\eta}') + \mathrm{I}_{\{\boldsymbol{0}\}}(\boldsymbol{\lambda} - \boldsymbol{\lambda}') \right] \\ &= \inf_{\boldsymbol{\lambda}', \boldsymbol{\eta}'} \left[\psi^*(\boldsymbol{\lambda}', \boldsymbol{\eta}') + \mathrm{I}(\boldsymbol{\lambda} = \boldsymbol{\lambda}') \right] = \inf_{\boldsymbol{\eta}'} \psi^*(\boldsymbol{\lambda}, \boldsymbol{\eta}') \ . \end{aligned}$$

Combining with Eq. (7.22) yields the result of the proposition. ■

Hence

$$\mathrm{U}^*(\boldsymbol{s}, \boldsymbol{\lambda}) = \inf_{\substack{\boldsymbol{s}^{(\mathrm{group})} \in \mathbb{R}^{\mathcal{G}} \\ \boldsymbol{\eta} \in \mathbb{R}^{\mathcal{G} \times \mathcal{J}}}} \mathrm{U}''^*(\boldsymbol{s}, \boldsymbol{s}^{(\mathrm{group})}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \tag{7.23}$$

$$= \inf_{\substack{\boldsymbol{s}^{(\mathrm{group})} \in \mathbb{R}^{\mathcal{G}} \\ \boldsymbol{\eta} \in \mathbb{R}^{\mathcal{G} \times \mathcal{J}}}} \mathrm{U}'^* \left(\bar{\bar{\mathbf{M}}}^{-\top} \begin{pmatrix} \boldsymbol{s} \\ \boldsymbol{s}^{(\mathrm{group})} \end{pmatrix}, (\bar{\bar{\mathbf{M}}}^{-\top} \otimes \mathbf{I}_{\mathcal{J}}) \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\eta} \end{pmatrix} \right) \ . \tag{7.24}$$

Eq. (7.23) follows by Proposition 7.1, using the fact that U' is polyhedral (because V and $\mathrm{U}_{\bar{g}}$ are polyhedral and the polyhedral property is preserved by the perspective transformation). Eq. (7.24) follows by Eq. (2.14).

To obtain an expression for U^* in terms of V^*, U^*_y, and U^*_g, it now suffices to combine Eqs. (7.17) and (7.24). Before doing so, we briefly discuss the form of the

matrix $\bar{\bar{\mathbf{M}}}^{-\top}$. This matrix will be denoted $\mathbf{E}$, because it converts class and group parameters s_y, $\boldsymbol{\lambda}_y$, $s_g^{(\text{group})}$, $\boldsymbol{\eta}_g$ appearing in Eq. (7.24), into *class effects* and *group effects* s'_y, $\boldsymbol{\lambda}'_y$, s'_g, $\boldsymbol{\lambda}'_g$ appearing in Eq. (7.17). These effects, rather than the actual parameters, are regularized by the conjugates V^*, U_y^*, U_g^*. To explicitly derive $\mathbf{E}$, note that

$$\bar{\bar{\mathbf{M}}}^{\top}\begin{pmatrix}\mathbf{I}_{\mathcal{Y}} & -\mathbf{M}\\ \mathbf{0}_{\mathcal{G}}\mathbf{0}_{\mathcal{Y}}^{\top} & \mathbf{I}_{\mathcal{G}}\end{pmatrix}=\begin{pmatrix}\mathbf{I}_{\mathcal{Y}} & \mathbf{M}\\ \mathbf{0}_{\mathcal{G}}\mathbf{0}_{\mathcal{Y}}^{\top} & \mathbf{I}_{\mathcal{G}}\end{pmatrix}\begin{pmatrix}\mathbf{I}_{\mathcal{Y}} & -\mathbf{M}\\ \mathbf{0}_{\mathcal{G}}\mathbf{0}_{\mathcal{Y}}^{\top} & \mathbf{I}_{\mathcal{G}}\end{pmatrix}=\mathbf{I}_{\bar{\mathcal{G}}}\ .$$

Thus,

$$\mathbf{E}=\bar{\bar{\mathbf{M}}}^{-\top}=\begin{pmatrix}\mathbf{I}_{\mathcal{Y}} & -\mathbf{M}\\ \mathbf{0}_{\mathcal{G}}\mathbf{0}_{\mathcal{Y}}^{\top} & \mathbf{I}_{\mathcal{G}}\end{pmatrix}\ .$$

Therefore, class and group effects are

$$s'_y=\mathbf{E}_y\begin{pmatrix}\boldsymbol{s}\\ \boldsymbol{s}^{(\text{group})}\end{pmatrix}=s_y-\sum_{g:y\in g}s_g^{(\text{group})}\qquad s'_g=\mathbf{E}_g\begin{pmatrix}\boldsymbol{s}\\ \boldsymbol{s}^{(\text{group})}\end{pmatrix}=s_g^{(\text{group})}\tag{7.25}$$

$$\boldsymbol{\lambda}'_y=(\mathbf{E}_y\otimes\mathbf{I}_{\mathcal{J}})\begin{pmatrix}\boldsymbol{\lambda}\\ \boldsymbol{\eta}\end{pmatrix}=\boldsymbol{\lambda}_y-\sum_{g:y\in g}\boldsymbol{\eta}_g\qquad \boldsymbol{\lambda}'_g=(\mathbf{E}_g\otimes\mathbf{I}_{\mathcal{J}})\begin{pmatrix}\boldsymbol{\lambda}\\ \boldsymbol{\eta}\end{pmatrix}=\boldsymbol{\eta}_g\ .\tag{7.26}$$

Note that the expressions for $\boldsymbol{\lambda}'_y$ and $\boldsymbol{\lambda}'_g$ in Eq. (7.26) are regularized in the dual of ℓ_1-regularized HME, Eq. (7.2). Combining Eqs. (7.17), (7.24), (7.25), and (7.26), we finally obtain the expression for the conjugate potential

$$\mathrm{U}^*(\boldsymbol{s},\boldsymbol{\lambda})=\inf_{\substack{\boldsymbol{s}^{(\text{group})}\in\mathbb{R}^{\mathcal{G}}\\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}}\mathrm{V}^*\left(\left\langle s_y-\sum_{g:y\in g}s_g^{(\text{group})}+\mathrm{U}_y^*\Big(\boldsymbol{\lambda}_y-\sum_{g:y\in g}\boldsymbol{\eta}_g\Big)\right\rangle_{y\in\mathcal{Y}},\left\langle s_g^{(\text{group})}+\mathrm{U}_g^*(\boldsymbol{\eta}_g)\right\rangle_{g\in\mathcal{G}}\right).\tag{7.27}$$

Note that the only step relying on the condition that V, U_y, and U_g are polyhedral is Eq. (7.23). If Eq. (7.23) holds for specific non-polyhedral potentials then Eq. (7.27) holds as well.

7.5 Polyhedral HME with Fixed Class Probabilities

In this section we derive a duality result and a generalization lemma for a class of potentials which includes ℓ_1-regularized HME. Specifically, we study the case when potentials U_y, U_g are polyhedral and V is a point indicator, specified by a probability distribution c on $\mathcal{Y}$,

$$\mathrm{V}(\boldsymbol{t},\boldsymbol{t}^{(\text{group})})=\sum_{y\in\mathcal{Y}}\mathrm{I}(t_y=c(y))+\sum_{g\in\mathcal{G}}\mathrm{I}(t_g^{(\text{group})}=c(g))\ .$$

This setting corresponds to fixing class probabilities to $c(y)$. Using the fact that $p(y) = c(y)$, the objective of HME, Eq. (7.3), can be rewritten as

$$\min_{\substack{p \in \Delta \\ \text{s.t. } p_{\mathcal{Y}} = c_{\mathcal{Y}}}} \left[\mathrm{D}(c \parallel q_{0;\mathcal{Y}}) + \sum_{y \in \mathcal{Y}} c(y) \Big[\mathrm{D}(p_y \parallel q_{0;y}) + \mathrm{U}_y(\mathbf{E}_p[\boldsymbol{f} \mid y]) \Big] + \sum_{g \in \mathcal{G}} c(g) \mathrm{U}_g(\mathbf{E}_p[\boldsymbol{f} \mid g]) \right] \tag{7.28}$$

where $q_{0;\mathcal{Y}}$ is the marginal distribution of q_0 over classes. Note that the first term of Eq. (7.28) is constant. The second term is a sum of single-class maxent objectives, weighted by $c(y)$. The third term is a sum of group constraints, weighted by $c(g)$. Thus, components $c(y)$ are perhaps better understood as measures of importance of individual classes rather than the actual probabilities.

We use Eq. (7.27) to derive the conjugate potential, which will in turn be used to derive a duality result and a generalization lemma. As the first step, note that for $\boldsymbol{s}''' \in \mathbb{R}^{\mathcal{Y}}$ and $\boldsymbol{s}'''^{(\text{group})} \in \mathbb{R}^{\mathcal{G}}$

$$\mathrm{V}^*(\boldsymbol{s}''', \boldsymbol{s}'''^{(\text{group})}) = \sum_{y \in \mathcal{Y}} c(y) s'''_y + \sum_{g \in \mathcal{G}} c(g) s'''^{(\text{group})}_g \quad .$$

Thus the inner expression in Eq. (7.27) can be rewritten as

$$\begin{aligned}
&\mathrm{V}^*\Big(\Big\langle s_y - \sum_{g : y \in g} s_g^{(\text{group})} + \mathrm{U}_y^*\big(\boldsymbol{\lambda}_y - \sum_{g : y \in g} \boldsymbol{\eta}_g\big) \Big\rangle_{y \in \mathcal{Y}}, \Big\langle s_g^{(\text{group})} + \mathrm{U}_g^*(\boldsymbol{\eta}_g) \Big\rangle_{g \in \mathcal{G}} \Big) \\
&\quad = \sum_{y \in \mathcal{Y}} c(y) s_y - \sum_{y \in \mathcal{Y}, g \in \mathcal{G}} \mathbb{1}(y \in g) c(y) s_g^{(\text{group})} + \sum_{y \in \mathcal{Y}} c(y) \mathrm{U}_y^*\big(\boldsymbol{\lambda}_y - \sum_{g : y \in g} \boldsymbol{\eta}_g\big) \\
&\qquad + \sum_{g \in \mathcal{G}} c(g) s_g^{(\text{group})} + \sum_{g \in \mathcal{G}} c(g) \mathrm{U}_g^*(\boldsymbol{\eta}_g) \\
&\quad = \sum_{y \in \mathcal{Y}} c(y) s_y + \sum_{y \in \mathcal{Y}} c(y) \mathrm{U}_y^*\big(\boldsymbol{\lambda}_y - \sum_{g : y \in g} \boldsymbol{\eta}_g\big) + \sum_{g \in \mathcal{G}} c(g) \mathrm{U}_g^*(\boldsymbol{\eta}_g) \quad .
\end{aligned}$$

