

Article

## Consistency and Generalization Bounds for Maximum Entropy Density Estimation

Shaojun Wang<sup>1,\*</sup>, Russell Greiner<sup>2</sup> and Shaomin Wang<sup>3</sup>

<sup>1</sup> Kno.e.sis Center, Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435, USA

<sup>2</sup> Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada; E-Mail: rgreiner@ualberta.ca

<sup>3</sup> Visa Inc., San Francisco, CA 94128, USA; E-Mail: shaomin.wang@gmail.com

\* Author to whom correspondence should be addressed; E-Mail: shaojun.wang@wright.edu; Tel.: +1-937-775-5140; Fax: +1-937-775-5133.

Received: 9 July 2013; in revised form: 13 November 2013 / Accepted: 3 December 2013 /

Published: 9 December 2013

---

**Abstract:** We investigate the statistical properties of maximum entropy density estimation, both for the complete data case and the incomplete data case. We show that under certain assumptions, the generalization error can be bounded in terms of the complexity of the underlying feature functions. This allows us to establish the universal consistency of maximum entropy density estimation.

**Keywords:** maximum entropy principle; density estimation; generalization bound; consistency

---

### 1. Introduction

The maximum entropy (ME) principle, originally proposed by Jaynes in 1957 [1], is an effective method for combining different sources of evidence from complex, yet structured, natural systems. It has since been widely applied in science, engineering and economics. In machine learning, ME was first popularized by Della Pietra *et al.* [2], who applied it to induce overlapping features for a Markov random field model of natural language. Later, it was applied to other machine learning areas, such as information fusion [3] and reinforcement learning [4]. It is now well known that for complete data, the ME principle is equivalent to maximum likelihood estimation (MLE) in a Markov random

field. In fact, these two problems are exact duals of one another. Recently, Wang *et al.* [5] proposed the latent maximum entropy (LME) principle to extend Jaynes' maximum entropy principle to deal with hidden variables, and demonstrated its effectiveness in many statistical models, such as mixture models, Boltzmann machines and language models [6]. We show that LME is different from both Jaynes' maximum entropy principle and maximum likelihood estimation, but it often yields better estimates in the presence of hidden variables and limited training data. This paper investigates the statistical properties of maximum entropy density estimation for both the complete and incomplete data cases.

Large sample asymptotic convergence results for MLE are typically based on point estimation analysis [7] in parametric models. Although point estimators have been extensively studied in the statistics literature since Fisher, these analyses typically do not consider generalization ability. Vapnik and Chervonenkis famously reformulated the problem of MLE for density estimation in the framework of empirical risk minimization and provided the first necessary and sufficient conditions for consistency [8,9]. However, the model they considered is still in a Fisher–Wald parametric setting. Barron and Sheu [10] considered a density estimation problem very similar to the one we address here, but only restricted to the one-dimensional case within a bounded interval. Their analysis cannot be easily generalized to a high dimensional case. Recently, Dudik *et al.* [11] analyzed regularized maximum entropy density estimation with inequality constraints and derived generalization bounds for this model. However, once again, their analysis does not easily extend beyond the specific model considered.

Some researchers have studied the consistency of maximum likelihood estimators under the Hellinger divergence [12], which is a particularly convenient measure for studying maximum likelihood estimation in a general distribution-free setting. However, Kullback–Leibler divergence is a more natural measure for probability distributions and is closely related to the perplexity measure used in language modeling and speech recognition research [13,14]. Moreover, convergence in the Kullback–Leibler divergence always establishes consistency in terms of Hellinger divergence [12], but not *vice versa*. Therefore, we concentrate on using the Kullback–Leibler divergence in our analysis.

In this paper, we investigate consistency and generalization bounds for maximum entropy density estimation with respect to the Kullback–Leibler divergence. The main technique we use in our analysis is Rademacher complexity, first used by Koltchinskii and Panchenko [15] to analyze the generalization error of combined classification methods, such as boosting, support vector machines and neural networks. Since then, the convenience of Rademacher analysis has been exploited by many to analyze various learning problems in classification and regression. For example, Rakhlin *et al.* [16] have used this technique to derive risk bounds for the density estimation of mixture models, which basically belong to directed graphical models using a conditional parameterization. Here, we use the Rademacher technique to analyze the generalization error of maximum entropy density estimation for general Markov random fields.

## 2. Maximum Entropy Density Estimation: Complete Data

Let  $X \in \mathcal{X}$  be a random variable. Given a set of feature functions  $\mathcal{F}(x) = \{f_1(x), \dots, f_N(x)\}$  specifying properties one would like to match in the data, the maximum entropy principle states that we

should select a probability model,  $p(x)$ , from the space of all probability distributions,  $\mathcal{P}(x)$ , over  $\mathcal{X}$ , to maximize entropy subject to the constraint that the feature expectations are preserved:

$$\max_{p(x) \in \mathcal{P}(x)} \left[ - \int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) \right] \tag{1}$$

$$\text{s.t. } \int_{x \in \mathcal{X}} f_i(x) p(x) \mu(dx) = \int_{x \in \mathcal{X}} f_i(x) p_0(x) \mu(dx); \quad i = 1 \dots N, \tag{2}$$

where  $p_0(x)$  denotes the unknown underlying true density and  $\mu$  denotes a given  $\sigma$ -finite measure on  $\mathcal{X}$ . If  $\mathcal{X}$  is finite or countably infinite, then  $\mu$  is the counting measure, and integrals reduce to sums. If  $\mathcal{X}$  is a subset of a finite dimensional space,  $\mu$  is the Lebesgue measure. If  $\mathcal{X}$  is a combination of both cases,  $\mu$  will be a combination of both measures. The dual problem is:

$$\min_{\lambda \in \Omega} \left[ - \int_{x \in \mathcal{X}} p_0(x) \log p_\lambda(x) \mu(dx) \right] \tag{3}$$

where  $p_\lambda(x) = \Phi_\lambda^{-1} \exp \left( \sum_{i=1}^N \lambda_i f_i(x) \right)$  and  $\Phi_\lambda = \int_{x \in \mathcal{X}} \exp \left( \sum_{i=1}^N \lambda_i f_i(x) \right) \mu(dx)$  is a normalizing constant that ensures  $\int_{x \in \mathcal{X}} p_\lambda(x) \mu(dx) = 1$ .

We will use the following notation and terminology throughout the analysis below. Define:

$$\Omega = \left\{ \lambda \in \mathbb{R}^N : \int_{x \in \mathcal{X}} \exp \left( \sum_{i=1}^N \lambda_i f_i(x) \right) \mu(dx) < \infty \right\}$$

and let:

$$\mathcal{E}(x) = \left\{ p_\lambda(x) \in \mathcal{P}(x) : p_\lambda(x) = \frac{1}{\Phi_\lambda} \exp \left( \sum_{i=1}^N \lambda_i f_i(x) \right), \quad \lambda \in \Omega \right\}$$

denote the exponential family induced by the set of feature functions. The restriction,  $\lambda \in \Omega$ , will guarantee that the maximum likelihood estimate is an interior point of the set of  $\lambda$ 's for which  $p_\lambda(x)$  is defined. The optimal solution,  $p_{\hat{\lambda}}(x)$ , of Equation (1) or Equation (3) is called the information projection [10,17] of  $p_0(x)$  to the exponential family,  $\mathcal{E}(x)$ .

In practice, the true distribution,  $p_0(x)$ , is not known, but instead, a collection of data  $\tilde{\mathcal{X}} = (x_1, \dots, x_M)$  sampled from  $p_0(x)$  is given. Therefore, instead of using the true distribution,  $p_0(x)$ , we use the empirical distribution,  $\tilde{p}(x)$ , to calculate the feature expectations. The ME principle then becomes:

$$\max_{p(x) \in \mathcal{P}} \left[ - \int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) \right] \tag{4}$$

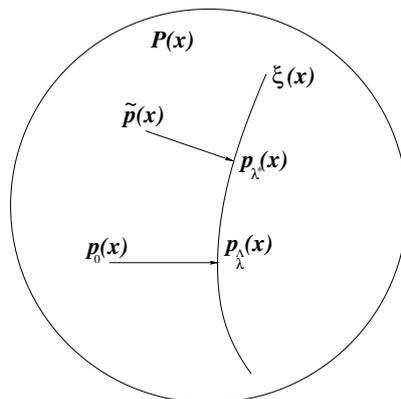
$$\text{s.t. } \int_{x \in \mathcal{X}} f_i(x) p(x) \mu(dx) = \sum_{x \in \tilde{\mathcal{X}}} \tilde{p}(x) f_i(x); \quad i = 1 \dots N, \tag{5}$$

and the dual problem using these constraints becomes:

$$\min_{\lambda \in \Omega} \left[ - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right] \tag{6}$$

The optimal solution,  $p_{\lambda^*}(x)$ , of Equation (4) or Equation (6) is the information projection of  $\tilde{p}(x)$  to the exponential family,  $\mathcal{E}(x)$ ; see Figure 1.

**Figure 1.** In the space of all probability distributions,  $\mathcal{P}(x)$  over  $\mathcal{X}$ ,  $p_{\hat{\lambda}}(x)$  is the information projection of the true (but unknown) distribution,  $p_0(x)$ , to the exponential family,  $\mathcal{E}$ , and  $p_{\lambda^*}(x)$  is the information projection of the empirical distribution,  $\tilde{p}(x)$ , to  $\mathcal{E}$ .



### 2.1. Consistency and Generalization Bounds of Estimation Error

There are two standard ways to measure the quality of the estimate,  $p_{\lambda^*}(x)$ .

One approach, based on the Kullback–Leibler divergence, was first considered by Barron and Sheu [10]. Basically, it uses the following well-known Pythagorean property (see Lemma 3 in [10]):

$$D(p_0(x) \| p_{\lambda}(x)) = D(p_0(x) \| p_{\hat{\lambda}}(x)) + D(p_{\hat{\lambda}}(x) \| p_{\lambda}(x)) \quad \forall p_{\lambda}(x) \in \mathcal{E} \tag{7}$$

and, in particular, the decomposition:

$$D(p_0(x) \| p_{\lambda^*}(x)) = D(p_0(x) \| p_{\hat{\lambda}}(x)) + D(p_{\hat{\lambda}}(x) \| p_{\lambda^*}(x)) \tag{8}$$

Here, the first term,  $D(p_0(x) \| p_{\hat{\lambda}}(x))$ , is the approximation error introduced by the bias of the set of feature functions,  $\mathcal{F}$ , which measures how closely the feature functions are able to approximate the true probability distribution. The second term,  $D(p_{\hat{\lambda}}(x) \| p_{\lambda^*}(x))$ , is the estimation error for densities in the exponential family, introduced by the variance of using a finite sample size. These two terms resemble the bias-variance tradeoff in least squares cost estimation [18]. The approximation error always exists unless the set of feature functions,  $\mathcal{F}$ , is rich enough [19]. In this paper, we assume that the set of feature functions is given and study how close  $p_{\hat{\lambda}}(x)$  is to  $p_{\lambda^*}(x)$ .

