Sample Complexity of Composite Likelihood

Joseph K. Bradley **Carlos Guestrin**
Carnegie Mellon University

Abstract

We present the first PAC bounds for learning parameters of Conditional Random Fields [12] with general structures over discrete and real-valued variables. Our bounds apply to composite likelihood [14], which generalizes maximum likelihood and pseudolikelihood [3]. Moreover, we show that the only existing algorithm with a PAC bound for learning high-treewidth discrete models [1] can be viewed as a computationally inefficient method for computing pseudolikelihood. We present an extensive empirical study of the statistical efficiency of these estimators, as predicted by our bounds. Finally, we use our bounds to show how to construct computationally and statistically efficient composite likelihood estimators.

1 Introduction

Markov Randoms Fields (MRFs) (c.f., [11]) and Conditional Random Fields (CRFs) [12] are models of distributions over random variables and are commonly used in machine learning, natural language processing, and other domains. MRFs and CRFs encode distributions by using conditional independence structure which permits simpler representations. Unfortunately, the task of *inference* (i.e., computing probabilistic queries over variables) in these models is #P-hard (and NP-hard to approximate) in general [16].

The difficulty of inference poses a particular problem for *parameter learning*, i.e., estimating the values in a distribution given its structure and a training dataset. A popular method for parameter learning is *maximum likelihood estimation (MLE)*, which maximizes the expected likelihood of the data. However, the MLE optimization problem requires inference and so is computationally intractable in general. When approximate inference is used, MLE generally loses its statistical guarantees. Some methods, such as *maximum pseudo-likelihood estimation (MPLE)* [3], estimate parameters without intractable inference (i.e., using only tractable probabilistic queries) but—before our work—only had asymptotic guarantees [13].

We present the first strong finite sample guarantees for learning the parameters of general CRFs (which generalize MRFs) when the target distribution is in our model class. Specifically, we analyze *maximum composite likelihood estimation (MCLE)* [14], a generalization of MLE and MPLE which permits learning without intractable inference. We prove learning bounds for MCLE w.r.t. the parameter estimation error and log loss within the probably approximately correct (PAC) framework [21]: we can achieve high accuracy with polynomial sample and computational complexity. In the large sample limit, our bounds match existing asymptotic normality results for MCLE [13].

To our knowledge, only one other method for learning parameters of high-treewidth discrete models has PAC guarantees [1]. We prove that the method actually computes MPLE. Thus, our analysis covers the only existing methods for PAC-learning CRF parameters.

We end with a detailed empirical analysis of our theoretical results. We show that our bounds accurately capture MCLE behavior in terms of a few problem properties, and we study how those properties vary with model structure and parameters. Finally, we improve upon the traditional use of MCLE by showing that careful choice of the MCLE loss structure can provide computationally efficient estimators with better statistical properties and empirical performance.

2 Related Work

Many works have shown MPLE to be empirically successful (c.f., [18–20]), with lower statistical but higher computational efficiency than MLE. Theoretical analyses have predicted such behavior (c.f., [6, 10, 13]), but only in the asymptotic setting.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Our analysis is most closely related to that of [15], who proved sample complexity bounds for parameter estimation error in regression problems $Y_i \sim X$ in Ising models. We adapt their analysis to handle general log linear CRFs, structured losses, and shared parameters (Sec. 4.4), and we use the resulting bounds on parameter estimation error to prove bounds on the log loss. Our bounds resemble asymptotic normality results such as those of [13] and [4].

Other methods for learning without intractable inference include piecewise likelihood and piecewise pseudolikelihood [19, 20], as well as stochastic composite likelihood (which generalizes MCLE) [4]. These methods are often empirically successful but, to our knowledge, do not have PAC guarantees. The one exception is [1], who use a (complicated) technique dubbed the canonical parametrization to PAC-learn bounded-degree factor graphs over discrete variables. In Sec. 5, we show that their algorithm in fact computes MPLE.

Intractable inference in learning may be replaced with approximate inference methods such as sampling and variational inference (c.f., [7, 11, 22]). In general, sampling lacks finite-time guarantees on the approximation accuracy of inference results. Variational inference sometimes includes problem-specific guarantees computable at runtime, but not a priori PAC bounds.

3 Learning CRF Parameters

We write general log-linear CRFs in the form

$$P_\theta(Y|X) = \frac{1}{Z(X;\theta)} \exp\left(\theta^T \phi(Y,X)\right), \tag{1}$$

where Y and X are sets of *output* and *input* variables, respectively. Y and X may be discrete or real-valued. θ is an r-vector of *parameters*; ϕ is an r-vector of *features* which are functions of Y and X with bounded range; and Z is the *partition function*. Note that an MRF $P(Y)$ would be represented by letting $X = \emptyset$.