Hence the expression for V^* does not depend on $\boldsymbol{s}^{(\text{group})}$. By Eq. (7.27), we obtain

$$\mathrm{U}^*(\boldsymbol{s}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{\eta} \in \mathbb{R}^{\mathcal{G} \times \mathcal{J}}} \Big[\sum_{y \in \mathcal{Y}} c(y) s_y + \sum_{y \in \mathcal{Y}} c(y) \mathrm{U}_y^*\big(\boldsymbol{\lambda}_y - \sum_{g : y \in g} \boldsymbol{\eta}_g\big) + \sum_{g \in \mathcal{G}} c(g) \mathrm{U}_g^*(\boldsymbol{\eta}_g) \Big] \quad . \tag{7.29}$$

Now, we are ready to state and prove the duality result for polyhedral HME with fixed class probabilities. We consider the primal objective Eq. (7.28) without the constant term $\mathrm{D}(c \parallel q_{0;\mathcal{Y}})$

$$P_{\text{HME}}(p) = \mathrm{I}(p_{\mathcal{Y}} = c) + \sum_{y \in \mathcal{Y}} c(y) \Big[\mathrm{D}(p_y \parallel q_{0;y}) + \mathrm{U}_y(\mathbf{E}_p[\boldsymbol{f} \mid y]) \Big] + \sum_{g \in \mathcal{G}} c(g) \mathrm{U}_g(\mathbf{E}_p[\boldsymbol{f} \mid g])$$

and show that its minimization is equivalent to the maximization of the weighted version of the objective Q of the single-class maxent (see Section 2.5) with group regularization terms

$$Q_{\mathrm{HME}}(\boldsymbol{\lambda},\boldsymbol{\eta}) = \sum_{y\in\mathcal{Y}} c(y)\Big[-\ln Z_{\boldsymbol{\lambda}_y;y} - \mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y + \textstyle\sum_{g:y\in g}\boldsymbol{\eta}_g\big)\Big] - \sum_{g\in\mathcal{G}} c(g)\mathrm{U}_g^*(-\boldsymbol{\eta}_g) \ .$$

Theorem 7.2. *Let P_{HME}, Q_{HME} be defined as above. Assume that U_y, U_g are polyhedral, $c(y)\neq 0$ for all y, and P_{HME} is proper. Then*

$$\min_{p\in\Delta} P_{\mathrm{HME}}(p) = \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}}\\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}} Q_{\mathrm{HME}}(\boldsymbol{\lambda},\boldsymbol{\eta}) \ . \tag{i}$$

Moreover, for a sequence $\boldsymbol{\lambda}^{(1)},\boldsymbol{\eta}^{(1)},\boldsymbol{\lambda}^{(2)},\boldsymbol{\eta}^{(2)},\cdots$ such that

$$\lim_{t\to\infty} Q_{\mathrm{HME}}(\boldsymbol{\lambda}^{(t)},\boldsymbol{\eta}^{(t)}) = \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}}\\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}} Q_{\mathrm{HME}}(\boldsymbol{\lambda},\boldsymbol{\eta})$$

the sequence of q_t, where $q_t(x,y) = c(y)q_{\boldsymbol{\lambda}_y^{(t)};y}(x)$, has a limit and

$$P_{\mathrm{HME}}\Big(\lim_{t\to\infty} q_t\Big) = \min_{p\in\Delta} P_{\mathrm{HME}}(p) \ . \tag{ii}$$

Note the similarity of Theorem 7.2 (HME duality) and Theorem 2.4 (maxent duality). Indeed, maxent duality will be the crucial step in the proof of HME duality. Before embarking on the proof of Theorem 7.2, we point out why it is not an immediate corollary of maxent duality under the reduction of HME to generalized maxent. Using Eq. (7.10), we find that the objective of generalized maxent is

$$P(p) = \mathrm{D}(p\parallel q_0) + \mathrm{I}(p_{\mathcal{Y}} = c_{\mathcal{Y}}) + \sum_{y\in\mathcal{Y}} \mathbf{E}_p[\mathbb{1}_y]\mathrm{U}_y\left(\frac{\mathbf{E}_p[\boldsymbol{h}_y]}{\mathbf{E}_p[\mathbb{1}_y]}\right) + \sum_{g\in\mathcal{G}} (\mathbf{M}_g\mathbb{1}_{\mathcal{Y}})\mathrm{U}_g\left(\frac{(\mathbf{M}_g\otimes\mathbf{I}_{\mathcal{J}})\mathbf{E}_p[\boldsymbol{h}]}{\mathbf{M}_g\mathbf{E}_p[\mathbb{1}_{\mathcal{Y}}]}\right)$$

$$\begin{aligned} &= \mathrm{D}(p_{\mathcal{Y}}\parallel q_{0;\mathcal{Y}}) + \sum_{y\in\mathcal{Y}} c(y)\mathrm{D}(p_y\parallel q_{0;y}) + \mathrm{I}(p_{\mathcal{Y}} = c_{\mathcal{Y}}) \\ &\qquad + \sum_{y\in\mathcal{Y}} c(y)\mathrm{U}_y\left(\frac{\mathbf{E}_p[\boldsymbol{h}_y]}{c(y)}\right) + \sum_{g\in\mathcal{G}} c(g)\mathrm{U}_g\left(\frac{(\mathbf{M}_g\otimes\mathbf{I}_{\mathcal{J}})\mathbf{E}_p[\boldsymbol{h}]}{c(g)}\right) \end{aligned} \tag{7.30}$$

where Eq. (7.30) follows by replacing $p(y)$ by $c(y)$ since these are identical whenever P is finite. Note that by the reduction, Eq. (7.30) equals the objective of Eq. (7.28), and hence differs from P_{HME} only by the constant $\mathrm{D}(p_{\mathcal{Y}}\parallel q_{0;\mathcal{Y}})$.

We can derive the matching dual objective using the expression for the conjugate

potential, Eq. (7.29),

$$Q(\boldsymbol{s},\boldsymbol{\lambda}) = -\ln Z_{\boldsymbol{s},\boldsymbol{\lambda}} - \inf_{\boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}\Big[-\sum_{y\in\mathcal{Y}} c(y)s_y + \sum_{y\in\mathcal{Y}} c(y)\mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y + \sum_{g:y\in g}\boldsymbol{\eta}_g\big) + \sum_{g\in\mathcal{G}} c(g)\mathrm{U}_g^*(-\boldsymbol{\eta}_g)\Big] \tag{7.31}$$

where $Z_{\boldsymbol{s},\boldsymbol{\lambda}}$ is the normalization constant of the Gibbs distribution $q_{\boldsymbol{s},\boldsymbol{\lambda}}$ defined over $\mathcal{X}\times\mathcal{Y}$

$$q_{\boldsymbol{s},\boldsymbol{\lambda}}(x,y) = \frac{q_0(x,y)e^{\sum_{y'\in\mathcal{Y}}\left[s_{y'}\mathbb{1}_{y'}(x,y)+\boldsymbol{\lambda}_{y'}\cdot\boldsymbol{h}_{y'}(x,y)\right]}}{Z_{\boldsymbol{s},\boldsymbol{\lambda}}} = \frac{q_0(x,y)e^{s_y+\boldsymbol{\lambda}_y\cdot\boldsymbol{f}(x,y)}}{Z_{\boldsymbol{s},\boldsymbol{\lambda}}} \ .$$

(The last equality follows from the definition of $\mathbb{1}_y$ and $\boldsymbol{h}_y$.)

Notice that Q is a function of $\boldsymbol{s}$ and $\boldsymbol{\lambda}$, whereas Q_{HME} is a function of $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$. The proof below shows maximization of Q and Q_{HME} is equivalent up to the constant difference of $\mathrm{D}(p_{\mathcal{Y}} \,\|\, q_{0;\mathcal{Y}})$ as in the primal.

Proof of Theorem 7.2. As mentioned above, to prove part (i), we use the reduction to generalized maxent and appeal to maxent duality:

$$\begin{aligned}
&\mathrm{D}(c \,\|\, q_{0;\mathcal{Y}}) + \min_{p\in\Delta} P_{\mathrm{HME}}(p) \\
&= \min_{p\in\Delta} P(p) = \sup_{\substack{\boldsymbol{s}\in\mathbb{R}^{\mathcal{Y}} \\ \boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}}}} Q(\boldsymbol{s},\boldsymbol{\lambda}) && (7.32)\\
&= \sup_{\substack{\boldsymbol{s}\in\mathbb{R}^{\mathcal{Y}} \\ \boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}}}} \Big[-\ln Z_{\boldsymbol{s},\boldsymbol{\lambda}} - \inf_{\boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}\Big[-\sum_{y\in\mathcal{Y}} c(y)s_y + \sum_{y\in\mathcal{Y}} c(y)\mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y + \sum_{g:y\in g}\boldsymbol{\eta}_g\big) + \sum_{g\in\mathcal{G}} c(g)\mathrm{U}_g^*(-\boldsymbol{\eta}_g)\Big]\Big] \\
&= \sup_{\substack{\boldsymbol{s}\in\mathbb{R}^{\mathcal{Y}} \\ \boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}}}} \sup_{\boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}} \Big[-\ln Z_{\boldsymbol{s},\boldsymbol{\lambda}} + \sum_{y\in\mathcal{Y}} c(y)s_y - \sum_{y\in\mathcal{Y}} c(y)\mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y + \sum_{g:y\in g}\boldsymbol{\eta}_g\big) - \sum_{g\in\mathcal{G}} c(g)\mathrm{U}_g^*(-\boldsymbol{\eta}_g)\Big] \\
&= \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}} \\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}} \Big[\mathrm{D}(c \,\|\, q_{0;\mathcal{Y}}) + \sum_{y\in\mathcal{Y}} c(y)\Big[-\ln Z_{\boldsymbol{s},\boldsymbol{\lambda}} - \mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y + \sum_{g:y\in g}\boldsymbol{\eta}_g\big)\Big] - \sum_{g\in\mathcal{G}} c(g)\mathrm{U}_g^*(-\boldsymbol{\eta}_g)\Big] && (7.33)\\
&= \mathrm{D}(c \,\|\, q_{0;\mathcal{Y}}) + \sup_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^{\mathcal{Y}\times\mathcal{J}} \\ \boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}}} Q_{\mathrm{HME}}(\boldsymbol{\lambda},\boldsymbol{\eta}) \ . && (7.34)
\end{aligned}$$

Eq. (7.32) follows by maxent duality where P and Q are given in Eqs. (7.30) and (7.31). Eq. (7.33) follows by setting the partial derivatives of the objective with respect to s_y equal to zero. Specifically, the partial derivatives with respect to s_y are equal to zero if and only if

$$s_y = \ln Z_{\boldsymbol{s},\boldsymbol{\lambda}} + \ln\frac{c(y)}{q_0(y)} - \ln Z_{\boldsymbol{\lambda}_y;y} \tag{7.35}$$

and hence the term $\sum_{y\in\mathcal{Y}} c(y)s_y$ becomes

$$\ln Z_{\boldsymbol{s},\boldsymbol{\lambda}} + \mathrm{D}(c \parallel q_{0;\mathcal{Y}}) - \sum_{y\in\mathcal{Y}} c(y) \ln Z_{\boldsymbol{\lambda}_y;y} \ ,$$

yielding Eq. (7.33). Eq. (7.34) follows from the definition of Q_{HME} and proves part (i) of the theorem.