Another way to evaluate the quality of  $p_{\lambda^*}(x)$  is to measure the difference between the best expected likelihood and the empirical likelihood of the estimate, which is a more desirable approach, because we can directly calculate the empirical likelihood of the estimate from the training data.

Both approaches, in fact, fall under the umbrella of Vapnik’s empirical risk minimization principle [8] for density estimation. Here, we use the first approach to show that the maximum entropy solution converges to the best possible solution,  $p_{\hat{\lambda}}(x)$ , and the second approach to show that the value of empirical likelihood converges to the maximum expected likelihood.

2.1.1. Maximum Entropy Principle

Exploiting tools from empirical process theory, including symmetrization, concentration and contraction inequalities, Koltchinskii and Panchenko [15] were able to give bounds on the generalization error of boosting, support vector machines (SVMs) and neural networks. We show through the following theorem that we can use similar tools to establish generalization bounds for the estimation error of maximum entropy density estimation. All proofs can be found in the Appendix.

**Theorem 1.** *Assuming  $\sup_{\lambda \in \Omega} \|\lambda\|_1 < \infty$  and  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f(x)\|_\infty < \infty$ , then there exist  $0 < \zeta < \alpha < \infty$ , such that, for any  $\eta \in (0, 1)$ , with a probability of at least  $1 - \eta$ ,*

$$\begin{aligned}
 D(p_{\hat{\lambda}}(x) \| p_{\lambda^*}(x)) &= D(p_0(x) \| p_{\lambda^*}(x)) - D(p_0(x) \| p_{\hat{\lambda}}(x)) \\
 &\leq \frac{4C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + C_2 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}}
 \end{aligned}
 \tag{9}$$

where  $M$  is the number of instances,  $\mathcal{N}(\mathcal{F}(x), \epsilon, d_x)$  is the random covering number [20,21] for linear combinations of functions in  $\mathcal{F}(x)$  at scale  $\epsilon$  with empirical Euclidean distance  $d_x$  on sample  $\tilde{\mathcal{X}}$ , and  $C_1$  and  $C_2$  are positive constants that do not depend on the instance.

If the linear combination of feature functions belongs to a VC-subgraph with Vapnik-Chervonenkis (VC) dimension  $V$ , it is well known that the Dudley integral is bounded by  $\sqrt{V}$  [12,20,22].

Rakhlin *et al.* [16] first applied a similar technique to derive generalization bounds for the density estimation of mixture models. Here, we apply the technique to derive bounds for maximum entropy density estimation in general Markov random fields. Even though the analysis we use is standard, our results show that the generalization bound is not related to the log partition function, but instead is upper bounded by the covering number of linear combinations of feature functions. This is an important observation, because the log partition function—a fundamental quantity associated with any graph-structured distribution—is NP-hard to compute in general [23].

Although the assumption that  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f(x)\|_\infty < \infty$  seems rather restrictive, most of the graphical models studied in machine learning [23] are, in fact, *discrete* Markov random fields, which always satisfy this condition.

To eliminate the assumption of boundedness on the parameters and feature functions, we use a result adapted from [24].

**Theorem 2.** *Assume that there exists a positive number,  $K(\mathcal{F})$ , such that for all  $\tau > 0$ ,*

$$\log E_{p_0(x)} \left( p_{\lambda^\bullet}(x)^{2\tau} - \frac{1}{p_{\lambda^\bullet}(x)^{2\tau}} \right) \leq \left( \tau K(\mathcal{F}) \right)^2
 \tag{10}$$

where  $\lambda^\bullet$  are parameters, such that  $E_{p_{\lambda^\bullet}(x)}(f(x)) = f(x)$ . Then, for all  $\lambda \in \Omega$ , we have, with a probability of at least  $1 - \eta$ ,

$$\begin{aligned}
 &E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \\
 &\leq E_{\tilde{\mathcal{X}}} \sup_{\lambda \in \Omega, f(x) \in \mathcal{F}(x)} \left( E_{p_0(x)} \langle \lambda, f(x) \rangle - E_{\tilde{p}(x)} \langle \lambda, f(x) \rangle \right) + K(\mathcal{F}) \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}}
 \end{aligned}
 \tag{11}$$

By the above result, Theorem 1 can be proven by replacing  $C_2$  with  $K(\mathcal{F})$ . Since the value,  $K(\mathcal{F})$ , is hard to determine in practice, we state our results in terms of the bound on feature functions instead. Nevertheless, the reader should keep in mind that this bound can be replaced by  $K(\mathcal{F})$  in all our results.

Using the result of [21], the following theorem gives a useful bound on the covering number.

**Theorem 3.** Assume  $\sup_{\lambda \in \Omega} \|\lambda\|_2 < a$  and  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f(x)\|_2 < b$ , then:

$$\log_2 \mathcal{N}(\mathcal{F}(x), \epsilon, d_x) \leq \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \left( \min(\log_2(2M + 1), \log_2(2N + 1)) \right) \tag{12}$$

We then have the following corollaries. The first is a direct consequence of Theorem 1. The second provides a generalization bound on the expected value of the estimation error, which can be shown using the proof in [16,20].

**Corollary 1.** Universal consistency: If  $\int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon$  is bounded, then as  $M \rightarrow \infty$ ,  $p_{\lambda^*}(x)$  will converge to  $p_{\hat{\lambda}}(x)$  in terms of the Kullback–Leibler divergence with rate  $\mathcal{O}(\frac{1}{\sqrt{M}})$ , independent of the form of the true distribution,  $p_0$ .

**Corollary 2.** Under the conditions of Theorem 1, the generalization bound of the expected estimation error is:

$$\begin{aligned} E_{\tilde{\mathcal{X}}} D(p_{\hat{\lambda}}(x) \| p_{\lambda^*}(x)) &= E_{\tilde{\mathcal{X}}} D(p_0(x) \| p_{\lambda^*}(x)) - D(p_0(x) \| p_{\hat{\lambda}}(x)) \\ &\leq \frac{4C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + \frac{C_2^2}{M} \end{aligned} \tag{13}$$

Next, we turn to the second approach to evaluating the quality of  $p_{\lambda^*}(x)$ ; namely, by measuring the difference between the expected log-likelihood and the empirical log-likelihood. This is the approach used by Dudik *et al.* [11].

Define  $\mathcal{L}_0(\lambda) = -\int_{x \in \mathcal{X}} p_0(x) \log p_{\lambda}(x) \mu(dx)$  and  $\tilde{\mathcal{L}}(\lambda) = -\frac{1}{M} \sum_{j=1}^M \log p_{\lambda}(x_j)$ . Then, from Theorem 1 and the McDiarmid concentration inequality, we obtain the following result for the sample log-likelihood, which is similar to Dudik *et al.* [11].

**Theorem 4.** There are  $0 < \zeta < \alpha < \infty$ , such that, with a probability of at least  $1 - \eta$ :

$$\begin{aligned} \left| \mathcal{L}_0(\hat{\lambda}) - \tilde{\mathcal{L}}(\lambda^*) \right| &= \left| E_{p_0(x)} \log p_{\hat{\lambda}}(x) - \frac{1}{M} \sum_{j=1}^M \log p_{\lambda^*}(x_j) \right| \\ &\leq \frac{6C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + 3C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}} \end{aligned} \tag{14}$$

By the Glivenko–Cantelli theorem [9], we know that the empirical distribution converges to the true distribution. Therefore, under certain conditions, the entropy of empirical distribution will also converge to the entropy of the true distribution. Whenever such convergence holds, we can combine this with Theorem 3 to show that  $D(p_0(x) \| p_{\hat{\lambda}}(x)) - D(\tilde{p}(x) \| p_{\lambda^*}(x)) \rightarrow 0$  as  $M \rightarrow \infty$ , establishing a stronger form of consistency than Corollary 1.

2.1.2. Regularized Maximum Entropy Principle

In many statistical modeling settings, the constraints used in the ME principle are subject to errors from the empirical nature of the data. This is particularly true in domains with sparse, high dimensional data. One way to gain robustness, though, is to relax the constraints and add a penalty to the entropy, leading to the *regularized* ME (RME) principle [25]:

$$\max_{p(x), \mathbf{a}} \left[ - \int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) - U(\mathbf{a}) \right] \tag{15}$$

$$\text{s.t. } \int_{x \in \mathcal{X}} p(x) f_i(x) \mu(dx) = \sum_x \tilde{p}(x) f_i(x) + a_i; \quad i = 1, \dots, N \tag{16}$$

Here,  $\mathbf{a} = (a_1, \dots, a_N)^T$ ,  $a_i$  is the error for each constraint, and  $U : \Re^N \rightarrow \Re$  is a cost function that has a value of zero at **zero**. The function penalizes any constraint violations, and can be used to penalize deviations in the more reliably observed constraints to a greater degree than deviations in less reliably observed constraints.

The dual problem of RME becomes:

$$\min_{\lambda \in \Omega} \left[ - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) + U^*(\lambda) \right] \tag{17}$$

which is equivalent to *maximum a posteriori* (MAP) estimation with prior  $U^*(\lambda)$ . Note that the prior,  $U^*(\lambda)$ , is derived from the penalty function,  $U$ , over errors  $\mathbf{a}$ , by setting  $U^*$  to the convex (Fenchel) conjugate [26,27] of  $U$ , i.e.,  $U^*(\lambda) = \sup_{\mathbf{a}} \langle \lambda, \mathbf{a} \rangle - U(\mathbf{a})$ . This function is always convex, regardless of the convexity of  $U$ . Conversely, given the convex conjugate cost function,  $U^*$ , the corresponding penalty function,  $U$ , can be derived by using the property of *Fenchel biconjugation* [26]; that is, the conjugate of the conjugate of a convex function is the original convex function,  $U = U^{**}$ .

Conventionally,  $U(\mathbf{a})$  is chosen to be nonnegative and have a minimum of zero at **zero**. To illustrate, consider a quadratic penalty  $U(\mathbf{a}) = \sum_{i=1}^N \frac{1}{2} \sigma_i^2 a_i^2$ . Here, the convex conjugate  $U^*(\lambda) = \sum_{i=1}^N \frac{\lambda_i^2}{2\sigma_i^2}$  can be determined by setting  $a_i = \frac{\lambda_i}{\sigma_i^2}$ , which specifies a Gaussian prior on  $\lambda$ . A different example can be obtained by considering the Laplacian prior on  $\lambda$ ,  $U^*(\lambda) = \|\lambda\|_1 = \sum_{i=1}^N |\lambda_i|$ , which leads to the penalty function:

$$U(\mathbf{a}) = \begin{cases} 0 & \|\mathbf{a}\|_\infty = \max_{i=1}^N |a_i| \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

that forces hard inequality constraints.