The feature vector ϕ implicitly defines the structure of the CRF: each element ϕ_t is a function of some subset of Y, X and defines edges in the CRF graph connecting its arguments. We call each $\theta_t^T \phi_t(Y,X)$ a *factor*.

In *parameter learning*, our goal is to estimate θ from a set of i.i.d. samples $\{y^{(i)}, x^{(i)}\}$ from a *target distribution* $P(X)P_{\theta^*}(Y|X)$, where θ^* are the *target parameters*. Note we assume $P_{\theta^*}(Y|X)$ is in our model family in Eq. (1). We learn parameters by minimizing a *loss* $\hat{\ell}$, which is a function of the samples and θ, plus a regularization term:

$$\min_\theta \hat{\ell}(\theta) + \lambda \|\theta\|_p. \tag{2}$$

Above, $\lambda \geq 0$ is a regularization parameter, and $p \in \{1, 2\}$ specifies the L_1 or L_2 norm. We write $\hat{\ell}$ for the loss computed w.r.t. the n training samples and ℓ for the loss w.r.t. the target distribution.

The choice of loss function defines the estimator. MLE minimizes the *log loss*:

$$\ell_L(\theta) = \mathbf{E}_{P(X)P_{\theta^*}(Y|X)} \left[-\log P_\theta(Y|X)\right]. \tag{3}$$

MPLE [3] minimizes the *pseudolikelihood loss*, which is the sum over variables Y_i of their likelihoods conditioned on neighbors in $Y_{\setminus i} \doteq Y \setminus \{Y_i\}$ and in X:

$$\ell_{PL}(\theta) = \mathbf{E}_{P(X)P_{\theta^*}(Y|X)} \left[-\sum_i \log P_\theta(Y_i|Y_{\setminus i}, X)\right]. \tag{4}$$

In general, computing the conditional probabilities $P(A|\cdot)$ in these losses takes time exponential in $|A|$. Thus, computing the log loss can take time exponential in $|Y|$, while computing the pseudolikelihood loss takes time linear in $|Y|$.

MCLE [14] minimizes the *composite likelihood loss*:

$$\ell_{CL}(\theta) = \mathbf{E}_{P(X)P_{\theta^*}(Y|X)} \left[-\sum_i \log P_\theta(Y_{A_i}|Y_{\setminus A_i}, X)\right], \tag{5}$$

where Y_{A_i} is a set of variables defining the i^{th} *likelihood component* and $Y_{\setminus A_i} \doteq Y \setminus Y_{A_i}$. With a single component $Y_{A_i} = Y$, MCLE reduces to MLE; if each component is a single variable $Y_{A_i} = \{Y_i\}$, MCLE reduces to MPLE. MCLE thus permits a range of losses between MLE and MPLE with varying computational complexity. We discuss the choice of likelihood components in Sec. 7.

If a feature ϕ_t is a function of $Y_i \in Y_{A_i}$, then we say ϕ_t and θ_t *participate* in component A_i, and we write $\theta_t \in \theta_{A_i}$. Also, some works permit more general components conditioned on $Y_B \subseteq Y_{\setminus A_i}$. We restrict $Y_B = Y_{\setminus A_i}$ for simplicity, and our analysis only requires that the set of components $\mathcal{A} \doteq \{A_i\}$ forms a consistent estimator (i.e., an estimator which recovers the target parameters with probability approaching 1 as the training set size approaches infinity).

4 Sample Complexity Bounds

This section presents our main theoretical results: PAC bounds for learning parameters for general CRFs using MCLE. We give bounds in terms of parameter estimation error, which we then use to bound log loss. The appendix has detailed proofs of all results.

We refer to our bounds as PAC bounds, although the PAC framework requires learning to be polytime. MLE and MCLE may or may not be polytime; our bounds are thus statistical statements. We later discuss choosing tractable MCLE estimators, for which our bounds are technically PAC bounds.

4.1 Parameter Estimation: MLE

For comparison, we first give bounds for MLE. Our bounds are written in terms of a few model properties: r (the number of parameters), $\phi_{max} \doteq \max_{t,y,x} \phi_t(y,x)$ (the maximum magnitude of any feature vector element), and C_{min} (a lower bound on $\Lambda_{min}\left(\nabla^2 \ell_L(\theta^*)\right)$, the minimum eigenvalue of the Hessian of the log loss w.r.t. the target distribution).[1] Bounding the eigenvalues away from 0 prevents variable interactions from being deterministic.