To prove part (ii), we use maxent duality as well. We set $\boldsymbol{s}^{(t)}$ according to Eq. (7.35) and show that the resulting sequence $Q(\boldsymbol{s}^{(t)}, \boldsymbol{\lambda}^{(t)})$ maximizes the generalized dual. First, notice that $Q(\boldsymbol{s}^{(t)}, \boldsymbol{\lambda}^{(t)})$ is bounded from below by $\mathrm{D}(c \parallel q_{0;\mathcal{Y}}) + Q_{\mathrm{HME}}(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\eta}^{(t)})$:

$$\begin{aligned}
&Q(\boldsymbol{s}^{(t)}, \boldsymbol{\lambda}^{(t)}) \\
&= -\ln Z_{\boldsymbol{s}^{(t)},\boldsymbol{\lambda}^{(t)}} - \inf_{\boldsymbol{\eta}\in\mathbb{R}^{\mathcal{G}\times\mathcal{J}}} \left[-\sum_{y\in\mathcal{Y}} c(y) s_y^{(t)} + \sum_{y\in\mathcal{Y}} c(y) \mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y^{(t)} + \sum_{g:y\in g} \boldsymbol{\eta}_g\big) + \sum_{g\in\mathcal{G}} c(g) \mathrm{U}_g^*(-\boldsymbol{\eta}_g) \right] \\
&\geq -\ln Z_{\boldsymbol{s}^{(t)},\boldsymbol{\lambda}^{(t)}} - \left[-\sum_{y\in\mathcal{Y}} c(y) s_y^{(t)} + \sum_{y\in\mathcal{Y}} c(y) \mathrm{U}_y^*\big(-\boldsymbol{\lambda}_y^{(t)} + \sum_{g:y\in g} \boldsymbol{\eta}_g^{(t)}\big) + \sum_{g\in\mathcal{G}} c(g) \mathrm{U}_g^*(-\boldsymbol{\eta}_g^{(t)}) \right] \\
&= \mathrm{D}(c \parallel q_{0;\mathcal{Y}}) + Q_{\mathrm{HME}}(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\eta}^{(t)}) \ . \qquad (7.36)
\end{aligned}$$

Eq. (7.36) follows similarly to Eqs. (7.33) and (7.34), since $\boldsymbol{s}^{(t)}$ is set according to Eq. (7.35). Taking the limit on the right-hand side and the limit inferior on the left-hand side of Eq. (7.36) yields

$$\begin{aligned}
\liminf_{t\to\infty} Q(\boldsymbol{s}^{(t)}, \boldsymbol{\lambda}^{(t)}) &\geq \mathrm{D}(c \parallel q_{0;\mathcal{Y}}) + \lim_{t\to\infty} Q_{\mathrm{HME}}(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\eta}^{(t)}) \\
&= \mathrm{D}(c \parallel q_{0;\mathcal{Y}}) + \sup_{\boldsymbol{\lambda},\boldsymbol{\eta}} Q_{\mathrm{HME}}(\boldsymbol{\lambda}, \boldsymbol{\eta}) \qquad (7.37)
\end{aligned}$$

where the last equality follows by the assumption of the theorem. Previously, in Eqs. (7.32)–(7.34), we showed that

$$\sup_{\boldsymbol{s},\boldsymbol{\lambda}} Q(\boldsymbol{s}, \boldsymbol{\lambda}) = \mathrm{D}(c \parallel q_{0;\mathcal{Y}}) + \sup_{\boldsymbol{\lambda},\boldsymbol{\eta}} Q_{\mathrm{HME}}(\boldsymbol{\lambda}, \boldsymbol{\eta}) \ . \qquad (7.38)$$

Combining Eqs. (7.37) and (7.38), we thus obtain that $Q(\boldsymbol{\lambda}^{(t)}, \boldsymbol{s}^{(t)})$ maximizes the generalized dual. Therefore, the distributions $q_{\boldsymbol{s}^{(t)},\boldsymbol{\lambda}^{(t)}}$ converge to the primal solution. However, for our choice of $\boldsymbol{s}^{(t)}$, we have $q_{\boldsymbol{s}^{(t)},\boldsymbol{\lambda}^{(t)}} = q_t$, hence part (ii) follows. ■

Similar to the dual of generalized maxent, the HME dual Q_{HME} can be rewritten in a shifted form

$$Q_{\mathrm{HME}}(\boldsymbol{\lambda}, \boldsymbol{\eta}) = \sum_{y\in\mathcal{Y}} c(y) \Big[-\mathrm{L}_{r_y}(\boldsymbol{\lambda}_y; y) - \mathrm{U}_{y;r}^*\big(\boldsymbol{\lambda}_y - \textstyle\sum_{g:y\in g} \boldsymbol{\eta}_g\big) \Big] - \sum_{g\in\mathcal{G}} c(g) \mathrm{U}_{k;r}^*(\boldsymbol{\eta}_g)$$

where

$$\mathrm{L}_{r_y}(\boldsymbol{\lambda}_y; y) = \mathbf{E}_{r_y}\left[-\ln \frac{q_{\boldsymbol{\lambda}_y; y}}{q_{0;y}}\right]$$

is log loss of the class distributions $q_{\boldsymbol{\lambda}_y;y}$ on r_y, and $\mathrm{U}^*_{y;r}$, $\mathrm{U}^*_{g;r}$ are conjugates of potentials shifted according to conditional expectations:

$$\mathrm{U}_{y;r}(\boldsymbol{u}) = \mathrm{U}_y(\mathbf{E}_r[\boldsymbol{f} \mid y] - \boldsymbol{u}) \ , \qquad \mathrm{U}_{g;r}(\boldsymbol{u}) = \mathrm{U}_g(\mathbf{E}_r[\boldsymbol{f} \mid g] - \boldsymbol{u}) \ .$$

If we shift the dual to $\tilde{\pi}$ then it can be interpreted as a regularized log likelihood problem (similar to the single-class case discussed in Section 2.5.2). Substituting the box constraints, and setting $c = \tilde{\pi}_{\mathcal{Y}}$, we obtain ℓ_1-regularized HME and its dual from Section 7.1.

In addition to the duality, we can also prove a generalization lemma, similar to the one derived in Section 3.1. Analogous to single-class maxent, we compare performance of HME solutions against the best performance among all Gibbs distributions, weighted by $c(y)$. Here, we only state a version of Lemma 3.1(ii), but the remaining parts can be derived similarly.

Lemma 7.3 (HME Generalization Lemma)**.** *Let $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{\eta}}$ maximize Q_{HME}. Then for arbitrary $\boldsymbol{\lambda}^\star$, $\boldsymbol{\eta}^\star$*

$$\begin{aligned}\sum_{y\in\mathcal{Y}} c(y)\mathrm{L}_{\pi_y}(\hat{\boldsymbol{\lambda}}_y; y) \leq \sum_{y\in\mathcal{Y}} c(y)\mathrm{L}_{\pi_y}(\boldsymbol{\lambda}^\star_y; y) \\ + \sum_{y\in\mathcal{Y}} c(y)\Big[2\mathrm{U}_{y;\tilde{\pi}}(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f} \mid y] - \mathbf{E}_{\pi}[\boldsymbol{f} \mid y]) \\ + \mathrm{U}^*_{y;\tilde{\pi}}\big(\boldsymbol{\lambda}^\star_y - \textstyle\sum_{g:y\in g} \boldsymbol{\eta}^\star_g\big) + \mathrm{U}^*_{y;\tilde{\pi}}\big(-\boldsymbol{\lambda}^\star_y + \textstyle\sum_{g:y\in g} \boldsymbol{\eta}^\star_g\big)\Big] \\ + \sum_{g\in\mathcal{G}} c(g)\Big[2\mathrm{U}_{g;\tilde{\pi}}(\mathbf{E}_{\tilde{\pi}}[\boldsymbol{f} \mid g] - \mathbf{E}_{\pi}[\boldsymbol{f} \mid g]) + \mathrm{U}^*_{g;\tilde{\pi}}(\boldsymbol{\eta}^\star_g) + \mathrm{U}^*_{g;\tilde{\pi}}(-\boldsymbol{\eta}^\star_g)\Big] \ .\end{aligned}$$

Proof. The result follows immediately by Lemma 3.1(ii). ■

In the next section, we will apply the foregoing general results to ℓ_1-regularized HME.

7.6 ℓ_1-Regularized HME

We consider a hierarchical generalization of ℓ_1-regularized maxent. Specifically, we consider fixed class probabilities, $p(y) = c(y)$, and the box potentials representing

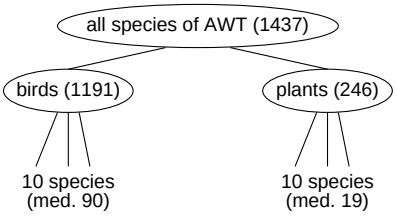


Figure 7.1. *The hierarchy of species in the Australian wet tropics dataset.* Numbers in parentheses indicate the number of training records. At the lowest level, we list only the number of species and report the median number of training records.

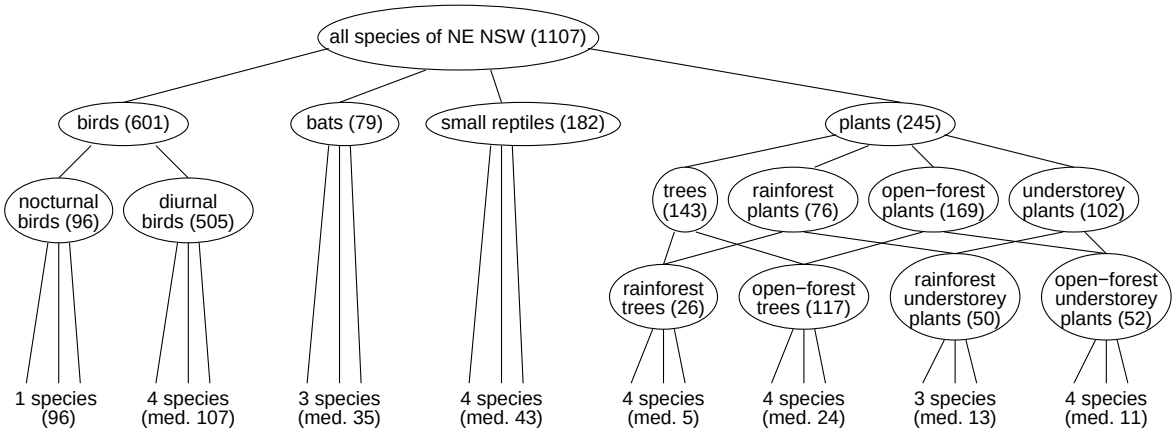


Figure 7.2. *The hierarchy of species in the North-east New South Wales dataset.* Numbers in parentheses indicate the number of training records. At the lowest level, we list only the number of species and report the median number of training records. Note that the children of *plants* correspond to overlapping groups. This hierarchy, therefore, cannot be represented as a tree.

inequality constraints

$$\left|\mathbf{E}_{\tilde{\pi}}[f_j \mid y] - \mathbf{E}_p[f_j \mid y]\right| \leq \beta_y \text{ for all } y \in \mathcal{Y},\ j \in \mathcal{J} \tag{7.39}$$

$$\left|\mathbf{E}_{\tilde{\pi}}[f_j \mid g] - \sum_{y \in g} \tilde{\pi}(y \mid g)\mathbf{E}_p[f_j \mid y]\right| \leq \beta_g \text{ for all } g \in \mathcal{G},\ j \in \mathcal{J}. \tag{7.40}$$

This slightly differs from the version introduced in Section 7.1. For simplicity, we assume that constraint widths depend only on the group or class, but not on the feature index j, and we allow class importance $c(y)$ to differ from $\tilde{\pi}(y)$.