However, there is an important aspect of the validity of choosing a legal (log) prior,  $U^*$ , which has been ignored in previous studies on the RME principle [11,25,28,29]. To see this, consider the following. By plugging the true distribution,  $p_0(x)$ , instead of the empirical distribution into Equation (16), one obtains:

$$\max_{p(x), \mathbf{a}} \left[ - \int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) - U(\mathbf{a}) \right] \tag{18}$$

$$\text{s.t.} \quad \int_{x \in \mathcal{X}} p(x) f_i(x) \mu(dx) = \int_{x \in \mathcal{X}} p_0(x) f_i(x) + a_i; \quad i = 1, \dots, N \quad (19)$$

The dual problem is:

$$\min_{\lambda \in \Omega} \left[ - \int_{x \in \mathcal{X}} p_0(x) \log p_\lambda(x) + U^*(\lambda) \right] \quad (20)$$

Since  $p_{\hat{\lambda}}(x)$  is the information projection of  $p_0(x)$  to the exponential family,  $\mathcal{E}(x)$ , we must have  $U(\hat{\mathbf{a}}) = U^*(\hat{\lambda}) = 0$  and  $\hat{\mathbf{a}} = 0$ , where  $\hat{\mathbf{a}}$  denotes the  $\mathbf{a}$  corresponding to  $\hat{\lambda}$ . Moreover, as a penalty term, the prior,  $U^*(\lambda)$ , should be nonnegative with a minimum of zero. We call the prior satisfying these conditions a *legal prior*. Both the standard Gaussian prior  $U^*(\lambda) = \sum_{i=1}^N \frac{\lambda_i^2}{2\sigma_i^2}$  and standard Laplacian prior  $U^*(\lambda) = \|\lambda\|_1$  do *not* satisfy  $U^*(\hat{\lambda}) = 0$ . The correct choices for these priors should be  $U^*(\lambda - \hat{\lambda}) = \sum_{i=1}^N \frac{(\lambda_i - \hat{\lambda}_i)^2}{2\sigma_i^2}$  for a Gaussian prior and  $U^*(\lambda - \hat{\lambda}) = \|\lambda - \hat{\lambda}\|_1$  for a Laplacian prior, respectively. Consequently,  $U(\mathbf{a})$  should be chosen as  $U(\mathbf{a}) + \langle \mathbf{a}, \hat{\lambda} \rangle$ . Note that  $U(\mathbf{a}) + \langle \mathbf{a}, \hat{\lambda} \rangle$  does have a value of zero at **zero**, but it is no longer nonnegative.

If we let  $p_{\lambda^\Delta}$  denote the solution to Equation (17), we then obtain the following generalization bound on estimation error without any restrictions on  $U(\mathbf{a})$  or  $U^*(\lambda)$ .

**Theorem 5.** Assume  $\sup_{\lambda \in \Omega} \|\lambda\|_1 < \infty$  and  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f(x)\|_\infty < \infty$ . Then, there exist  $0 < \zeta < \alpha < \infty$ , such that with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} D(p_{\hat{\lambda}}(x) \| p_{\lambda^\Delta}(x)) &= D(p_0(x) \| p_{\lambda^\Delta}(x)) - D(p_0(x) \| p_{\hat{\lambda}}(x)) \\ &\leq \frac{4C_1}{\sqrt{M}} E_{\tilde{x}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + C_2 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} \\ &\quad + U^*(\hat{\lambda}) - U^*(\lambda^\Delta) \end{aligned} \quad (21)$$

We then have the following result that guarantees the universal consistency of RME.

**Corollary 3.** *Universal consistency:* If  $\int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon$  is bounded and  $U^*(\hat{\lambda}) \leq U^*(\lambda^\Delta)$ , then as  $M \rightarrow \infty$ ,  $p_{\lambda^\Delta}(x)$  will converge to  $p_{\hat{\lambda}}(x)$  in terms of their Kullback–Leibler divergence with rate  $O(\frac{1}{\sqrt{M}})$ , for any true distribution,  $p_0(x)$ , and prior distribution.

If we choose  $U^*(\lambda)$  to be a legal prior, then we also have a further result.

**Corollary 4.** Assume  $\sup_{\lambda \in \Omega} \|\lambda\|_1 < \infty$ ,  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f(x)\|_\infty < \infty$  and that  $U^*(\lambda)$  is a legal prior. Then, there exist  $0 < \zeta < \alpha < \infty$ , such that, with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} D(p_{\hat{\lambda}}(x) \| p_{\lambda^\Delta}(x)) &= D(p_0(x) \| p_{\lambda^\Delta}(x)) - D(p_0(x) \| p_{\hat{\lambda}}(x)) \\ &\leq \frac{4C_1}{\sqrt{M}} E_{\tilde{x}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + C_2 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} - U^*(\lambda^\Delta) \end{aligned} \quad (22)$$

Moreover, as  $M \rightarrow \infty$ , both  $U^*(\lambda^\Delta) \rightarrow 0$  and  $p_{\lambda^\Delta}(x)$  will converge to  $p_{\hat{\lambda}}(x)$  in terms of their Kullback–Leibler divergence with rate  $O(\frac{1}{\sqrt{M}})$ , for any true distribution,  $p_0(x)$ , and prior distribution.

The legal prior guarantees RME’s consistency without introducing any regularization parameter, as commonly done in machine learning [9,19]. Since  $U^*(\lambda^\Delta) \geq 0$  for any legal prior, this result shows that RME always obtains a lower generalization bound than ME, even without assuming the truth of the prior.

Using the result of Theorem 5 and the McDiarmid concentration inequality, we are able to derive a generalization bound on the difference between the best expected log-likelihood and the log of the MAP probability. This holds for any regularization cost function, as long as  $U^*(\hat{\lambda}) \leq U^*(\lambda^\Delta)$ , without requiring that  $U^*(\lambda)$  be convex. We note that Dudik *et al.* [11] gave a similar result, but only for the special case of using the  $l_1$  norm as a penalty in the dual formulation; see Equation (17).

For a broad family of regularization functions, *i.e.*, log priors in Equation (17), such that the Hessian matrices of  $\lambda$  (the prior information matrices) are positive definite, we can improve the convergence rate from  $\mathcal{O}(\frac{1}{\sqrt{M}})$  to  $\mathcal{O}(\frac{1}{M})$ . The techniques needed to prove the  $\mathcal{O}(\frac{1}{M})$  convergence rate were first proposed by Zhang [19,30,31] to show the consistency and generalization bounds of various classification methods based on convex risk optimization with the  $l_2$ -penalty.

Define the convex function:

$$L_\lambda(x) = -\log p_\lambda(x) = -\log \frac{\exp(\langle \lambda, f(x) \rangle)}{\int_{x'} \exp(\langle \lambda, f(x') \rangle) \mu(dx')} = \log \int_{x'} \exp(\langle \lambda, f(x') - f(x) \rangle) \mu(dx') \quad (23)$$

Consider training samples  $\tilde{\mathcal{X}}_{M+1} = (x_1, \dots, x_{M+1})$ . Let  $\tilde{\lambda}$  be the solution of Equation (24) with training data  $\tilde{\mathcal{X}}_{M+1}$ .

$$\min_{\lambda \in \Omega} \left[ \frac{1}{M} \sum_{j=1}^{M+1} L_\lambda(x_j) + U^*(\lambda) \right] \quad (24)$$

Let  $\tilde{\lambda}_k$  be the solution of Equation (24) with training sample  $x_k$  removed from the set,  $\tilde{\mathcal{X}}_{M+1}$ . We have the following lemma that extends the quadratic case considered in [19,31]. We prove the result in its full generality, though our proof is essentially a variant of the stability results by Zhang [19,30,31].

**Lemma 1.** *Assume the Hessian matrix of  $U^*(\lambda)$  is positive definite with smallest eigenvalue  $\kappa$ . Then,  $\|\tilde{\lambda} - \tilde{\lambda}_k\|_2 \leq \frac{2}{\kappa M} |\nabla L_{\tilde{\lambda}}(x_k)|$ .*

Finally, we can obtain the following leave-one-out error bound, which has a convergence rate of  $\mathcal{O}(\frac{1}{M})$ .

**Theorem 6.** *Let  $C = \sup_{x,x'} \|f(x) - f(x')\|$ . When the legal prior,  $U(\lambda)$ , is chosen, such that its Hessian matrix of  $\lambda$  is positive definite with smallest eigenvalue  $\kappa$ , the expected generalization error can then be bounded as:*

$$\begin{aligned} E_{\tilde{\mathcal{X}}} D(p_{\tilde{\lambda}}(x) \| p_{\lambda^\Delta}(x)) &= E_{\tilde{\mathcal{X}}} D(p_0(x) \| p_{\lambda^\Delta}(x)) - D(p_0(x) \| p_{\tilde{\lambda}}(x)) \\ &\leq \frac{C^2}{\kappa M} \left( 1 - \exp(-E_X L_{\tilde{\lambda}}(X)) \right) \end{aligned} \quad (25)$$

### 3. Maximum Entropy Density Estimation: Incomplete Data

Here, we consider a variable to be “latent” or “hidden” if it is never observed, but causally effects the observations [6]. In practice, many of the natural patterns we wish to classify are the result of causal processes that have a hidden hierarchical structure, yielding data that does not report the value of *latent* variables [6]. Obtaining fully labeled data is tedious or impossible in many realistic cases. This motivates us to propose an *unsupervised* statistical learning technique, the latent maximum entropy (LME) principle, which is still formulated in terms of maximizing entropy, except that we must now change the problem formulation to respect hidden variables.

Following the terminology of the expectation-maximization (EM) algorithm [6], let  $Y \in \mathcal{Y}$  be the observed incomplete data,  $Z \in \mathcal{Z}$  be the missing data and then  $X = (Y, Z) \in \mathcal{X}$  be the random variable denoting the complete data. That is,  $X = (Y, Z)$ . For example,  $Y$  might be observed natural language in the form of text and  $Z$  might be its missing syntactic and semantic information. If we let  $p(x)$  and  $p(y)$  denote the densities of  $X$  and  $Y$ , respectively, and let  $p(z|y)$  denote the conditional density of  $Z$  given  $Y$ , then  $p(y) = \int_{z \in \mathcal{Z}} p(x) \mu(dz)$ , and  $p(x) = p(y)p(z|y)$ . The LME principle can be stated as follows.

Given features  $f_1(x), \dots, f_N(x)$ , specifying the properties that we would like to match in the data, select a joint probability model,  $p(x)$ , from the space of all probability distributions,  $\mathcal{P}(x)$ , over  $\mathcal{X}$ , to maximize the entropy:

$$\max_{p(x) \in \mathcal{P}} \left[ - \int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) \right] \tag{26}$$

$$\text{s.t. } \int_{x \in \mathcal{X}} f_i(x) p(x) \mu(dx) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) p(z|y) \mu(dz), \quad i = 1, \dots, N \tag{27}$$

where  $x = (y, z)$  and  $\tilde{p}(y)$  is the empirical distribution of the observed data,  $\mathcal{Y} = (y_1, \dots, y_M)$  denotes the set of observed  $Y$  values and  $p(z|y)$  is the conditional distribution of latent variables given the observed data. Intuitively, the constraints specify that we require the expectations of  $f_i(X)$  in the joint model to match their empirical expectations on the incomplete data,  $Y$ , taking into account the structure of the implied dependence of the unobserved component,  $Z$ , on  $Y$ .