Our first theorem is a PAC bound for MLE w.r.t. parameter estimation error.

Theorem 4.1. ***(MLE PAC bound)*** *Assume a lower bound on the minimum eigenvalue of the log loss: $\Lambda_{min}\left(\nabla^2 \ell_L(\theta^*)\right) \geq C_{min} > 0$. Suppose we learn parameters $\hat{\theta}$ by minimizing Eq. (2) using the log loss ℓ_L w.r.t. $n > 1$ i.i.d. samples, using regularization $\lambda = \frac{C_{min}^2}{2^6 r^2 \phi_{max}^3} n^{-\xi/2}$, where $\xi \in (0,1)$. Then $\hat{\theta}$ will be close to the target parameter vector θ^*:*

$$\left\| \hat{\theta} - \theta^* \right\|_1 \leq \frac{C_{min}}{4 r \phi_{max}^3} n^{-\xi/2} \tag{6}$$

with probability at least

$$1 - 2r(r+1)\exp\left(-\frac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi}\right). \tag{7}$$

In Theorem 4.1, the constant ξ trades off the convergence rate of the parameters with the probability of success. As $n \longrightarrow \infty$, we may let $\xi \longrightarrow 1$ while keeping the probability of success high; as $\xi \longrightarrow 1$, the convergence rate approaches $n^{-1/2}$, the asymptotic rate [13]. The next corollary eliminates ξ by converting the PAC bound into a sample complexity bound.

Corollary 4.2. ***(MLE sample complexity)*** *Given the assumptions from Theorem 4.1, to estimate the parameters within L_1 error ϵ with probability at least $1-\delta$, it suffices to have a training set of size*

$$n \geq \frac{2^9 \phi_{max}^2}{C_{min}^2} \frac{1}{(\epsilon/r)^2} \log \frac{2r(r+1)}{\delta}. \tag{8}$$

This sample complexity result implies that parameters are easier to learn when the minimum eigenvalue bound C_{min} is large; i.e., estimators with large C_{min} are more *statistically efficient*.[2] Asymptotic results (e.g., [13, 15]) have related statistical efficiency to the loss' Hessian. The above bound is expressed in terms of ϵ/r, the error for the full parameter vector normalized by the number of parameters r; keeping this normalized parameter error ϵ/r constant, the bound increases only logarithmically with r. Also, note that changing ϕ_{max} essentially rescales parameters.

[1] If the features are overcomplete, then r is the intrinsic dimensionality of the feature space, computed as the number of non-zero eigenvalues of the log loss Hessian. C_{min} is a lower bound on the non-zero eigenvalues.

[2] We use "statistical efficiency" w.r.t. finite sample sizes, borrowing the term from asymptotic analysis.

4.2 Parameter Estimation: MCLE

We now present bounds generalized to MCLE. Since MCLE can use multiple likelihood components A_i, our bounds contain additional quantities. M_{max} denotes the maximum number of components in which any feature participates. The bound C_{min} applies to all $A_i \in \mathcal{A}$. We also write ρ_{min} as a lower bound on the sum of minimum eigenvalues for likelihood components (w.r.t. the target distribution) in which any parameter θ_t participates: $\rho_{min} \doteq \min_t \rho_t$, where $\rho_t \leq \sum_{i\,:\,\theta_t \in \theta_{A_i}} \Lambda_{min}(\nabla^2 [\ell_{CL}(\theta^*)]_{A_i})$. Here, $[\ell_{CL}(\theta^*)]_{A_i}$ denotes the loss term contributed by component A_i.

Intuitively, ρ_{min} generalizes C_{min} from MLE to MCLE. Recall that MLE uses a single likelihood component to estimate θ, and the minimum eigenvalue of the component's Hessian (C_{min}) affects the statistical efficiency of MLE. In MCLE, each parameter may be estimated using multiple likelihood components, and the sum of the minimum eigenvalues for those components (ρ_{min}) affects the statistical efficiency of MCLE.

We can now present our PAC bound for MCLE.