HME with class importance can be interpreted as follows: samples (x_i, y_i) come from an unknown distribution π, but our goal is to perform well relative to the distribution μ, which weights individual classes according to their importance,

$$\mu(x, y) = c(y)\pi(x \mid y) \ .$$

Equivalently, $\mu(y) = c(y)$ and $\mu(x \mid y) = \pi(x \mid y)$.

Under this interpretation, Eq. (7.39) reflects the assumption $\mu(x \mid y) = \pi(x \mid y)$ and captures the approximation

$$\mathbf{E}_{\tilde{\pi}}[f_j \mid y] \approx \mathbf{E}_{\pi}[f_j \mid y] = \mathbf{E}_{\mu}[f_j \mid y] \ .$$

Eq. (7.40) expresses the approximation

$$\mathbf{E}_{\pi}[f_j \mid g] \approx \sum_{y \in g} \tilde{\pi}(y \mid g) \mathbf{E}_{\pi}[f_j \mid y] = \sum_{y \in g} \tilde{\pi}(y \mid g) \mathbf{E}_{\mu}[f_j \mid y] \ .$$

Notice that the potentials U_g resulting from Eq. (7.40) are not in terms of feature expectations conditioned on the group membership. In order to derive an HME formulation, instead of f_j we consider

$$f'_j(x, y) = \frac{\tilde{\pi}(y)}{c(y)} f_j(x, y) \ .$$

The constraints of Eqs. (7.39) and (7.40) can then be rewritten as

$$\left| \mathbf{E}_{\tilde{\pi}}[f_j \mid y] - \frac{c(y)}{\tilde{\pi}(y)} \mathbf{E}_p[f'_j \mid y] \right| \leq \beta_y \text{ for all } y \in \mathcal{Y},\ j \in \mathcal{J} \tag{7.41}$$

$$\left| \mathbf{E}_{\tilde{\pi}}[f_j \mid g] - \frac{c(g)}{\tilde{\pi}(g)} \mathbf{E}_p[f'_j \mid g] \right| \leq \beta_g \text{ for all } g \in \mathcal{G},\ j \in \mathcal{J}. \tag{7.42}$$

Applying the duality results of the previous section, we obtain the dual of ℓ_1-regularized HME.

Theorem 7.4. *Let* $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{\mathcal{Y} \times \mathcal{J}}$, $\hat{\boldsymbol{\eta}} \in \mathbb{R}^{\mathcal{G} \times \mathcal{J}}$ *optimize*

$$\begin{aligned} \sup_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^{\mathcal{Y} \times \mathcal{F}} \\ \boldsymbol{\eta} \in \mathbb{R}^{\mathcal{G} \times \mathcal{F}}}} \Bigg\{ & \frac{1}{m} \sum_{i=1}^{m} \frac{c(y_i)}{\tilde{\pi}(y_i)} \ln q_{\boldsymbol{\lambda}_{y_i; y_i}}(x_i) \\ & - \sum_{y \in \mathcal{Y}} c(y) \beta_y \left\| \boldsymbol{\lambda}_y - \sum_{g : y \in g} \frac{\tilde{\pi}(y \mid g)}{c(y \mid g)} \boldsymbol{\eta}_g \right\|_1 - \sum_{g \in \mathcal{G}} c(g) \beta_g \|\boldsymbol{\eta}_g\|_1 \Bigg\} \ . \end{aligned} \tag{7.43}$$

Then $p(x, y) = c(y) q_{\hat{\boldsymbol{\lambda}}_{y;y}}(x)$ *minimizes* P_{HME} *with the constraints* (7.39) *and* (7.40)*.*

Proof. The result follows by Theorem 7.2, applied to features f'_j and constraints in Eqs. (7.41) and (7.42), using the dual objective shifted to the empirical distribution. ■

7.6.1 Performance Guarantees

Performance guarantees for ℓ_1-regularized HME can be derived similar to the guarantees of Sections 3.2.1 and 3.4. For example, for a finite set of features bounded in

$[0,1]$ we can derive an analog of Theorem 3.3. Below, we use m_y and m_g to denote the number of samples with $y_i = y$ and $y_i \in g$.

Theorem 7.5. *Assume that $f_j : \mathcal{X} \times \mathcal{Y} \to [0,1]$. Let $\delta > 0$ and let $\hat{\boldsymbol{\lambda}}$ maximize the regularized likelihood of Eq. (7.43) with $\beta_y = \beta_0/\sqrt{m_y}, \beta_g = \beta_0/\sqrt{m_g}$ where*

$$\beta_0 = \sqrt{\ln(2|\mathcal{Y}||\mathcal{J}| + 2|\mathcal{G}||\mathcal{J}|)/2} \ .$$

Then with probability at least $1-\delta$, for all $\boldsymbol{\lambda}^\star \in \mathbb{R}^{\mathcal{Y}\times\mathcal{J}}, \boldsymbol{\eta}^\star \in \mathbb{R}^{\mathcal{G}\times\mathcal{J}}$,

$$\sum_{y\in\mathcal{Y}} c(y)\mathrm{L}_{\pi_y}(\hat{\boldsymbol{\lambda}}_y; y) \le \sum_{y\in\mathcal{Y}} c(y)\mathrm{L}_{\pi_y}(\boldsymbol{\lambda}^\star_y; y) + 2\beta_0 \sum_{y\in\mathcal{Y}} \frac{c(y)}{\sqrt{m_y}} \left\| \boldsymbol{\lambda}^\star_y - \sum_{g: y\in g} \frac{\tilde{\pi}(y \mid g)}{c(y \mid g)} \boldsymbol{\eta}^\star_g \right\|_1 + 2\beta_0 \sum_{g\in\mathcal{G}} \frac{c(g)}{\sqrt{m_g}} \left\| \boldsymbol{\eta}^\star_g \right\|_1 \ .$$

Proof. Instead of drawing pairs (x_i, y_i) independently from π, we first draw y_i's independently from π and then draw each x_i from π_{y_i}. It suffices to show that for any choice of y_i's, the statement of the theorem is true with probability at least $1-\delta$ over the draw of x_i's. We first consider the constraints on group expectations $\mathbf{E}_{\tilde{\pi}}[f_j \mid g]$. If the y_i's are fixed then for an arbitrary f_j and g, the empirical mean $\mathbf{E}_{\tilde{\pi}}[f_j \mid g]$ is an average of m_g independent (but not identically distributed!) random variables bounded in $[0,1]$. Expectation of this empirical mean, conditioned on the y_i's, is

$$\sum_{y\in g} \tilde{\pi}(y \mid g)\mathbf{E}_\pi[f_j \mid y] \ .$$

Thus, by Hoeffding's inequality, the probability that the deviation

$$\left| \mathbf{E}_{\tilde{\pi}}[f_j \mid g] - \sum_{y\in g} \tilde{\pi}(y \mid g)\mathbf{E}_\pi[f_j \mid y] \right|$$

exceeds β_g is at most $\delta/(|\mathcal{Y}||\mathcal{J}| + |\mathcal{G}||\mathcal{J}|)$. Similarly, the probability that any particular constraint on the class expectation $\mathbf{E}_{\tilde{\pi}}[f_j \mid y]$ is not satisfied is at most $\delta/(|\mathcal{Y}||\mathcal{J}|+|\mathcal{G}||\mathcal{J}|)$. Hence, by the union bound, the probability that this will happen for any $g \in \mathcal{G}$, $j \in \mathcal{J}$ or $y \in \mathcal{Y}$, $j \in \mathcal{J}$ is at most δ. ∎

Theorem 7.5 quantifies the benefits of multiple-density estimation. First, notice that even when a moderately large number of groups is introduced, β_0 increases only slowly. Thus, the performance does not deteriorate compared with an empty hierarchy, as long as the number of groups grows at most polynomially with the number of classes. In most applications, the number of groups is much smaller than the number of classes, so the increase in β_0 is relatively small.

Next, we show how group information improves learning. Consider a simple example of estimating distributions of several bird species with an equal number of occurrences and equal importance, i.e., $m_y = m/|\mathcal{Y}|$ and $c(y) = 1/|\mathcal{Y}|$ for all y. Further, assume that distributions of these birds are similarly influenced by about half the features, and distinctly influenced by the other half. For example, the birds are influenced in the same way by precipitation and vegetation, but different birds respond differently to temperature. Denote the first subset of features as *shared* and the second subset of features as *distinct*. We compare how our generalization guarantees change if we introduce the group *birds*.

First, fix parameters $\boldsymbol{\lambda}_y^\star$ of the optimal Gibbs distributions. Since species depend similarly on *shared*, we assume that slices of parameters $\boldsymbol{\lambda}_{y,shared}^\star$ corresponding to *shared* are roughly equal; denote the shared parameter values as $\boldsymbol{\lambda}_{shared}^\star$. For an empty hierarchy, the gap between maxent solutions and best Gibbs distributions weighted by $c(y)$ is

$$\begin{aligned} 2\beta_0 \sum_{y\in\mathcal{Y}} \frac{c(y)}{\sqrt{m_y}} \left\|\boldsymbol{\lambda}_y^\star\right\|_1 &= \frac{2\beta_0}{\sqrt{m|\mathcal{Y}|}} \sum_{y\in\mathcal{Y}} \left\|\boldsymbol{\lambda}_y^\star\right\|_1 \\ &= \frac{2\beta_0}{\sqrt{m|\mathcal{Y}|}} \sum_{y\in\mathcal{Y}} \left\|\boldsymbol{\lambda}_{y,distinct}^\star\right\|_1 + 2\beta_0 \sqrt{\frac{|\mathcal{Y}|}{m}} \left\|\boldsymbol{\lambda}_{shared}^\star\right\|_1 . \end{aligned} \tag{7.44}$$

Now, add the group *birds*, and set $\boldsymbol{\eta}_{birds,shared}^\star = \boldsymbol{\lambda}_{shared}^\star$ and $\boldsymbol{\eta}_{birds,distinct}^\star = \mathbf{0}$. According to Theorem 7.5, the gap between maxent solutions and best Gibbs distributions is now

$$\begin{aligned} 2\beta_0' \sum_{y\in\mathcal{Y}} \frac{c(y)}{\sqrt{m_y}} \left\|\boldsymbol{\lambda}_y^\star - \boldsymbol{\eta}_{birds}^\star\right\|_1 + 2\beta_0' \frac{c(birds)}{\sqrt{m_{birds}}} \|\boldsymbol{\eta}_{birds}^\star\| \\ = \frac{2\beta_0'}{\sqrt{m|\mathcal{Y}|}} \sum_{y\in\mathcal{Y}} \left\|\boldsymbol{\lambda}_{y,distinct}^\star\right\|_1 + \frac{2\beta_0'}{\sqrt{m}} \left\|\boldsymbol{\lambda}_{shared}^\star\right\|_1 . \end{aligned} \tag{7.45}$$

First note that the multipliers β_0' in Eq. (7.45) are slightly larger than β_0 in Eq. (7.44), as the number of groups has increased from zero to one. Apart from that, the first term of Eq. (7.45), accounting for distinct parameters of the bird species, is identical to the first term of Eq. (7.44). On the other hand, the second term, accounting for shared parameters of the bird species, is effectively divided by the square root of the number of species, taking advantage of the group information. Already for a moderate number of species, for example, 10 or 20, this decrease may be quite significant. Assuming that relevance of *distinct* is similar to the relevance of *shared*, i.e., $\|\boldsymbol{\lambda}_{y,distinct}\|_1 \approx \|\boldsymbol{\lambda}_{y,shared}\|_1$, the gap in performance between maxent distributions and best Gibbs distributions is reduced almost twofold. This means that by using

the group information, we obtain predictions which would require four times as many samples in single-class estimation. If *shared* features are more relevant than *distinct* features then the gap shrinks by an even larger amount, resulting in even more significant savings. Thus, as intuitively expected, larger amounts of information shared within groups yield larger savings over single-class maxent. Similarly, larger numbers of samples within groups yield larger savings.