Note that the conditional distribution,  $p(z|y)$ , implicitly encodes the latent structure and is a nonlinear mapping of  $p(x)$ . That is,  $p(z|y) = \frac{p(y,z)}{\int_{z' \in \mathcal{Z}} p(y,z') \mu(dz')} = \frac{p(x)}{\int_{z' \in \mathcal{Z}} p(x') \mu(dz')}$ , where  $x = (y, z)$  and  $x' = (y, z')$  by definition. Clearly,  $p(z|y)$  is a nonlinear function of  $p(x)$  because of the division.

Unfortunately, there is no simple optimal solution for  $p(x)$  in Equations (26) and (27). However, a good approximation can be obtained by restricting the model to  $p_\lambda(x) = \Phi_\lambda^{-1} \exp \left( \sum_{i=1}^N \lambda_i f_i(x) \right)$  where  $\Phi_\lambda = \int_{x \in \mathcal{X}} \exp \left( \sum_{i=1}^N \lambda_i f_i(x) \right) \mu(dx)$  is the normalization constant. This restriction provides a free parameter,  $\lambda_i$ , for each feature function,  $f_i(x)$ . By adopting such a “log-linear” restriction, it turns out that we can formulate a practical algorithm for approximately satisfying the LME principle.

In [5], we exploited the following connection between LME and maximum likelihood estimation (MLE) to derive a practical training algorithm.

**Theorem 7.** [5] *Under the log-linear assumption, maximizing the likelihood of log-linear models on incomplete data is equivalent to satisfying the feasibility constraints of the LME principle. That is, the*

only distinction between MLE and LME in log-linear models is that, among local maxima (feasible solutions), LME selects the model with the maximum entropy, whereas MLE selects the model with the maximum likelihood.

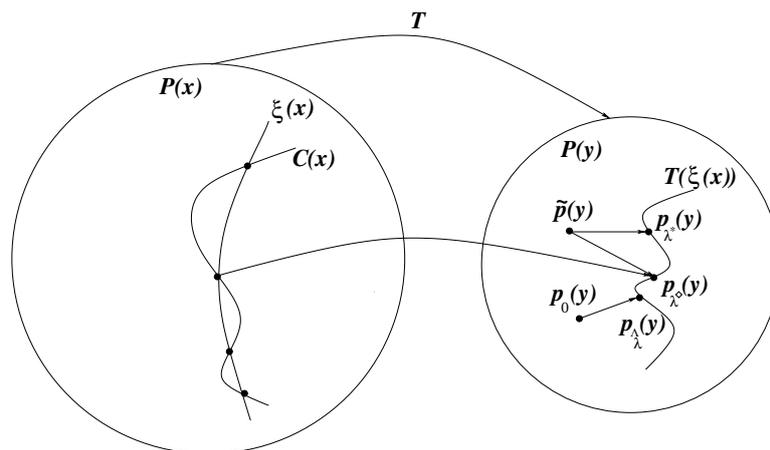
Define  $L_\lambda(y) = -\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \log p_\lambda(y)$  and  $H(\lambda, \lambda) = -\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p_\lambda(z|y) \log p_\lambda(z|y) \mu(dz)$ . The latter is the conditional entropy of a hidden variable over observed sample data  $\tilde{\mathcal{Y}}$  that measures the uncertainty of the hidden variables. Then, in the case where  $\lambda$  is a feasible log-linear solution according to Equation (27), we have the following relationship between likelihood over observed data and the entropy of the joint model.

**Corollary 5.** [5] *If  $\lambda$  is in the set of feasible solutions, then:*

$$L_\lambda(y) = H(p_\lambda(x)) - H(\lambda, \lambda) \tag{28}$$

We will use the following notation and terminology throughout the analysis below. Denote the manifold of the nonlinear constraint set in Equation (27) as  $\mathcal{C}$ . We then define  $p_{\tilde{\lambda}}(y) = \arg \max_{p_\lambda(x) \in \mathcal{E}} \int_{y \in \mathcal{Y}} p_0(y) \log p_\lambda(y)$  as the nearest point in terms of  $D(p_0(y) || p_\lambda(y))$  from  $p_0(y)$  to the marginalized exponential family over  $z$ , using  $p_\lambda(y) = \int_{z \in \mathcal{Z}} p_\lambda(y, z) \mu(dz)$ , where  $p_\lambda(x) \in \mathcal{E}(x)$ ; see Figure 2.

**Figure 2.** The operator,  $T$ , denotes the marginalization of  $p(x)$  over  $z$  and maps the entire space of all probability distributions,  $\mathcal{P}(x)$  over  $\mathcal{X}$ , into the space of all probability distributions,  $\mathcal{P}(y)$  over  $\mathcal{Y}$ . Here,  $p_{\tilde{\lambda}}(y)$  is the information projection of  $p_0(y)$  to the marginalized exponential family,  $\mathcal{E}$ ;  $p_{\lambda^*}(y)$  is the information projection of  $\tilde{p}(y)$  to the marginalized exponential family,  $\mathcal{E}$ ; and  $p_{\lambda^\circ}(y)$  is the distribution that in joint model space has the highest entropy among the intersection points of the exponential family,  $\mathcal{E}$ , and the nonlinear constraint set,  $\mathcal{C}$ .



In [29,32], we formulate the regularized latent maximum entropy principle (RLME) as the following:

$$\max_{p(x), \mathbf{a}} \left[ -\int_{x \in \mathcal{X}} p(x) \log p(x) \mu(dx) - U(\mathbf{a}) \right] \tag{29}$$

subject to:

$$\int_{x \in \mathcal{X}} p(x) f_i(x) \mu(dx) = \sum_y \tilde{p}(y) \int_z p(z|y) f_i(y, z) \mu(dz) + a_i; \quad i = 1, \dots, N \tag{30}$$

Again  $\mathbf{a} = (a_1, \dots, a_N)$ , where  $a_i$  is the error for each constraint, and  $U : \mathfrak{R}^N \rightarrow \mathfrak{R}$  is a convex function with its minimum at zero.

The standard *maximum a posteriori* (MAP) estimate minimizes the negative penalized log-likelihood  $R(\lambda) = -\sum_y \tilde{p}(y) \log p_\lambda(y) + U^*(\lambda)$ .

Our key result in [29,32] is that locally minimizing  $R(\lambda)$  is equivalent to satisfying the feasibility constraints in Equation (30) of the RLME principle.

**Theorem 8.** [29,32] *Under the log-linear assumption, locally maximizing the posterior probability of log-linear models on incomplete data is equivalent to satisfying the feasibility constraints of the RLME principle. That is, the only distinction between MAP and RLME in log-linear models is that, among local maxima (feasible solutions), RLME selects the model with the maximum regularized entropy, whereas MAP selects the model with the maximum posterior probability*

**Corollary 6.** [29,32] *If  $\lambda^*$  is in the set of feasible solutions of Equation (30), then  $R(\lambda) = H(p_\lambda) - U(\mathbf{a}) - H(\lambda, \lambda)$ .*

### 3.1. Consistency and Generalization Bounds for Estimation Error

To measure the quality of the maximum likelihood and maximum entropy estimates, we do not consider the divergence of the models in the original joint space,  $\mathcal{P}(x)$ . Instead, we consider the marginalized models in the observed data space,  $\mathcal{P}(y)$ . However, to measure the divergence between models in the observed data space, we have to take the difference of  $D(p_0(y)||p_{\lambda^*}(y))$  and  $D(p_0(y)||p_{\hat{\lambda}}(y))$ , even though, technically, the Pythagorean property no longer holds in this case. Nevertheless, this still gives a useful measure of the approximation quality.

#### 3.1.1. Maximum Likelihood Estimate

We first establish consistency and provide generalization bounds for the maximum likelihood density estimate,  $p_{\lambda^*}(y)$ . Note that if we attempt to use a technique similar to the complete data case here, we will obtain a bound that is governed by the covering number of the *log* of the marginal feature functions  $\mathcal{F}(y) = \int_{z \in \mathcal{Z}} \exp(\langle \lambda, f(y, z) \rangle) \mu(dz)$ . Bounding the covering number of  $\log\text{-int-exp}\mathcal{F}(x)$  is more difficult than bounding it for  $\mathcal{F}(y)$  directly. To cope with this issue, we use the refined version of the Rademacher comparison inequality proposed in [24] to eliminate the log function. This is a slightly different approach than that taken by Rakhlin *et al.* [16], who, instead, use the contraction technique of [15,20,33] to derive bounds for mixture model density estimation. Here, we pursue a streamlined analysis that avoids working with the likelihood ratio (also, see [34]), hence avoiding the second application of contraction, which results in tighter constants.

**Theorem 9.** *Assume for all  $\lambda \in \Omega$  and for all  $y \in \mathcal{Y}$ , we have  $0 < a \leq \mathcal{F}(y) \leq b$ . Then, there exist  $0 < \zeta < \alpha < \infty$  and positive constants,  $C_3, C_4 \in \mathfrak{R}^+$ , such that, with a probability of at least  $1 - \eta$ , for any dataset of size  $M$  drawn from  $p_0(y)$ ,*

$$\begin{aligned}
 & D(p_0(y)||p_{\lambda^*}(y)) - D(p_0(y)||p_{\hat{\lambda}}(y)) & (31) \\
 \leq & \frac{4C_3}{\sqrt{M}} E_{\tilde{y}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon \right] + C_4 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}}
 \end{aligned}$$

where  $\mathcal{N}(\mathcal{F}(y), \epsilon, d_y)$  is the random covering number of the marginal feature functions,  $\mathcal{F}(y) = \int_{z \in \mathcal{Z}} \exp(\langle \lambda, f(y, z) \rangle) \mu(dz)$ , at scale  $\epsilon$  with empirical Euclidean distance  $d_y$  on sample data  $\tilde{\mathcal{Y}}$ .

Similarly, we can eliminate the assumption of boundedness on the parameters and feature functions,  $\mathcal{F}(y)$ , by using a result adapted from [24].

**Theorem 10.** Assume that there exists a positive number,  $K(\mathcal{F})$ , such that for all  $\tau > 0$ :

$$\log E_{p_0(y)} \left( p_{\lambda^\bullet(y)}^{2\tau} - \frac{1}{p_{\lambda^\bullet(y)}^{2\tau}} \right) \leq (\tau K(\mathcal{F}))^2 \tag{32}$$

where  $\lambda^\bullet$  are the parameters  $\left( p_{\lambda^\bullet(y)}^{2\tau} - \frac{1}{p_{\lambda^\bullet(y)}^{2\tau}} \right)$  achieving the maximum subject to  $E_{p_{\lambda^\bullet(x)}}(f(x)) = E_{p_{\lambda^\bullet(z|y)}}(f(y, z))$ . Then, for all  $\lambda \in \Omega$ , we have with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} & E_{p_0(x)} \log p_\lambda(y) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(y_j) \\ & \leq E_{\tilde{\mathcal{Y}}} \sup_{\lambda \in \Omega, f(x) \in \mathcal{F}(x)} \left( E_{p_0(x)} \log \mathcal{F}(y) - E_{\tilde{p}(x)} \log \mathcal{F}(y) \right) + K(\mathcal{F}) \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} \end{aligned} \tag{33}$$

Using the above result, Theorem 9 can be proven by replacing  $C_4$  with  $K(\mathcal{F})$ . Again, since the value,  $K(\mathcal{F})$ , is hard to determine in practice, we will state our results below in terms of a bound on feature functions, but as before, the reader should bear in mind that the bound on feature functions can be replaced by  $K(\mathcal{F})$ .