We briefly discuss dependence of the guarantee of Theorem 7.5 on the number of features $|\mathcal{J}|$. Mainly notice that the guarantee grows very moderately with the number of features. In particular, as long as the number of features grows subexponentially with the number of training examples, our bound is nontrivial. However, when using data on species with very few samples, it may be necessary to restrict the number of features for some species. In the *AWT* example, this may mean dropping constraints on quadratic features for species with an insufficient number of records. Expectations of quadratic features can be still constrained, conditioned on groups that contain these species, provided that the total number of samples within the groups is sufficiently large.

In Theorem 7.5, we used Hoeffding's inequality and the union bound. Using other techniques, it is possible to prove bounds for potentially infinite feature classes. For instance, when $\mathcal{F}$ is a class of binary features with VC dimension d then Theorem A.4 yields a version of Theorem 7.5. The only change is in setting β_0 according to

$$\beta_0 = \sqrt{32\left[d\ln\left(\frac{em}{d}\right)+\ln\left(\frac{8|\mathcal{Y}|+8|\mathcal{G}|}{\delta}\right)\right]}\ . \tag{7.46}$$

Thus, we obtain guarantees for infinite feature classes, analogous to those of Section 3.4.

7.6.2 ℓ_1-Regularized HME as MAP with a Hierarchical Prior

So far, we have considered two interpretations of the HME problem. The first interpretation is the maximization of entropy subject to constraints on conditional expectations. The second interpretation is the maximization of regularized log likelihood. Here, we introduce a third interpretation. We show that when $\mathcal{G}$ describes a tree hierarchy, HME can be viewed as maximum *a posteriori* under a hierarchical Laplace prior. The HME interpretation is more general since it allows arbitrary groups. In addition, HME guides the process of choosing hyperparameters and provides insights into generalization properties.

In this section, we limit our attention to tree hierarchies, such as the *AWT* hierarchy in Fig. 7.1. For tree hierarchies, it is natural to set up a hierarchical model, in

which we associate a vector of Gibbs-distribution parameters $\boldsymbol{\lambda}_n$ with each node n. Let $\mathcal{N}$ denote the set of all nodes in the hierarchy, including leaves y corresponding to our individual classes. A hierarchical Laplace prior, conditioned on $y_1,\ldots,y_m$, can be specified as

$$\boldsymbol{\lambda}_{root} \sim e^{-\alpha_{root}\|\boldsymbol{\lambda}_{root}\|_1} \tag{7.47}$$

$$\boldsymbol{\lambda}_n \mid \boldsymbol{\lambda}_{parent(n)} \sim e^{-\alpha_n\|\boldsymbol{\lambda}_n-\boldsymbol{\lambda}_{parent(n)}\|_1} \text{ for all } n \neq root \tag{7.48}$$

$$x_i \mid \boldsymbol{\lambda}_{y_i} \sim q_{\boldsymbol{\lambda}_{y_i};y_i}(x_i) \text{ for all } i. \tag{7.49}$$

This corresponds to the directed graphical model with structure identical to the hierarchy, with a separate random variable $\boldsymbol{\lambda}_n$ assigned to each node. The root is distributed according to Eq. (7.47), the remaining nodes depend on their parents according to Eq. (7.48), and observations, described by Eq. (7.49), are attached at the bottom.

For example, in *AWT*, the process of drawing samples $x_1,\ldots,x_m$ given $y_1,\ldots,y_m$ can be described as first drawing the parameter $\boldsymbol{\lambda}_{all\ species}$ according to its prior, then choosing $\boldsymbol{\lambda}_{birds}$ and $\boldsymbol{\lambda}_{plants}$ conditioned on $\boldsymbol{\lambda}_{all\ species}$, then drawing $\boldsymbol{\lambda}_y$ conditioned on the respective groups, such as $\boldsymbol{\lambda}_{golden\ bowerbird}$ conditioned on $\boldsymbol{\lambda}_{birds}$, and finally choosing observations x_i in which $y_i = golden\ bowerbird$, conditioned on $\boldsymbol{\lambda}_{golden\ bowerbird}$.

To derive the equivalence of Eq. (7.43) with a hierarchical Laplace prior, we set class importance equal to empirical probabilities and multiply the objective by m:

$$\sum_{i=1}^{m} \ln q_{\boldsymbol{\lambda}_{y_i};y_i}(x_i) - \sum_{y\in\mathcal{Y}}\left(m_y\beta_y\left\|\boldsymbol{\lambda}_y - \sum_{g:y\in g}\boldsymbol{\eta}_g\right\|_1\right) - \sum_{g\in\mathcal{G}}\left(m_g\beta_g\|\boldsymbol{\eta}_g\|_1\right) \ . \tag{7.50}$$

To show that the regularization in Eq. (7.50) corresponds to the hierarchical prior described above, we identify each inner node n with the set $g(n) \subseteq \mathcal{Y}$ containing all classes y which are descendants of n. We set $\mathcal{G} = \{g(n) : n \text{ is an inner node}\}$ and establish the correspondence by setting $\boldsymbol{\lambda}_n$, for each inner node n, equal to the sum of contributions $\boldsymbol{\eta}_{g(n')}$ over n' on the path from the root to the node n. The second and third terms in Eq. (7.50) then become

$$-\sum_{y\in\mathcal{Y}}\left(m_y\beta_y\left\|\boldsymbol{\lambda}_y - \boldsymbol{\lambda}_{parent(y)}\right\|_1\right) - m_{root}\beta_{root}\left\|\boldsymbol{\lambda}_{root}\right\|_1 - \sum_{n\in\mathcal{N}\setminus\mathcal{Y}\setminus\{root\}}\left(m_n\beta_n\left\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}_{parent(n)}\right\|_1\right)$$

where m_n and β_n are shorthand for $m_{g(n)}$ and $\beta_{g(n)}$. The equivalence with the hierarchical Laplace prior is now obtained by setting $\alpha_n = m_n\beta_n$. Thus, similar to the single-class case, maximizing the regularized log likelihood corresponds to maximizing the posterior.

7.7 Experiments

We evaluate HME, specifically ℓ_1-regularized HME with fixed class importance, on synthetic and real data. In both cases, we use SUMMET of Section 4.1. Class importance in all of our experiments equals empirical probabilities. We use variance-based regularization parameters similar to the previous chapters

$$\beta_{y,j} = \beta_0 \sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j \mid y]/m_y} \tag{7.51}$$

$$\beta_{g,j} = \beta_0 \sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j \mid g]/m_g} \ , \tag{7.52}$$

where β_0 is a single tuning parameter and $\mathbf{V}'_{\tilde{\pi}}[f_j \mid y]$, $\mathbf{V}'_{\tilde{\pi}}[f_j \mid g]$ are unbiased empirical estimates of f_j's variance within a class y or a group g.

7.7.1 Synthetic Data

Experimental Design

We first study a synthetic toy-example which simulates species-distribution modeling. We consider a synthetic map consisting of 100 pixels described by two features: precipitation (*prec*) and temperature (*temp*). Values of *prec* are equally spaced in $[0,1]$ and values of *temp* are defined as $temp = (2 \cdot prec - 1)^2$ (we make no claims about physical plausibility of this model). We study two synthetic species: *icebird* and *sunbird*. Both prefer low precipitation, but they differ in their temperature requirements: *icebird* prefers low temperatures while *sunbird* prefers high temperatures. We assume that true distributions of *icebird* and *sunbird* are Gibbs distributions with parameters $\boldsymbol{\lambda}_{icebird} = (-5,-2)$, $\boldsymbol{\lambda}_{sunbird} = (-3,1)$.

We have 100 observations of *sunbird* and vary the number of observations of *icebird* between 3 and 10,000. For each number of occurrences, we estimate the distribution of *icebird* using both single-class maxent, and HME with a single group $birds = \{icebird, sunbird\}$. The tuning parameter β_0 is set to 0.5.

Results

In Figure 7.3, we present our results. For each HME run, we report values of the HME parameters of the class *icebird* and the group *birds*. For *temp*, the HME parameters of *icebird* agree with its single-class parameters. This matches the intuition behind the bound of Section 7.6.1: the temperature requirements of *icebird* and *sunbird* are different, so pooled estimates provide no advantage; the best setting of the *birds* parameter is zero and the best setting of the *icebird* parameter matches the single-class case. For *prec*, the situation is rather different. The parameter $\eta_{birds,prec}$

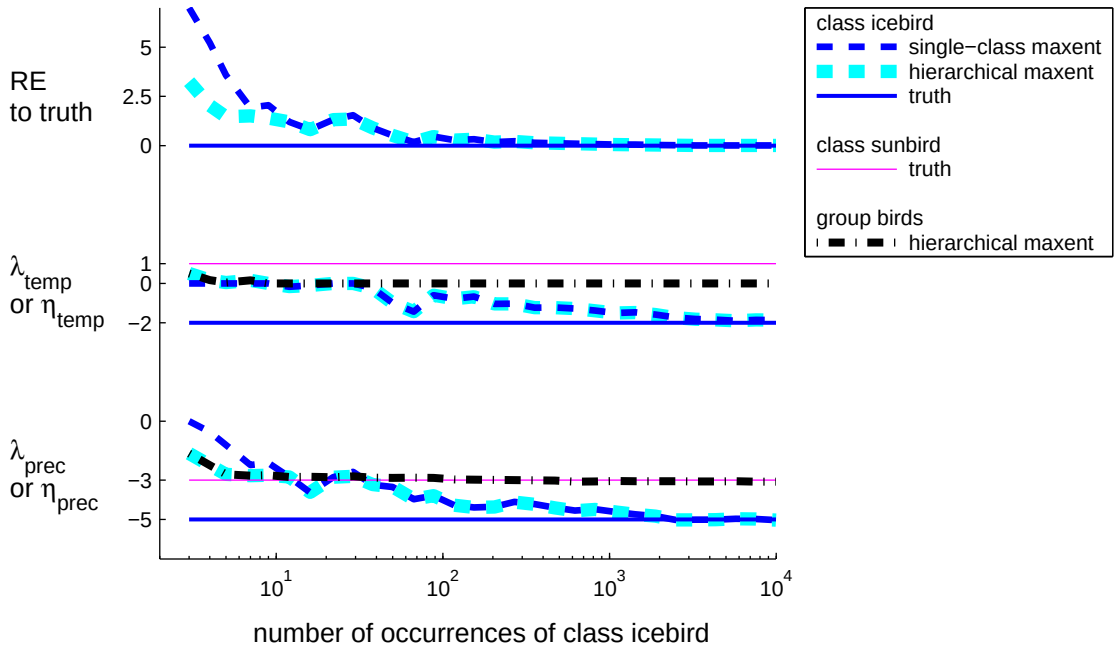


Figure 7.3. *Synthetic experiments.* The precipitation and temperature parameters of classes *icebird* and *sunbird*, and the group *birds*, are fitted by HME as the number of occurrences of *icebird* increases (the number of occurrences of *sunbird* is fixed at 100). Performance of the *icebird* models is reported in terms of relative entropy to the truth. Both the HME models and the single-species models of *icebird* converge to the truth, but HME performs better for small sample sizes, taking advantage of the group estimate of the precipitation parameter.

shows that *birds* prefer low precipitation. This information is used with small sample sizes of *icebird*: $\lambda_{icebird,prec}$ matches $\eta_{birds,prec}$. As the number of samples increases, single-class estimates for *icebird* become more accurate than group estimates, which is reflected in the HME parameters. In the top plot of Fig. 7.3, we see that the HME model performs better than the single-class model. As expected, the improvement is especially dramatic for small sample sizes. For moderate and large sample sizes, the HME estimates match single-class estimates exactly.

7.7.2 Real Data

Experimental Design

Next, we demonstrate the performance of HME on a real-world dataset, specifically on the regions *AWT* and *NSW* from the NCEAS dataset (see Section 5.4). To avoid problems with sample-selection bias, we use only the training (presence-only) portion of the data. We use a randomly chosen half of species in both *AWT* and *NSW* (we withhold the other half for future experiments) with linear and quadratic features derived from continuous environmental variables. We take advantage of the previous

tuning (Section 5.5) and use refined versions of Eqs. (7.51) and (7.52)

$$\beta_{y,j} = \beta'_0 \beta_{\mathrm{LQ}}(m_y) \sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j \mid y]/m_y}$$
$$\beta_{g,j} = \beta'_0 \beta_{\mathrm{LQ}}(m_g) \sqrt{\mathbf{V}'_{\tilde{\pi}}[f_j \mid g]/m_g} \ ,$$

where $\beta_{\mathrm{LQ}}(m_y)$ and $\beta_{\mathrm{LQ}}(m_g)$ are tuned regularization parameters (see the second line of Table 5.3) and β'_0 is a single tuning "hypermultiplier." Note that the region *NSW* contains one categorical variable (see Table 5.1). To simplify the evaluation, this variable is omitted from the experiments.

We evaluate the performance of HME in terms of log likelihood (negative log loss) and AUC using five-fold cross-validation. The complete hierarchies, with the average number of training occurrences across all folds, are given in Figs. 7.1 and 7.2.

We run HME with three types of hierarchy for *AWT* and four types of hierarchy for *NSW*. In both regions we consider empty hierarchies, hierarchies of depth one, with the single group *all species*, and hierarchies of depth two. In *AWT*, the hierarchy of depth two is the complete hierarchy, in *NSW*, the hierarchy of depth two includes the groups *all species*, *birds*, *bats*, *small reptiles*, and *plants*. In *NSW*, we also consider a hierarchy of depth four (the complete hierarchy). Note that this hierarchy contains overlapping groups, so it cannot be expressed as a tree; however, this is not a problem for our setup. The hierarchies are referred to as *h0*, *h1*, *h2*, and *h4*, according to their depth.

Results

In Figure 7.4, we report results for a range of smoothing parameters β'_0. In each region, we show the average across all species. In *AWT*, performance of HME improves, both in terms of log likelihood and AUC, as the hierarchy gets more specific. The improvement is observed across the majority of species and values of β'_0.[3]

In *NSW*, the plots indicate that on average *h1* differs very little from the empty hierarchy *h0*, whereas hierarchies *h2* and *h4* perform better, with *h2* being the best according to the AUC plot. However, on the species level, all three non-empty hierarchies lead to improvements.[4] We do not analyze the choice of the smoothing

[3]Specifically, log likelihood is improved by *h1* over *h0* on 19 out of 20 species, and by *h2* over *h1* on 15 out of 20 species, across all values of β'_0. Results for AUC are similar. The performance of *h1* improves over *h0* on 19 out of 20 species, and *h2* improves over *h1* on 14 out of 20 species.

[4]Specifically, *h1*, *h2*, and *h4* improve the log loss compared with *h0* on 18, 19, and 17 out of 27 species, respectively, across all values β'_0. The improvements in AUC are observed for 17, 16, and 17 out of 27 species, respectively. If we only consider $\beta'_0 \in [0.4, 0.6]$ then log loss improves for 19, 20, and 19 out of 27 species, respectively, and an improvement in AUC on 20, 21, and 19 out of 27 species, respectively.

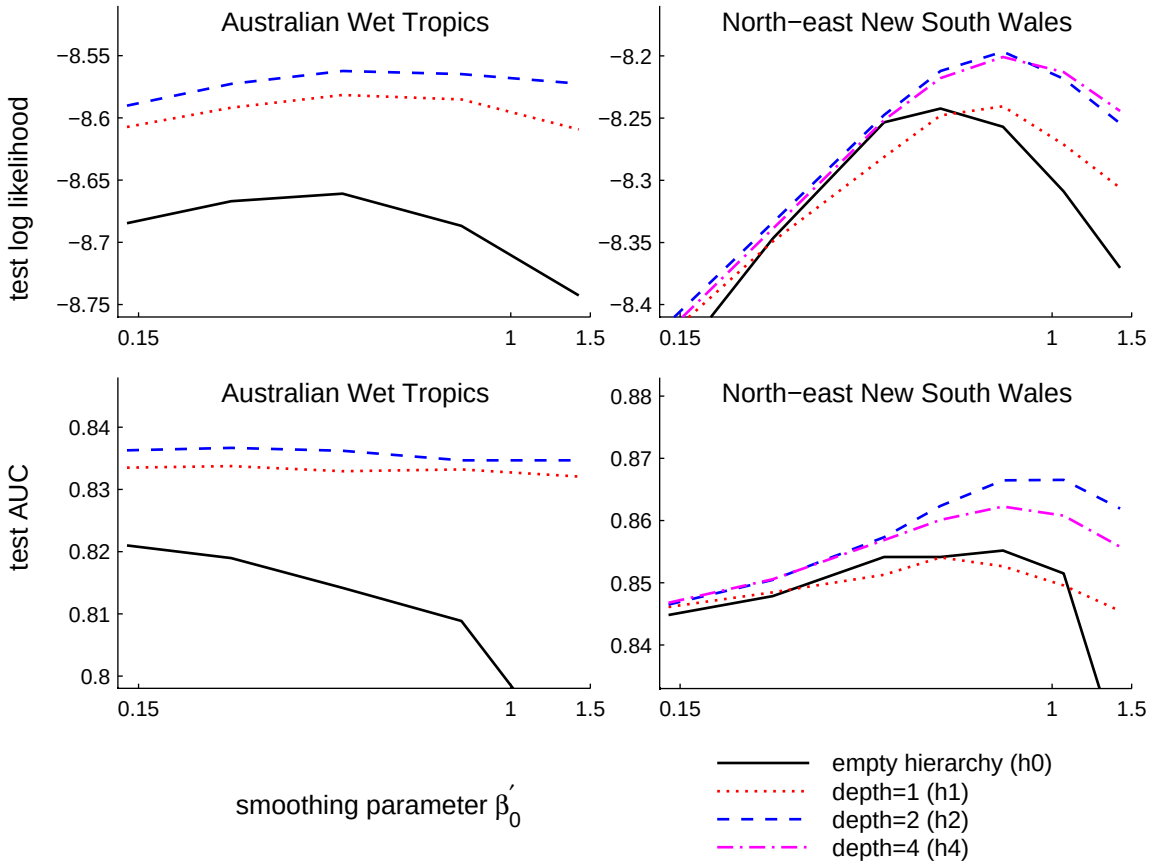


Figure 7.4. *Performance of hierarchies with different depth over a range of smoothing parameters* β_0'. In *AWT*, the hierarchies *h1* and *h2* perform consistently better and are more robust to changes in β_0' than the empty hierarchy *h0*; no hierarchy beyond depth two is available. In *NSW*, *h1*, *h2*, and *h4* perform better than *h0*. The average performance of *h1* and *h0* appears similar, but *h1* improves the log likelihood of 18 out of 27 species, a significant departure from random improvements.

parameter β_0'. We assume that in a concrete application, β_0' is set to a fixed value or determined by model selection.

The main benefit of HME should be observed on species with small numbers of samples. In Fig. 7.5, we show how the improvement due to the use of the group information varies across sample sizes. In *AWT*, we report results using *h2* with $\beta_0' = 0.4$; in *NSW*, we report results using *h2* with $\beta_0' = 0.6$. In *AWT*, the improvement is extremely consistent, and it appears to agree with the difference in relative entropy that we observed in synthetic experiments (Fig. 7.3). In *NSW*, we see the same trend on the vast majority of species. However, the performance seems significantly worse in one case. It is the species with the smallest number of training occurrences—four. This poor result may be as much due to the limits of our method as due to the variance in the evaluation results, since in the case of this species we are performing

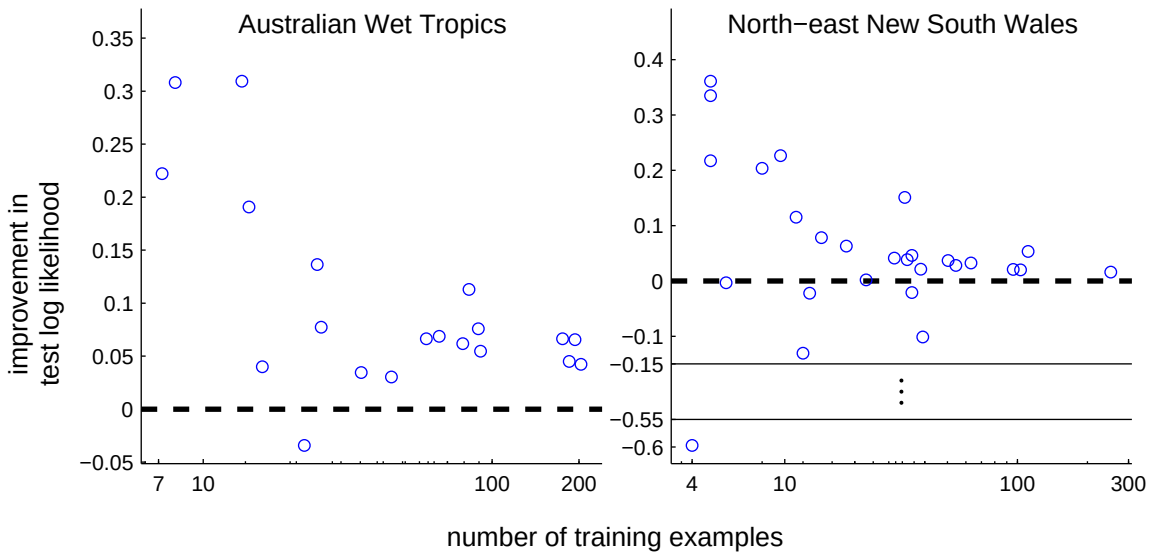


Figure 7.5. *Improvement in performance using HME.* We report the difference in test log likelihood between HME and single-class maxent for every species. The depth of the hierarchy is two. The improvement is the most dramatic for small sample sizes. The performance is significantly worse only in one case: a NE New South Wales species with only four training occurrence records.

five-fold cross-validation with only five samples. Among seven species with ten or fewer training records, this is the only species whose test performance significantly drops.