From the results above, we can establish the following consistency property.

**Corollary 7.** Universal consistency: If  $\int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon$  is bounded, then  $p_{\lambda^*}(y)$  will converge to  $p_{\hat{\lambda}}(y)$  (in terms of the difference of the Kullback–Leibler divergence to the true distribution,  $p_0(y)$ ) with rate  $\mathcal{O}(\frac{1}{\sqrt{M}})$ , regardless of the form of true distribution  $p_0(y)$ .

Similar to the complete data case, using the result of Theorem 9 and the McDiarmid concentration inequality, we are also able to derive the generalization bound for the difference of the best expected log-likelihood and the maximum empirical log-likelihood.

### 3.1.2. Latent Maximum Entropy Estimate

Let  $p_{\lambda^\diamond}(y)$  denote the maximum entropy estimate of Equation (26) over the exponential family,  $\mathcal{E}$ . We use similar techniques to the case of complete data regularized maximum entropy (Section 2.1.2) to prove consistency and generalization bounds for using the latent maximum entropy density estimate,  $p_{\lambda^\diamond}(y)$ .

**Theorem 11.** (LME principle) Assume for all  $\lambda \in \Omega$  and for all  $y \in \mathcal{Y}$ , we have  $0 < a \leq \mathcal{F}(y) \leq b$ . Then, there exist  $0 < \zeta < \alpha < \infty$ , such that with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} D(p_0(y) \| p_{\lambda^\diamond}(y)) - D(p_0(y) \| p_{\hat{\lambda}}(y)) & \leq \frac{4C_3}{\sqrt{M}} E_{\tilde{\mathcal{Y}}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon \right] \\ & \quad + C_4 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} + E_{\tilde{p}(y)} \log \frac{p_{\hat{\lambda}}(y)}{p_{\lambda^\diamond}(y)} \end{aligned} \tag{34}$$

Using this result, we can then easily establish the following consistency property.

**Corollary 8.** *Universal consistency: If  $\int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon$  is bounded and also  $E_{\hat{p}(y)} \log p_{\hat{\lambda}}(y) \leq E_{\hat{p}(y)} \log p_{\lambda^{\diamond}}(y)$ , then  $p_{\lambda^{\diamond}}(y)$  will converge to  $p_{\hat{\lambda}}(y)$  (in terms of the difference of the Kullback–Leibler divergence to the true distribution,  $p_0(y)$ ) with rate  $\mathcal{O}(\frac{1}{\sqrt{M}})$ , for any true distribution,  $p_0(y)$ .*

Corollary 8 gives a sufficient condition, i.e.,  $E_{\hat{p}(y)} \log p_{\hat{\lambda}}(y) \leq E_{\hat{p}(y)} \log p_{\lambda^{\diamond}}(y)$ , that leads to the universal consistency of latent maximum entropy estimation. This perhaps partially explains our observations of experimental results on synthetic data conducted in [5].

Note that in the proof of Theorem 11 and Corollary 8, it is not necessary to restrict  $p_{\lambda^{\diamond}}$  to be the model that has global maximum joint entropy over all feasible log-linear solutions. It turns out that the conclusion still holds for all feasible log-linear models,  $p_{\lambda}(y)$ , that have greater empirical log-likelihood,  $E_{\hat{p}(y)} \log p_{\lambda}(y)$ , than the empirical log-likelihood,  $E_{\hat{p}(y)} \log p_{\hat{\lambda}}(y)$ , of the optimal expected log-likelihood estimate,  $p_{\hat{\lambda}}(y)$ . That is, as the sample size grows, any of these feasible log-linear models will converge to  $p_{\hat{\lambda}}(y)$  (in terms of the difference of the Kullback–Leibler divergence to the true distribution,  $p_0(y)$ ) with rate  $\mathcal{O}(\frac{1}{\sqrt{M}})$ .

### 3.1.3. Maximum a Posteriori Estimate

In a similar manner, it is straightforward to have the following generalization bound for the MAP estimate,  $p_{\lambda^{\Delta}}(y)$ .

**Theorem 12.** (MAP principle) *Assume for all  $\lambda \in \Omega$  and for all  $y \in \mathcal{Y}$ ,  $0 < a \leq \mathcal{F}(y) \leq b$ . Then, with a probability of at least  $1 - \eta$ ,*

$$\begin{aligned}
 & D(p_0(y) \| p_{\lambda^{\Delta}}(y)) - D(p_0(y) \| p_{\hat{\lambda}}(y)) \tag{35} \\
 & \leq \frac{4C_3}{\sqrt{M}} E_{\hat{\chi}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_x)} d\epsilon \right] + C_4 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} + U^*(\hat{\lambda}) - U^*(\lambda^{\Delta})
 \end{aligned}$$

By the above theorem, one can easily obtain the following consistency result.

**Corollary 9.** *Universal consistency: If  $\int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon$  is bounded and  $U^*(\hat{\lambda}) \leq U^*(\lambda^{\Delta})$ , then  $p_{\lambda^{\Delta}}(y)$  will converge to  $p_{\hat{\lambda}}(y)$  in terms of the difference of the Kullback–Leibler divergence to the true distribution,  $p_0(y)$ , with rate  $\mathcal{O}(\frac{1}{\sqrt{M}})$  without assuming the form of the true distribution,  $p_0(y)$ , nor the true prior distribution.*

### 3.1.4. Regularized Latent Maximum Entropy Estimate

We can also, in a similar manner, establish the following generalization bound for the RLME estimate,  $p_{\lambda^{\diamond}}(y)$ .

**Theorem 13.** (RLME principle) *Assume for all  $\lambda \in \Omega$  and for all  $y \in \mathcal{Y}$ ,  $0 < a \leq \mathcal{F}(y) \leq b$ . Then, with a probability of at least  $1 - \eta$ ,*

$$\begin{aligned}
 D(p_0(y) \| p_{\lambda^{\diamond}}(y)) - D(p_0(y) \| p_{\hat{\lambda}}(y)) & \leq \frac{4C_3}{\sqrt{M}} E_{\hat{\chi}} \left[ \int_{\zeta}^{\alpha} \sqrt{\mathcal{N}(\mathcal{F}(y), \epsilon, d_x)} d\epsilon \right] + C_4 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} \tag{36} \\
 & + E_{\hat{p}(y)} \log p_{\hat{\lambda}}(y) - E_{\hat{p}(y)} \log p_{\lambda^{\diamond}}(y) + U^*(\hat{\lambda}) - U^*(\lambda^{\diamond})
 \end{aligned}$$

By the above theorem, we can easily obtain the following consistency result.

**Corollary 10.** *Universal consistency: If  $\int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon$  is bounded and  $E_{\tilde{p}(y)} \log p_{\hat{\lambda}}(y) + U^*(\hat{\lambda}) \leq E_{\tilde{p}(y)} \log p_{\lambda^{\diamond}}(y) + U^*(\lambda^{\diamond})$ , then  $p_{\lambda^{\diamond}}(y)$  will converge to  $p_{\hat{\lambda}}(y)$  in terms of the difference of the Kullback–Leibler divergence to the true distribution,  $p_0(y)$ , with rate  $\mathcal{O}(\frac{1}{\sqrt{M}})$  without assuming the form of true distribution  $p_0(y)$  and true prior distribution.*

#### 4. Conclusions

We have investigated the statistical properties of using the maximum entropy principle for density estimation, in both the complete and incomplete data cases. For complete data, maximum entropy is equivalent to maximum likelihood estimation in a Markov random field. Here, we derived bounds on the generalization error based on the complexity of linear combinations of feature functions, and used this to establish a form of universal consistency. We then provided a similar analysis for regularized maximum entropy estimation, which, interestingly, yields a better generalization bound (and maintains consistency) for any legal prior. Moreover, if the information matrix of the prior is positive definite, we can further show that the convergence rate can be improved to  $\mathcal{O}(\frac{1}{M})$  instead of  $\mathcal{O}(\frac{1}{\sqrt{M}})$ . For incomplete data, maximum entropy is no longer equivalent to maximum likelihood estimation, and the analysis becomes more difficult. Nevertheless, we established bounds on the generalization error of maximum likelihood in terms of the complexity of the *marginalized* feature functions, again achieving a form of universal consistency. With additional assumptions, we were able to extend this analysis to apply it to latent maximum entropy estimation and to prove its universal consistency, as well. Analogous conclusions can be drawn for regularized situations. Finally, we note that an alternative analysis can be based on replacing the Kullback–Leibler divergence with the more general Bregman distance [35,36]. The analysis here can be easily extended to this more general setting.

In our future work, we are planning to study the trade-off between approximation error and estimation error to select the best set of feature functions.

#### Acknowledgments

We thank Dale Schuurmans for his assistance and Tong Zhang for his comments. The research was supported by the Alberta Innovates Centre for Machine Learning, University of Alberta, Canada.

#### Conflicts of Interest

The authors declare no conflict of interest.

#### Appendix

In the appendix, we give proofs of the theorems, lemmas and corollaries.

**Proof of Theorem 1.**

*Proof.* The techniques we use are quite standard [15,16,20,37] and have appeared in many papers. To be concise, following lecture notes 14–15 in [37], the first key technique we are using is the method of bounded differences [38]. Define:

$$\begin{aligned}
 h(x_1, \dots, x_M) &= \sup_{p_\lambda(x) \in \mathcal{E}} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| \\
 &= \sup_{\lambda \in \Omega, f \in \mathcal{F}} \left| E_{p_0(x)} \langle \lambda, f(x) \rangle - E_{\tilde{p}(x)} \langle \lambda, f(x) \rangle \right|
 \end{aligned}$$

Then:

$$\begin{aligned}
 & \left| h(x_1, \dots, x_M) - h(x_1, \dots, x_{k-1}, x'_k, x_{k+1}, \dots, x_M) \right| \tag{37} \\
 &= \left| \sup_{\lambda \in \Omega, f \in \mathcal{F}} \left| E_{p_0(x)} \langle \lambda, f(x) \rangle - \frac{1}{M} \sum_{j=1}^M \langle \lambda, f(x_j) \rangle \right| \right. \\
 & \quad \left. - \sup_{\lambda \in \Omega, f \in \mathcal{F}} \left| E_{p_0(x)} \langle \lambda, f(x) \rangle - \frac{1}{M} \left( \sum_{j=1, j \neq k}^M \langle \lambda, f(x_j) \rangle + \langle \lambda, f(x'_k) \rangle \right) \right| \right| \\
 &\leq \sup_{\lambda \in \Omega, f \in \mathcal{F}} \frac{1}{M} \left| \langle \lambda, (f(x_k) - f(x'_k)) \rangle \right| \\
 &\leq \frac{2}{M} \sup_{\lambda \in \Omega, f \in \mathcal{F}} \|\lambda\|_1 \sup_{x \in \mathcal{X}} \|f(x)\|_\infty = \frac{C_2}{M}
 \end{aligned}$$