Chapter 8

Conclusion

In this dissertation, we have provided a unified and complete account of maxent with generalized regularization. We have proved general performance guarantees and proposed versions of iterative scaling that incorporate regularization. We have applied this unified analysis to problems of small-sample estimation, biased estimation, and multiple estimation, and extensively evaluated maxent in an application to species-distribution modeling.

In Chapter 3, we have carried out analysis of several regularization types and presented scenarios in which these regularizations may be useful. Theoretical analysis in Chapters 6 and 7 (biased estimation and multiple estimation) focused on methods derived from ℓ_1-regularized maxent, but it should be straightforward to generalize to convex regularizations using techniques of Chapter 3. For instance, considering HME with ℓ_2^2 regularization would yield analysis of maxent with a hierarchical Gaussian prior. Techniques presented in this dissertation apply to arbitrary log-concave priors, including many widely used ones, such as those in the exponential family. The maximum entropy interpretation enhances our understanding of their generalization properties.

In our experiments (Chapters 5–7), we saw that ℓ_1 regularization facilitated learning in many-dimensional spaces, and its principled extensions to biased estimation and multiple estimation lead to additional improvements. Further empirical study is needed to verify whether the theory derived for other regularization types corresponds to their performance. Note that the quality of regularization can be assessed from two different perspectives: performance over test data and running time. The tradeoff between statistical guarantees and computational efficiency is an interesting question open for future research. In particular, convergence rates of algorithms presented in this dissertation are not known.

We have explored one direction of generalizing maxent: replacing equality constraints by an arbitrary convex potential in the primal or, equivalently, adding a

convex regularization term to the maximum likelihood estimation in the dual. An alternative line of generalizations arises by replacing relative entropy in the primal objective by an arbitrary Bregman or Csiszár divergence along the lines of Altun and Smola (2006), and Collins, Schapire, and Singer (2002). Modified duality results and modified algorithms apply in the new setting, but performance guarantees do not directly translate to the case when divergences are derived from samples. Divergences of this kind are used in many cases of interest such as logistic regression (a conditional version of maxent) and boosting. Generalizing the presented approach to these settings would increase our understanding of regularization and could potentially lead to new algorithms for classification and regression.

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

Appendix A

Empirical Error Inequalities

In this appendix, we list the empirical error inequalities used throughout this dissertation. All of the results are adapted from Devroye et al. (1996).

Theorem A.1 (Hoeffding's inequality, Theorem 8.1 of Devroye et al., 1996; first in Hoeffding, 1963)**.** *Let $X_1,\ldots,X_m$ be independent random variables such that $X_i \in [0,1]$ with probability one. Denote their average by $\tilde{X}_m = \left(\sum_{i=1}^m X_i\right)/m$. Then, for any $\varepsilon > 0$,*

$$\mathbf{P}\left(\tilde{X}_m - \mathbf{E}[\tilde{X}_m] \geq \varepsilon\right) \leq e^{-2\varepsilon^2 m} \qquad \textit{and} \qquad \mathbf{P}\left(\tilde{X}_m - \mathbf{E}[\tilde{X}_m] \leq -\varepsilon\right) \leq e^{-2\varepsilon^2 m} \ .$$

Theorem A.2 (Bernstein's inequality, Theorem 8.2 of Devroye et al., 1996; first in Bernstein, 1946)**.** *Let $X_1,\ldots,X_m$ be independent real-valued random variables with zero mean such that $X_i \leq 1$ with probability one. Denote their average by $\tilde{X}_m = \left(\sum_{i=1}^m X_i\right)/m$, and the average variance by $\sigma^2 = \left(\sum_{i=1}^m \mathbf{V}[X_i]\right)/m$. Then, for any $\varepsilon > 0$,*

$$\mathbf{P}\left(\tilde{X}_m > \varepsilon\right) \leq \exp\left(-\frac{m\varepsilon^2}{2\sigma^2 + 2\varepsilon/3}\right) \ .$$

Theorem A.3 (McDiarmid's inequality, Theorem 9.2 of Devroye et al., 1996; first in McDiarmid, 1989)**.** *Let $X_1,\ldots,X_m$ be independent random variables taking values in a set A and assume that $s : A^m \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_m,x_i' \in A} \left|s(x_1,\ldots,x_m) - s(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_m)\right| \leq c_i \ , \ 1 \leq i \leq m \ .$$

Then, for any $\varepsilon > 0$,

$$\mathbf{P}\left\{s(X_1,\ldots,X_m) - \mathbf{E}[s(X_1,\ldots,X_m)] \geq \varepsilon\right\} \leq e^{-2\varepsilon^2 \Big/ \sum_{i=1}^m c_i^2}$$

$$\textit{and} \qquad \mathbf{P}\left\{\mathbf{E}[s(X_1,\ldots,X_m)] - s(X_1,\ldots,X_m) \geq \varepsilon\right\} \leq e^{-2\varepsilon^2 \Big/ \sum_{i=1}^m c_i^2} \ .$$

The next two theorems bound deviations of empirical frequencies from true probabilities; specifically, deviations between the following two measures derived from independent random variables $X_1, \ldots, X_m$:

$$\tilde{\pi}(A) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(X_i \in A) \tag{A.1}$$

$$\pi(A) = \mathbf{E}[\tilde{\pi}(A)] = \frac{1}{m} \sum_{i=1}^{m} \mathbf{P}(X_i \in A) \ . \tag{A.2}$$

Measure $\tilde{\pi}$ corresponds to the average empirical distribution, whereas π corresponds to the average distribution. The first measure is a random quantity, the second is its expectation.

Theorem A.4. *Let $X_1, \ldots, X_m$ be independent random variables and $\mathcal{A}$ a class of sets. Then, for any $\varepsilon > 0$,*

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} |\tilde{\pi}(A) - \pi(A)| > \varepsilon \right) \leq 8 s(\mathcal{A}, m) e^{-m\varepsilon^2/32}$$

where $\tilde{\pi}$ and π are defined in Eqs. (A.1) *and* (A.2)*, and s is the growth function (see Section 3.4.1).*

Theorem A.4 is a version of Theorem 12.5 of Devroye et al. (1996) (first in Vapnik and Chervonenkis, 1971), restated for independent, but not necessarily identically distributed random variables. The original theorem assumes that the random variables are identically distributed and independent, but its proof remains valid when the identical-distribution requirement is omitted.

Theorem A.5 (Theorem 12.8 of Devroye et al., 1996; first in Devroye, 1982)**.** *Let $X_1, \ldots, X_m$ be independent identically distributed random variables and $\mathcal{A}$ a class of sets. Then, for any $\varepsilon > 0$,*

$$\mathbf{P}\left(\sup_{A \in \mathcal{A}} |\tilde{\pi}(A) - \pi(A)| > \varepsilon \right) \leq 4 e^{8} s(\mathcal{A}, m^2) e^{-2m\varepsilon^2}$$

where $\tilde{\pi}$ and π are defined in Eqs. (A.1) *and* (A.2)*, and s is the growth function (see Section 3.4.1).*

Theorem A.5 improves the multiplicative constant in the exponent of Theorem A.4 at the cost of increasing the coefficient in front of the exponential and imposing an additional requirement that the random variables $X_1, \ldots, X_m$ be identically distributed.

Bibliography

Altun, Y. and A. Smola (2006). Unifying divergence minimization and statistical inference via convex duality. In *COLT 2006: Proceedings of the 19th Annual Conference on Learning Theory*.

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

Azoury, K. S. and M. K. Warmuth (2001). Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning 43*, 211–246.

Baxter, J. (2000). A model of inductive bias learning. *J. Artif. Intell. Res. 12*, 149–198.

Beneš, V. E. (1965). *Mathematical Theory of Connecting Networks and Telephone Traffic*. New York: Academic Press.

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

Bernstein, S. N. (1946). *Theory of Probability* (fourth ed.). Moscow-Leningrad: Gostekhizdad.

Boltzmann, L. (1871a). Analytischer Beweis des zweiten Haubtsatzes der mechanischen Wärmetheorie aus den Sätzen über das Gleichgewicht der lebendigen Kraft. *Wiener Berichte 63*, 712–732. In Boltzmann (1909), Volume 1, Paper 20. Cited in Uffink (2004).

Boltzmann, L. (1871b). Einige allgemeine Sätze über wärmegleichgewicht. *Wiener Berichte 63*, 679–711. In Boltzmann (1909), Volume 1, Paper 19. Cited in Uffink (2004).

Boltzmann, L. (1871c). Über das Wärmegleichgewicht zwischen mehratomigen Gasmolekülen. *Wiener Berichte 63*, 397–418. In Boltzmann (1909), Volume 1, Paper 18. Cited in Uffink (2004).

Boltzmann, L. (1909). *Wissenschaftliche Abhandlungen*. 3 vols. Edited by F. Hasenöhrl. Leipzig: Barth. Reprint, New York: Chelsea, 1969. Cited in Uffink (2004).

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

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *Zh. Vychisl. Mat. Mat. Fiz. 7*(3), 620–631. English translation in *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Caruana, R. (1993). Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning: Proceedings of the Tenth International Conference*, pp. 41–48. Morgan Kaufmann.

Censor, Y. and A. Lent (1981). An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications 34*(3), 321–353.

Censor, Y. and S. A. Zenios (1997). *Parallel Optimization: Theory, Algorithms, and Applications*. New York: Oxford University Press.

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

Chen, S. F. and R. Rosenfeld (2000, January). A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing 8*(1), 37–50.

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

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. Wiley.

Csiszár, I. (1975). I-Divergence geometry of probability distributions and minimization problems. *The Annals of Probability 3*(1), 146–158.

Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability 12*(3), 768–793.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics 19*(4), 2032–2066.

Csiszár, I. (1995). Maxent, mathematics, and information theory. In *Proceedings of the Fifteenth Internationl Workshop on Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer Academic Publishers.

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

Dekel, O., S. Shalev-Shwartz, and Y. Singer (2003). Smooth ϵ-insensitive regression by loss symmetrization. In *COLT/Kernel 2003: Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*, pp. 433–447. Springer.