By the McDiarmid concentration inequality in Equation [38], we have:

$$P\left(h(x_1, \dots, x_M) - E_{\tilde{\chi}} h(x_1, \dots, x_M) \geq \delta\right) \leq \exp\left(\frac{-2\delta^2}{\sum_{j=1}^M \left(\frac{C_2}{M}\right)^2}\right) = \exp\left(\frac{-2M\delta^2}{C_2^2}\right)$$

Let  $\eta = \exp\left(\frac{-2M\delta^2}{C_2^2}\right)$ , i.e.,  $\delta = C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}}$ . Then:

$$P\left(h(x_1, \dots, x_M) - E_{\tilde{\chi}} h(x_1, \dots, x_M) \geq C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}}\right) \leq \eta \tag{38}$$

Therefore, with a probability of at least  $1 - \eta$ ,

$$\begin{aligned}
 & \sup_{p_\lambda(x) \in \mathcal{E}(x)} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| \\
 &\leq E_{\tilde{\chi}} \sup_{\lambda \in \Omega, f(x) \in \mathcal{F}(x)} \left| E_{p_0(x)} \langle \lambda, f(x) \rangle - E_{\tilde{p}(x)} \langle \lambda, f(x) \rangle \right| + C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}}
 \end{aligned}$$

Next, we use the symmetrization technique of [33,39], which states that if:

$$Z(\tilde{\mathcal{X}}) = \sup_{g \in \mathcal{G}} \left| E g(x) - \frac{1}{M} \sum_{j=1}^M g(x_j) \right| \quad \text{and} \quad R(\tilde{\mathcal{X}}) = \sup_{g \in \mathcal{G}} \left| \frac{1}{M} \sum_{j=1}^M \omega_j g(x_j) \right|$$

then:

$$EZ_{\tilde{\mathcal{X}}}(\tilde{\mathcal{X}}) \leq 2ER_{\tilde{\mathcal{X}},\omega}(\tilde{\mathcal{X}})$$

where  $\omega = (\omega_1, \dots, \omega_M)$  is a Rademacher sequence. We then have with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} & \sup_{p_\lambda(x) \in \mathcal{E}(x)} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| \\ & \leq 2E_{\tilde{\mathcal{X}},\omega} \sup_{\lambda \in \Omega, f \in \mathcal{F}} \left| \frac{1}{M} \sum_{j=1}^M \omega_j \langle \lambda, f(x_j) \rangle \right| + C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}} \end{aligned}$$

A classical result of Dudley establishes that Rademacher averages over linear combinations in  $\mathcal{F}(x)$  are bounded by Dudley’s entropy integral [16,20,22,37],

$$E_\omega \sup_{\lambda \in \Omega, f(x) \in \mathcal{F}(x)} \left| \frac{1}{M} \sum_{j=1}^M \epsilon \langle \lambda, f(x_j) \rangle \right| \leq \frac{C_1}{\sqrt{M}} \int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon$$

where  $0 < \zeta < \alpha < \infty$ ; provided, as observed in [20,40], that  $\sup_{\lambda \in \Omega} \|\lambda\|_\infty$  and  $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} \|f(x)\|_\infty$  are bounded. One can then show that with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} & \sup_{p_\lambda(x) \in \mathcal{E}(x)} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| \tag{39} \\ & \leq 2 \frac{C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}} \end{aligned}$$

Therefore:

$$\begin{aligned} D(p_{\hat{\lambda}}(x) \| p_{\lambda^*}(x)) &= D(p_0(x) \| p_{\lambda^*}(x)) - D(p_0(x) \| p_{\hat{\lambda}}(x)) && \text{by Equation (8)} \\ &= \left( E_{p_0(x)} \log p_{\hat{\lambda}}(x) - E_{\tilde{p}(x)} \log p_{\hat{\lambda}}(x) \right) + \left( E_{\tilde{p}(x)} \log p_{\lambda^*}(x) \right. \\ &\quad \left. - E_{p_0(x)} \log p_{\lambda^*}(x) \right) + \left( E_{\tilde{p}(x)} \log p_{\hat{\lambda}}(x) - E_{\tilde{p}(x)} \log p_{\lambda^*}(x) \right) \\ &\leq 2 \sup_{p_\lambda(x) \in \mathcal{E}(x)} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| + \frac{1}{M} \sum_{j=1}^M \log \frac{p_{\hat{\lambda}}(x_j)}{p_{\lambda^*}(x_j)} \\ &\leq 2 \sup_{p_\lambda(x) \in \mathcal{E}(x)} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| \end{aligned}$$

where the second inequality comes from the fact that  $\frac{1}{M} \sum_{j=1}^M \log \frac{p_{\hat{\lambda}}(x_j)}{p_{\lambda^*}(x_j)} \leq 0$ , since  $p_{\lambda^*}(x)$  has maximum likelihood in the exponential family,  $\mathcal{E}(x)$ .

Therefore, with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} D(p_{\hat{\lambda}}(x) \| p_{\lambda^*}(x)) &= D(p_0(x) \| p_{\lambda^*}(x)) - D(p_0(x) \| p_{\hat{\lambda}}(x)) \\ &\leq \frac{4C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + C_2 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} \end{aligned}$$

□

**Proof of Theorem 2.**

*Proof.* Choose the function to be  $\log p_\lambda(x), \lambda \in \Omega$ , then  $\cosh\left(2\tau \log p_\lambda(x)\right) = \frac{1}{2}\left(p_\lambda(x)^{2\tau} - \frac{1}{p_\lambda(x)^{2\tau}}\right)$ . By Theorem 3 in [24], we have that if there exists a positive number,  $K(\mathcal{F})$ , such that for all  $\tau > 0$ ,

$$\log E_{p_0(x)} \sup_{\lambda \in \Omega} \left( p_\lambda(x)^{2\tau} - \frac{1}{p_\lambda(x)^{2\tau}} \right) \leq \left( \tau K(\mathcal{F}) \right)^2$$

then, for all  $\lambda \in \Omega$  with a probability of at least  $1 - \eta$ , (14) will hold.

Next, take the derivative of  $\left( p_\lambda(x)^{2\tau} - \frac{1}{p_\lambda(x)^{2\tau}} \right)$  with respect to  $\lambda$  and set this to zero. After some routine calculation, we obtain:

$$\left( p_\lambda(x)^{2\tau-1} - \frac{1}{p_\lambda(x)^{2\tau+1}} \right) p_\lambda(x) \left( E_{p_\lambda(x)}(f_i(x)) - f_i(x) \right) = 0; i = 1 \dots N$$

Thus,  $\lambda^\bullet = \arg \sup_{\lambda \in \Omega} \left( p_\lambda(x)^{2\tau} - \frac{1}{p_\lambda(x)^{2\tau}} \right)$  are those  $\lambda^\bullet$ , such that  $E_{p_{\lambda^\bullet}(x)}(f(x)) = f(x)$ , which can be uniquely determined by the maximum entropy approach for each fixed  $x$ . □

**Proof of Theorem 3.**

*Proof.* Define  $u_i = f_i(x), i = 1, \dots, N$ . We consider real valued linear function classes of the following form:

$$L(\lambda, u) = \lambda \cdot u = \sum_{i=1}^N \lambda_i u_i \tag{40}$$

Zhang showed that  $\log \mathcal{N}_2(L(\lambda, u), \epsilon, d_u) \leq \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \log(2N + 1)$  (Theorem 3 in [21]) and  $\log N_2(L(\lambda, u), \epsilon, d_u) \leq \left\lceil \frac{a^2 b^2}{\epsilon^2} \right\rceil \log(2M + 1)$  (Corollary 3 in [21]) Since  $\mathcal{N}_2(L(\lambda, u), \epsilon, d_u) = \mathcal{N}_2(\mathcal{F}(x), \epsilon, d_x)$ , we have the conclusion. □

**Proof of Theorem 4.**

*Proof.*

$$\begin{aligned} \left| \mathcal{L}_0(\hat{\lambda}) - \tilde{\mathcal{L}}(\lambda^*) \right| &= \left| E_{p_0(x)} \log p_{\hat{\lambda}}(x) - \frac{1}{M} \sum_{j=1}^M \log p_{\lambda^*}(x_j) \right| \\ &\leq \left| E_{p_0(x)} \log p_{\hat{\lambda}}(x) - E_{p_0(x)} \log p_{\lambda^*}(x) \right| + \left| E_{p_0(x)} \log p_{\lambda^*}(x) - \frac{1}{M} \sum_{j=1}^M \log p_{\lambda^*}(x_j) \right| \end{aligned}$$

Thus, combining the inequalities of Equation (9) and the uniform bound Equation (12), we have with probability  $1 - \eta$ ,

$$\left| \mathcal{L}_0(\hat{\lambda}) - \tilde{\mathcal{L}}(\lambda^*) \right| \leq \frac{6C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + 3C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}}$$

□

**Proof of Theorem 5.**

*Proof.* The proof is similar to the ME case. Consider the chain of inequalities:

$$\begin{aligned}
 D(p_{\hat{\lambda}}(x)||p_{\lambda^\Delta}(x)) &= D(p_0(x)||p_{\lambda^\Delta}(x)) - D(p_0(x)||p_{\hat{\lambda}}(x)) && \text{by (7)} \\
 &= \left( E_{p_0(x)} \log p_{\hat{\lambda}}(x) - E_{\bar{p}(x)} \log p_{\hat{\lambda}}(x) \right) + \left( E_{\bar{p}(x)} \log p_{\lambda^\Delta}(x) \right. \\
 &\quad \left. - E_{p_0(x)} \log p_{\lambda^\Delta}(x) \right) + \left( E_{\bar{p}(x)} \log p_{\hat{\lambda}}(x) - E_{\bar{p}(x)} \log p_{\lambda^\Delta}(x) \right) \\
 &\leq 2 \sup_{p_\lambda(x) \in \mathcal{E}} \left| E_{p_0(x)} \log p_\lambda(x) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(x_j) \right| \\
 &\quad + \left( [E_{\bar{p}(x)} \log p_{\hat{\lambda}}(x) - U^*(\hat{\lambda})] - [E_{\bar{p}(x)} \log p_{\lambda^\Delta}(x) - U^*(\lambda^\Delta)] \right) \\
 &\quad + U^*(\hat{\lambda}) - U^*(\lambda^\Delta)
 \end{aligned}$$

where the second inequality follows from the fact that  $[E_{\bar{p}} \log p_{\hat{\lambda}} - U^*(\lambda)] - [E_{\bar{p}} \log p_{\lambda^\Delta} - U^*(\lambda^\Delta)] \leq 0$ , since  $p_{\lambda^\Delta}$  maximizes the *a posteriori* objective over the exponential family,  $\mathcal{E}$ .