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

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

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

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

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

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

Dudík, M., D. M. Blei, and R. E. Schapire (2007). Hierarchical maximum entropy density estimation. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 249–256. New York: ACM Press. Also available at http://doi.acm.org/10.1145/1273496.1273528.

Dudík, M., S. J. Phillips, and R. E. Schapire (2004). Performance guarantees for regularized maximum entropy density estimation. In *COLT 2004: Proceedings of the 17th Annual Conference on Learning Theory*, pp. 472–486. Berlin: Springer-Verlag.

Dudík, M., S. J. Phillips, and R. E. Schapire (2007). Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research 8*, 1217–1260. Also available at http://jmlr.csail.mit.edu/papers/volume8/dudik07a/dudik07a.pdf.

Dudík, M. and R. E. Schapire (2006). Maximum entropy distribution estimation with generalized regularization. In *COLT 2006: Proceedings of the 19th Annual Conference on Learning Theory*, pp. 123–138. Berlin: Springer-Verlag.

Dudík, M., R. E. Schapire, and S. J. Phillips (2006). Correcting sample selection bias in maximum entropy density estimation. In *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, pp. 323–330. Cambridge, MA: MIT Press.

Elith, J. (2002). Quantitative methods for modeling species habitat: Comparative performance and an application to Australian plants. In S. Ferson and M. Burgman (Eds.), *Quantitative Methods for Conservation Biology*, pp. 39–58. New York: Springer-Verlag.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography 29*(2), 129–151.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978.

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

Ferrier, S., G. Watson, J. Pearce, and M. Drielsma (2002). Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. I. Sommunity-level modelling. *Biodiversity and Conservation 11*, 2275–2307.

Fréchet, M. (1910). Extension au cas d'intégrales multiples d'une définition de l'intégrale due à Stieltjes. *Nouv. Ann. Math. (sér. 4) 10*, 241–256. Cited in *Encyclopaedia of Mathematics* (Hazewinkel, 1987), s.v. "Vitali variation".

Freund, Y. and L. Mason (1999). The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 124–133. Morgan Kaufmann.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics 19*(1), 1–141.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics, Developed with Especial Reference to the Rational Foundations of Thermodynamics*. New York: C. Scribner's sons.

Golea, M., W. S. L. Peter L. Bartlett, and L. Mason (1998). Generalization in decision trees and DNF: Does size matter? In *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, Cambridge, MA, pp. 259–265. MIT Press.

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

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

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

Groves, R. M. (1989). *Survey Errors and Survey Costs*. Wiley.

Grünwald, P. (2001). Strong entropy concentration, game theory, and algorithmic randomness. In *COLT/EuroCOLT 2001: Proceedings of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pp. 320–336. London: Springer-Verlag.

Hájek, A. (2007, Fall). Interpretations of probability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. First published in 2002. Available at http://plato.stanford.edu/archives/fall2007/entries/probability-interpret/. Accessed at http://plato.stanford.edu/entries/probability-interpret/, July 26, 2007.

Hanley, J. A. and B. S. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*, 29–36.

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

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Science+Business Media.

Hazewinkel, M. (Ed.) (1987). *Encyclopaedia of Mathematics*. 10 vols. Dordrecht: Kluwer Academic Publishers. An updated and annotated translation of the Soviet "Mathematical Encyclopaedia" (ed. M. Vinogradov, 5 vols., Soviet Encyclopaedia Publishing House, 1977–1985). Also available at http://eom.springer.de/.

Heckman, J. J. (1979, January). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Hoeffding, W. (1963, March). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*(301), 13–30.

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

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

Ivanov, V. K. (1962). On linear problems which are not well-posed. *Soviet Math. Dokl. 3*(4), 981–983. Translated from *Dokl. Akad. Nauk SSSR*, 145(2): 270–272, 1962. Cited in Vapnik (1999), pp. 9 and 236; and *Encyclopaedia of Mathematics* (Hazewinkel, 1987), s.v. "ill-posed problems".

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Stat. Prob.*, Volume 1, pp. 311–319.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review 106*(4), 620–630. Reprinted in Jaynes (1983), pp. 6–16.

Jaynes, E. T. (1978). Where do we stand on maximum entropy? In R. D. Levine and M. Tribus (Eds.), *The Maximum Entropy Formalism*, pp. 15–118. Cambridge, MA: MIT Press. Reprinted in Jaynes (1983), pp. 210–314.

Jaynes, E. T. (1983). *Papers on Probability, Statistics, and Statistical Physics*. Edited by R. D. Rosenkrantz. Dordrecht, Holland: D. Reidel Publishing Company.

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

Kapur, J. N. and H. K. Kesavan (1992). *Entropy Optimization Principles with Applications*. Academic Press.

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

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

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

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

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

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

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

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

Lebesgue, H. (1910). Sur l'intégration des fonctions discontinues. *Ann. Sci. École Norm. Sup. 27*(3), 361–450. Cited in *Encyclopaedia of Mathematics* (Hazewinkel, 1987), s.v. "Vitali variation".

Lee, S.-I., H. Lee, P. Abbeel, and A. Y. Ng (2006, July). Efficient L_1 regularized logistic regression. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA, pp. 1–9.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (second ed.). New York: Springer-Verlag.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). Wiley.

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

McCallum, A., R. Rosenfeld, T. M. Mitchell, and A. Y. Ng (1998). Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 359–367. Morgan Kaufmann.

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

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

New, M., M. Hulme, and P. Jones (1999). Representing twentieth-century space-time climate variability. Part 1: Development of a 1961-90 mean monthly terrestrial climatology. *Journal of Climate 12*, 829–856.

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

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pp. 615–622. New York: ACM Press. Also available at http://doi.acm.org/10.1145/1015330.1015435.

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

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

Phillips, D. L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM 9*(1), 84–97. Also available at http://doi.acm.org/10.1145/321105.321114.

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

Phillips, S. J. and M. Dudík (2007). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Preprint, submitted for peer review.

Phillips, S. J., M. Dudík, J. Elith, C. Graham, A. Lehman, J. Leathwick, and S. Ferrier (2007). Sample selection bias and presence-only models of species distributions. Preprint, submitted for peer review.

Phillips, S. J., M. Dudík, and R. E. Schapire (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pp. 655–662. New York: ACM Press. Also available at http://doi.acm.org/10.1145/1015330.1015412.

Phillips, S. J., M. Dudík, and R. E. Schapire (2007). Maxent software for species habitat modeling. http://www.cs.princeton.edu/~schapire/maxent.

Raina, R., A. Y. Ng, and D. Koller (2006). Constructing informative priors using transfer learning. In *Proc. of the Twenty-Third International Conference on Machine Learning*, pp. 713–720.

Raxworthy, C. J., E. Martinez-Meyer, N. Horning, R. A. Nussbaum, G. E. Schneider, M. A. Ortega-Huerta, and A. T. Peterson (2003). Predicting distributions of known and unknown reptile species in Madagascar. *Nature 426*, 837–841.

Reddy, S. and L. M. Dávalos (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography 30*, 1719–1727.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton University Press.

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

Rosset, S. and E. Segal (2003). Boosting density estimation. In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, pp. 641–648. Cambridge, MA: MIT Press.

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

Sanov, I. N. (1957). On the probability of large deviations of random variables. *Mat. Sbornik 42*, 11–44.

Sauer, J. R., J. E. Hines, and J. Fallon (2001). The North American breeding bird survey, results and analysis 1966–2000, Version 2001.2. http://www.mbr-pwrc.usgs.gov/bbs/bbs.html. USGS Patuxent Wildlife Research Center, Laurel, MD.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory Series A 13*, 145–147.

Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423, 623–656.

Shore, J. E. and R. W. Johnson (1980, January). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory IT-26*(1), 26–37.

Skilling, J. (1988). The axioms of maximum entropy. In G. J. Erickson and C. R. Smith (Eds.), *Maximum-Entropy and Bayesian Methods in Science and Engineering*, Volume 1, pp. 173–187. Kluwer Academic Publishers.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Stat. Prob.*, Volume 1, pp. 197–206.

Strong, D. and T. Chan (2003). Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems 19*, S165–ŰS187. Also available at http://stacks.iop.org/IP/19/S165.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge, MA. MIT Press.

Tibshirani, R. (1996). Regression shrikage and selection via the lasso. *Journal of the Royal Statistical Society (Series B) 58*(1), 267–288.

Tikhonov, A. N. (1963a). Regularization of incorrectly posed problems. *Soviet Math. Dokl. 4*, 1624–1627. Translated from *Dokl. Akad. Nauk SSSR*, 153(1), 49–52, 1963. Cited in *Encyclopaedia of Mathematics* (Hazewinkel, 1987), s.v. “ill-posed problems”.

Tikhonov, A. N. (1963b). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl. 4*, 1035–1038. Translated from *Dokl. Akad. Nauk SSSR*, 151(3), 501–504, 1963. Cited in Vapnik (1999), pp. 9 and 235; and *Encyclopaedia of Mathematics* (Hazewinkel, 1987), s.v. "ill-posed problems".

Topsøe, F. (1979). Information theoretical optimization techniques. *Kybernetika 15*(1), 8–27.

Uffink, J. (2004, Winter). Boltzmann's work in statistical physics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/win2004/entries/statphys-Boltzmann/.

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

Van Campenhout, J. M. and T. M. Cover (1981). Maximum entropy and conditional probability. *IEEE Transactions on Information Theory IT-27*, 483–489.

van de Geer, S. A. (2006, June). High-dimensional generalized linear models and the lasso. Technical Report 133, ETH Seminar für Statistik.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory* (second ed.). New York: Springer-Verlag.

Vapnik, V. N. and A. Ya. Chervonenkis (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Dokl. Akad. Nauk SSSR 181*(4). In Russian.

Vapnik, V. N. and A. Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications 16*(2), 264–280.

Vapnik, V. N. and A. Ya. Chervonenkis (1974). *Theory of Pattern Recognition*. Moscow: Nauka. In Russian.

Vitali, G. (1908). Sui gruppi di punti e sulle funzioni di variabili reali. *Atti Accad. Sci. Torino 43*, 75–92. Cited in *Encyclopaedia of Mathematics* (Hazewinkel, 1987), s.v. "Vitali variation".

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

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

Welling, M., R. S. Zemel, and G. E. Hinton (2003). Self supervised boosting. In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, pp. 665–672. Cambridge, MA: MIT Press.

Wiley, E. O., K. M. McNyset, A. T. Peterson, C. R. Robins, and A. M. Stewart (2003). Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography 16*(3), 120–127.

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

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (Series B) 68*(1), 49–67.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, pp. 903–910. New York: ACM Press. Also available at http://doi.acm.org/10.1145/1015330.1015425.

Zadrozny, B., J. Langford, and N. Abe (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 435–442.

Zaniewski, A. E., A. Lehmann, and J. M. Overton (2002). Predicting species spatial distributions using presence-only data: A case study of native New Zealand ferns. *Ecological Modelling 157*, 261–280.

Zhang, T. (2005). Class-size independent generalization analysis of some discriminative multi-category classification. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge, MA, pp. 1625–1632. MIT Press.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society (Series B) 67*, 301–320.