Therefore, with a probability of at least  $1 - \eta$ ,

$$\begin{aligned}
 D(p_{\hat{\lambda}}(x)||p_{\lambda^\Delta}(x)) &= D(p_0(x)||p_{\lambda^\Delta}(x)) - D(p_0(x)||p_{\hat{\lambda}}(x)) \\
 &\leq \frac{4C_1}{\sqrt{M}} E_{\tilde{\mathcal{X}}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(x), \epsilon, d_x)} d\epsilon \right] + C_2 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} + U^*(\hat{\lambda}) - U^*(\lambda^\Delta)
 \end{aligned}$$

□

**Proof of Corollary 4.**

*Proof.* Since  $U^*(\lambda)$  is a legal prior,  $U^*(\hat{\lambda}) = 0$ . We thus have the inequality by the last theorem. As  $M \rightarrow \infty$ , the right-hand side of the last inequality goes to  $-U^*(\lambda^\Delta)$ , which is nonpositive; however, the left-hand side is nonnegative. Therefore, we must have  $U^*(\lambda^\Delta) = 0$ . □

**Proof of Lemma 1.**

*Proof.* By the definition of  $\tilde{\lambda}$ , we have:

$$\frac{1}{M} \sum_{i=1}^{M+1} \nabla L_{\tilde{\lambda}}(x_i) + \nabla U^*(\tilde{\lambda}) = 0 \tag{41}$$

The Bregman divergence of  $L_\lambda$  for the Markov random field model can be written as:

$$B_{L(x)}(\lambda_1, \lambda_2) = L_{\lambda_2}(x) - L_{\lambda_1}(x) - \nabla L_{\lambda_1}(x)^T (\lambda_2 - \lambda_1)$$

which is always nonnegative.

Then, we have:

$$\begin{aligned}
 \frac{1}{M} \sum_{i=1, i \neq k}^{M+1} L_{\tilde{\lambda}_k}(x_i) &\geq \frac{1}{M} \sum_{i=1, i \neq k}^{M+1} \left[ L_{\tilde{\lambda}_k}(x_i) - B_{L(x)}(\tilde{\lambda}(x_i), \tilde{\lambda}_k(x_i)) \right] \\
 &= \frac{1}{M} \sum_{i=1, i \neq k}^{M+1} L_{\tilde{\lambda}}(x_i) + \frac{1}{M} \sum_{i=1, i \neq k}^M \nabla L_{\tilde{\lambda}}(x)^T (\tilde{\lambda}_k - \tilde{\lambda})
 \end{aligned} \tag{42}$$

By Taylor’s expansion, we know there exists  $\theta^\Delta \in \Omega \subseteq \mathfrak{R}^L$ , such that:

$$\begin{aligned} U^*(\tilde{\lambda}_k) &= U^*(\tilde{\lambda}) + \nabla U^*(\tilde{\lambda})(\tilde{\lambda}_k - \tilde{\lambda}) + \frac{1}{2}(\tilde{\lambda}_k - \tilde{\lambda})^T \nabla^2 U^*(\theta^\Delta)(\tilde{\lambda}_k - \tilde{\lambda}) \\ &\geq U^*(\tilde{\lambda}) + \nabla U^*(\tilde{\lambda})(\tilde{\lambda}_k - \tilde{\lambda}) + \frac{\kappa}{2} \|\tilde{\lambda}_k - \tilde{\lambda}\|_2^2 \end{aligned} \tag{43}$$

where the last inequality is due to the assumption that the Hessian matrix of  $U^*(\lambda)$  is positive definite with the smallest eigenvalue  $\kappa > 0$ .

Furthermore, note that by the definition of  $\tilde{\lambda}_k$ , we have:

$$\frac{1}{M} \sum_{i=1, i \neq k}^{M+1} L_{\tilde{\lambda}}(x_i) + U^*(\tilde{\lambda}) \geq \frac{1}{M} \sum_{i=1, i \neq k}^{M+1} L_{\tilde{\lambda}_k}(x_i) + U^*(\tilde{\lambda}_k) \tag{44}$$

Combining the results of three inequalities in Equations (42), (43) and (44), we then obtain:

$$\begin{aligned} \frac{\kappa}{2} \|\tilde{\lambda} - \tilde{\lambda}_k\|_2^2 &\leq \frac{1}{M} \sum_{i=1, i \neq k}^{M+1} \nabla L_{\tilde{\lambda}}(x_i)^T (\tilde{\lambda} - \tilde{\lambda}_k) + U^*(\tilde{\lambda})^T (\tilde{\lambda} - \tilde{\lambda}_k) \\ &\leq \left\| \frac{1}{M} \sum_{i=1, i \neq k}^{M+1} \nabla L_{\tilde{\lambda}}(x_i) + U^*(\tilde{\lambda}) \right\| \|\tilde{\lambda} - \tilde{\lambda}_k\| \\ &= \frac{1}{M} |\nabla L_{\tilde{\lambda}}(x_k)| \|\tilde{\lambda} - \tilde{\lambda}_k\| \end{aligned}$$

The last equality follows from Equation (41). By canceling  $\|\tilde{\lambda} - \tilde{\lambda}_k\|$  from the above inequality, we obtained the desired bound. □

**Proof of Theorem 6.**

*Proof.* We use the same leave-one-out technique in [19,30,31]. It follows from Lemma 1 that:

$$\|\tilde{\lambda}_k - \tilde{\lambda}\| \leq \frac{1}{\kappa M} \|\nabla L_{\tilde{\lambda}}(x_k)\| \leq \frac{C}{\kappa M} \left(1 - \exp(-L_{\tilde{\lambda}}(x_k))\right)$$

Therefore:

$$L_{\tilde{\lambda}_k}(x_k) - L_{\tilde{\lambda}}(x_k) \leq \frac{C^2}{\kappa M} \left(1 - \exp(-L_{\tilde{\lambda}}(x_k))\right)$$

After summing over  $k$  from one to  $M + 1$ , then taking the expectation with respect to the training data and using Jensen’s inequality, we obtain:

$$\begin{aligned} E_{\tilde{x}} E_X L_{\lambda^\Delta}(X) &= E_{\tilde{x}_{M+1}} E_{x_k} L_{\tilde{\lambda}_k}(x_k) \\ &\leq E_{\tilde{x}_{M+1}} \frac{1}{M+1} \sum_{k=1}^{M+1} L_{\tilde{\lambda}}(x_k) + \frac{C^2}{\kappa M} \left(1 - E_{\tilde{x}_{M+1}} \frac{1}{M+1} \sum_{k=1}^{M+1} \exp(-L_{\tilde{\lambda}}(x_k))\right) \\ &\leq E_{\tilde{x}_{M+1}} \frac{1}{M+1} \sum_{k=1}^{M+1} L_{\tilde{\lambda}}(x_k) + \frac{C^2}{\kappa M} \left(1 - \exp\left(-E_{\tilde{x}_{M+1}} \frac{1}{M+1} \sum_{k=1}^{M+1} L_{\tilde{\lambda}}(x_k)\right)\right) \end{aligned} \tag{45}$$

Since  $U^*(\lambda)$  is a legal prior,  $U^*(\lambda) \geq 0$  and  $U^*(\hat{\lambda}) = 0$ , and since  $\tilde{\lambda}$  is the optimal solution of Equation (24), we have:

$$\begin{aligned} E_{\tilde{\lambda}_{M+1}} \frac{1}{M+1} \sum_{k=1}^{M+1} L_{\tilde{\lambda}}(x_k) &\leq \frac{M}{M+1} E_{\tilde{\lambda}_{M+1}} \left( \frac{1}{M} \sum_{k=1}^{M+1} L_{\tilde{\lambda}}(x_k) + U^*(\tilde{\lambda}) \right) \\ &\leq \frac{M}{M+1} E_{\tilde{\lambda}_{M+1}} \left( \frac{1}{M} \sum_{k=1}^{M+1} L_{\tilde{\lambda}}(x_k) + U^*(\hat{\lambda}) \right) = E_X L_{\tilde{\lambda}}(X) \end{aligned} \tag{46}$$

Combining the results of inequalities in Equations (45) and (46), yields:

$$\begin{aligned} E_{\tilde{\lambda}} D(p_{\tilde{\lambda}}(x) \| p_{\lambda^\Delta}(x)) &= E_{\tilde{\lambda}} D(p_0(x) \| p_{\lambda^\Delta}(x)) - D(p_0(x) \| p_{\tilde{\lambda}}(x)) \\ &= E_{\tilde{\lambda}} E_X L_{\lambda^\Delta}(X) - E_X L_{\tilde{\lambda}}(X) \leq \frac{C^2}{\kappa M} \left( 1 - \exp(-E_X L_{\tilde{\lambda}}(X)) \right) \end{aligned}$$

□

**Proof of Theorem 9.**

*Proof.* By working with  $p_\lambda(y)$  and using the same techniques as before, i.e., the McDiarmid concentration inequality [38] and symmetrization [33,39], we have with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} &\sup_{\lambda \in \Omega} \left| E_{p_0(y)} \log p_\lambda(y) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(y_j) \right| \tag{47} \\ &\leq 2E_{\tilde{y}, \omega} \sup_{\lambda \in \Omega, f(x) \in \mathcal{F}(x)} \left| \frac{1}{M} \sum_{j=1}^M \omega_j \log \mathcal{F}(y) \right| + C_2 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}} \end{aligned}$$

Now, we apply the refined version of the Rademacher comparison inequality proposed in Theorem 7 of Meir and Zhang [24], which says that for  $l$ -Lipschitz functions  $\phi_i : \mathfrak{R} \rightarrow \mathfrak{R}, i = 1, \dots, M$ , one obtains the inequality:

$$E_\omega \left( \sup_{f \in \mathcal{F}} \sum_{i=1}^M \omega_i \phi_i(f_i) \right) \leq l E_\omega \left( \sup_{f \in \mathcal{F}} \sum_{i=1}^M \omega_i f_i \right)$$

By the arguments in [20,33], it is easy to show that the absolute value version is valid as:

$$E_\omega \left( \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^M \omega_i \phi_i(f_i) \right| \right) \leq 2l E_\omega \left( \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^M \omega_i f_i \right| \right)$$

Let  $\phi(x) = \log(x)$ , where  $a \leq x \leq b$ . It is easy to verify that  $\phi(x)$  is  $\frac{1}{a}$ -Lipschitz, so:

$$E_{\tilde{y}, \omega} \sup_{\lambda \in \Omega} \left| \frac{1}{M} \sum_{j=1}^M \omega_j \log \mathcal{F}_\lambda(y_j) \right| \leq \frac{2}{a} E_{\tilde{y}, \omega} \sup_{\lambda \in \Omega} \left| \frac{1}{M} \sum_{j=1}^M \omega_j \mathcal{F}_\lambda(y_j) \right|$$

Combining this inequality with that of Dudley’s entropy integral, one can then establish that with a probability of at least  $1 - \eta$ ,

$$\begin{aligned} &\sup_{\lambda \in \Omega} \left| E_{p_0(y)} \log p_\lambda(y) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(y_j)(y_j) \right| \tag{48} \\ &\leq 2 \frac{C_3}{\sqrt{M}} \int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon + C_4 \sqrt{\frac{\log(\frac{1}{\eta})}{2M}} \end{aligned}$$

where  $0 < \zeta < \alpha < \infty$ . These two quantities depend on  $\sup_{\lambda \in \Omega} \|\lambda\|_\infty$ ,  $a$  and  $b$ . Furthermore,  $C_3$  is the constant in the bound of Rademacher averages over linear combinations in  $\mathcal{F}(y)$ , obtained by by Dudley’s entropy modified by  $a$  and  $b$ . Finally, the constant,  $C_4$ , depends on  $a, b$  and  $\eta$ .

Putting the pieces together, with a probability of at least  $1 - \eta$ , we obtain:

$$\begin{aligned} & D(p_0(y) \| p_{\lambda^*}(y)) - D(p_0(y) \| p_{\hat{\lambda}}(y)) \\ &= \left( E_{p_0(y)} \log p_{\hat{\lambda}}(y) - E_{\bar{p}(y)} \log p_{\hat{\lambda}}(y) \right) + \left( E_{\bar{p}(y)} \log p_{\lambda^*}(y) - E_{p_0(y)} \log p_{\lambda^*}(y) \right) \\ & \quad + \left( E_{\bar{p}(y)} \log p_{\hat{\lambda}}(y) - E_{\bar{p}(y)} \log p_{\lambda^*}(y) \right) \\ &\leq 2 \sup_{\lambda \in \Omega} \left| E_{p_0(y)} \log p_\lambda(y) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(y_j) \right| + \frac{1}{M} \sum_{j=1}^M \log \frac{p_{\hat{\lambda}}(y_j)}{p_{\lambda^*}(y_j)} \\ &\leq \frac{4C_3}{\sqrt{M}} E_{\bar{y}} \left[ \int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon \right] + C_4 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} \end{aligned}$$

□

**Proof of Theorem 10.**

*Proof.* Choose the function to be  $\log p_\lambda(y)$ ,  $\lambda \in \Omega$ . Then,  $\cosh \left( 2\tau \log p_\lambda(x) \right) = \frac{1}{2} \left( p_\lambda(y)^{2\tau} - \frac{1}{p_\lambda(y)^{2\tau}} \right)$ . By Theorem 3 in [24], we have that if there exists a positive number,  $K(\mathcal{F})$ , such that for all  $\tau > 0$ ,

$$\log E_{p_0(y)} \sup_{\lambda \in \Omega} \left( p_\lambda(y)^{2\tau} - \frac{1}{p_\lambda(y)^{2\tau}} \right) \leq \left( \tau K(\mathcal{F}) \right)^2$$

then, for all  $\lambda \in \Omega$  with a probability of at least  $1 - \eta$ , (41) will hold.

Next, we take the derivative of  $\left( p_\lambda(y)^{2\tau} - \frac{1}{p_\lambda(y)^{2\tau}} \right)$  with respect to  $\lambda$  and set this to zero. After some routine calculation, we obtain for  $i = 1 \dots N$  :

$$\left( p_\lambda(x)^{2\tau-1} - \frac{1}{p_\lambda(y)^{2\tau+1}} \right) p_\lambda(y) \left( E_{p_\lambda(x)} \left( f_i(x) \right) - E_{p_\lambda(z|y)} \left( f_i(y, z) \right) \right) = 0$$

Thus,  $\lambda^\bullet = \arg \sup_{\lambda \in \Omega} \left( p_\lambda(y)^{2\tau} - \frac{1}{p_\lambda(y)^{2\tau}} \right)$  are those parameters, such that they achieve  $E_{p_{\lambda^\bullet(x)}} \left( f(x) \right) = E_{p_{\lambda^\bullet(z|y)}} \left( f_i(y, z) \right)$ . Moreover, they achieve the maximum of  $\left( p_\lambda(y)^{2\tau} - \frac{1}{p_\lambda(y)^{2\tau}} \right)$ . Even though these parameters can be uniquely determined for each fixed  $y$ , unlike the complete data case, they may not be the same as the MLE or LME estimates, due to the existence of multiple feasible solutions. □

**Proof of Theorem 11.**

*Proof.* The coefficients are the same as in Theorem 8, and the proof is similar.

$$\begin{aligned} & D(p_0(y) \| p_{\lambda^\circ}(y)) - D(p_0(y) \| p_{\hat{\lambda}}(y)) \\ &= \left( E_{\bar{p}(y)} \log p_{\hat{\lambda}}(y) - E_{p_0(y)} \log p_{\hat{\lambda}}(y) \right) + \left( E_{\bar{p}(y)} \log p_{\lambda^\circ}(y) - E_{p_0(y)} \log p_{\lambda^\circ}(y) \right) \\ & \quad + \left( E_{\bar{p}(y)} \log p_{\hat{\lambda}}(y) - E_{\bar{p}(y)} \log p_{\lambda^\circ}(y) \right) \\ &\leq 2 \sup_{\lambda \in \Omega} \left| E_{p_0(y)} \log p_\lambda(y) - \frac{1}{M} \sum_{j=1}^M \log p_\lambda(y_j) \right| + \frac{1}{M} \sum_{j=1}^M \log \frac{p_{\hat{\lambda}}(y_j)}{p_{\lambda^\circ}(y_j)} \end{aligned}$$

Therefore, by using inequality in Equation (48), we have with a probability of at least  $1 - \eta$ ,

$$D(p_0(y)||p_{\lambda^*}(y)) - D(p_0(y)||p_{\hat{\lambda}}(y)) \leq \frac{4C_3}{\sqrt{M}} E_{\tilde{y}} \left[ \int_{\zeta}^{\alpha} \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)} d\epsilon \right] \\ + C_4 \sqrt{\frac{2 \log(\frac{1}{\eta})}{M}} + E_{\tilde{p}(y)} \log \frac{p_{\hat{\lambda}}(y)}{p_{\lambda^*}(y)}$$

□

## References

- Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
- Della Pietra, S.; Della Pietra, V.; Lafferty, J. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 380–393.
- Palmieri, F.; Ciunzo, D. Objective priors from maximum entropy in data classification. *Inf. Fusion* **2013**, *14*, 186–198.
- Ziebart, B.; Maas, A.; Bagnell, J.; Dey, A. Maximum Entropy Inverse Reinforcement Learning. In Proceedings of The 23rd National Conference on Artificial Intelligence (AAAI), Chicago, IL, USA, 13–17 July 2008; pp.1433–1438.
- Wang, S.; Schuurmans, D.; Zhao, Y. The latent maximum entropy principle. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, No. 8, 1–42.
- McLachlan, G.; Krishnan, T. *The EM Algorithm and Extensions*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2008.
- Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer-Verlag: New York, NY, USA, 1998.
- Vapnik, V.; Chervonenkis, A. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognit. Image Anal.* **1991**, *1*, 283–305.
- Vapnik, V.N. *Statistical Learning Theory*; Wiley-Interscience: Hoboken, NJ, USA, 1998.
- Barron, A.; Sheu, C. Approximation of density functions by sequences of exponential families. *Ann. Stat.* **1991**, *19*, 1347–1369.
- Dudik, M.; Phillips, S.; Schapire, R. Performance Guarantees for Regularized Maximum Entropy Density Estimation. In Proceedings of The 17th Annual Conference on Learning Theory (COLT), Banff, Canada, 1–4 July 2004; pp. 472–486.
- Van de Geer, S. *Empirical Processes in M-Estimation*; Cambridge University Press: Cambridge, UK, 2000.
- Jelinek, F. *Statistical Methods for Speech Recognition*; MIT Press: Cambridge, MA, USA, 1998.
- Rosenfeld, R. A maximum entropy approach to adaptive statistical language modeling. *Comput. Speech Lang.* **1996**, *10*, 187–228.
- Koltchinskii, V.; Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Stat.* **2002**, *30*, 1–50.
- Rakhlin, A.; Panchenko, D.; Mukherjee, S. Risk bounds for mixture density estimation. *ESAIM: Probab. Stat.* **2005**, *9*, 220–229.
- Csiszar, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158.

18. Barron, A. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **1994**, *14*, 115–133.
19. Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.* **2004**, *32*, 56–85.
20. Panchenko, D. *Class Lecture Notes of MIT 18.465: Statistical Learning Theory*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2002.
21. Zhang, T. Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res.* **2002**, *2*, 527–550.
22. Dudley, R. *Uniform Central Limit Theorems*; Cambridge University Press: New York, NY, USA, 1999.
23. Wainwright, M.; Jordan, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **2008**, *1*, 1–305 .
24. Meir, R.; Zhang, T. Generalization error bounds for Bayesian mixture algorithms. *J. Mach. Learn. Res.* **2003**, *4*, 839–860.
25. Chen, S.F.; Rosenfeld, R. A survey of smoothing techniques for ME models. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 37–50.
26. Borwein, J.; Lewis, A. *Convex Analysis and Nonlinear Optimization: Theory and Examples*; Springer-Verlag: New York, NY, USA, 2000.
27. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
28. Lebanon, G.; Lafferty, J. Boosting and maximum likelihood for exponential models. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 447–454.
29. Wang, S.; Schuurmans, D.; Peng, F.; Zhao, Y. Learning mixture models with the regularized latent maximum entropy principle. *IEEE Trans. Neural Netw.* **2004**, *15*, 903–916.
30. Zhang, T. Leave-one-out bounds for kernel methods. *Neural Comput.* **2003**, *15*, 1397–1437.
31. Zhang, T. Class-size independent generalization analysis of some discriminative multi-category classification methods. *Adv. Neural Inf. Process. Syst.* **2004**, *17*, 1625–1632.
32. Wang, S.; Schuurmans, D.; Peng, F.; Zhao, Y. Combining statistical language models via the latent maximum entropy principle. *Mach. Learn. J.* **2005**, *60*, 229–250.
33. Ledoux, M.; Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag: Berlin and Heidelberg, Germany, 1991.
34. Opper, M.; Haussler, D. Worst Case Prediction over Sequences under Log Loss. In *The Mathematics of Information Coding, Extraction and Distribution*; Cybenko, G., O’Leary, D., Rissanen, J., Eds; Springer-Verlag: New York, NY, USA, 1998.
35. Lafferty, J.; Della Pietra, V.; Della Pietra, S. Statistical Learning Algorithms Based on Bregman Distance. In Proceedings of the Canadian Workshop on Information Theory, Toronto, Ontario, Canada, 3–6 June 1997; pp. 77–80.
36. Wang, S.; Schuurmans, D. Learning Continuous Latent Variable Models with Bregman Divergences. In Proceedings of The 14th International Conference on Algorithmic Learning Theory (ALT), Sapporo, Japan, 17–19 October 2003; pp. 190–204.

37. Bartlett, P. *Class Lecture Notes of UC-Berkeley CS281B/Stat241B: Statistical Learning Theory*; University of California, Berkeley: Berkeley, CA, USA, 2003.
38. McDiarmid, C. On the Method of Bounded Differences. In *Surveys in Combinatorics*; Siemons, J., Ed.; Cambridge University Press: Cambridge, UK, 1989; pp. 148–188.
39. Van der Vaart, A.; Wellner, J. *Weak Convergence and Empirical Processes*; Springer-Verlag: New York, NY, USA, 1996.
40. Zhang, T.; Yu, B. Boosting with early stopping: Convergence and consistency. *Ann. Stat.* **2005**, *33*, 1538–1579.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).