Theorem 4.3. ***(MCLE PAC bound)*** *Assume we use a consistent MCLE estimator ℓ_{CL}. Assume bounds $\min_i \Lambda_{min}\left(\nabla^2 [\ell_{CL}(\theta^*)]_{A_i}\right) \geq C_{min} > 0$, and let $\rho_{min} = \min_t \rho_t$. Suppose we learn parameters $\hat{\theta}$ by minimizing Eq. (2) using the MCLE loss ℓ_{CL} w.r.t. $n > 1$ i.i.d. samples, using regularization $\lambda = \frac{C_{min}^2}{2^6 r^2 M_{max} \phi_{max}^3} n^{-\xi/2}$, where $\xi \in (0,1)$. Then $\hat{\theta}$ will be close to the target parameter vector θ^*:*

$$\left\| \hat{\theta} - \theta^* \right\|_1 \leq \frac{\rho_{min}}{4 r M_{max} \phi_{max}^3} n^{-\xi/2} \tag{9}$$

with probability at least

$$1 - 2r(|\mathcal{A}|r+1)\exp\left(-\frac{C_{min}^4}{2^{13} r^4 M_{max}^4 \phi_{max}^8} n^{1-\xi}\right). \tag{10}$$

If the number of likelihood components is

$$|\mathcal{A}| \leq \frac{1}{2r^2}(2r)^{\left[\frac{2^8 C_{min}^2 M_{max}^2}{\rho_{min}^2}\right]}, \tag{11}$$

then Eq. (9) holds with probability at least

$$1 - 4r\exp\left(-\frac{\rho_{min}^4 n^{1-\xi}}{2^{13} r^4 M_{max}^4 \phi_{max}^8}\right). \tag{12}$$

We can see that using multiple likelihood components can worsen the bound by increasing M_{max} and $|\mathcal{A}|$

but can also improve the bound by increasing ρ_{min}. Sec. 7 shows how careful choices of components can improve this tradeoff. We include the special condition Eq. (11) since it permits us to replace C_{min} with ρ_{min} in the probability bound, which will later prove helpful in explaining the behavior of MCLE with overlapping components. Eq. (11) is a reasonable requirement in many cases, and it holds for our empirical tests.

The following corollary converts Theorem 4.3 into a sample complexity bound.

Corollary 4.4. ***(MCLE sample complexity)*** *Given the assumptions from Theorem 4.3, to estimate the parameters within L_1 error ϵ with probability at least $1-\delta$, it suffices to have a training set of size*

$$n \geq \frac{2^9 M_{max}^2 \phi_{max}^2 \rho_{min}^2}{C_{min}^4} \frac{1}{(\epsilon/r)^2} \log \frac{2r(|\mathcal{A}|r+1)}{\delta}. \quad (13)$$

If Eq. (11) holds, then it suffices to have

$$n \geq \frac{2^9 M_{max}^2 \phi_{max}^2}{\rho_{min}^2} \frac{1}{(\epsilon/r)^2} \log \frac{4r}{\delta}. \quad (14)$$

Theorem 4.3 generalizes Theorem 4.1: with MLE, $\rho_{min} = C_{min}$, $M_{max} = 1$, and $|\mathcal{A}| = 1$. For MPLE, we have $\rho_{min} \geq C_{min}$, $M_{max} = 2$, and $|\mathcal{A}| = |Y|$. Thus, MLE's statistical guarantees are stronger than those for MPLE and MCLE only up to problem-dependent constants. All three estimators' sample complexity bounds have the same dependence (up to log terms) on the problem properties r and ϕ_{max}, the desired error ϵ, and the probability of failure δ.

The estimators mainly differ in their spectral properties. Recall that ρ_{min} generalizes C_{min}. If each parameter θ_t participates in an equal number of MCLE components, then we may replace M_{max}/ρ_{min} with the minimum over parameters θ_t of the average of minimum eigenvalues for components in which θ_t participates: $\overline{\rho}_{min} \doteq \min_t \operatorname{avg}_{i\,:\,\theta_t \in \theta_{A_i}} \Lambda_{min}(\nabla^2[\ell_{CL}(\theta^*)]_{A_i})$. This substitution makes our MCLE sample complexity bound in Eq. (14) identical (up to log factors) to our MLE bound in Eq. (8), with $\overline{\rho}_{min}$ substituted for C_{min}. I.e., MCLE averages the effects of its various components. We show in Sec. 7 how this averaging can mitigate the negative impact of "bad" components (with small eigenvalues).

4.3 Loss Reduction

Thus far, we have only given bounds on parameter estimation error. We now show that a bound on parameter error may be used to bound the log loss.

Theorem 4.5. ***(Log loss, given parameter error)*** *Let Λ_{max} be the largest eigenvalue of the log loss Hessian at θ^*. If the parameter estimation error is small:*

$$\|\theta - \theta^*\|_1 \leq \frac{-\Lambda_{max} + \sqrt{\Lambda_{max}^2 + 4r\phi_{max}^4}}{4\phi_{max}^3}, \quad (15)$$

the log loss converges quadratically in the error:

$$\ell_L(\theta) \leq \ell_L(\theta^*) + \left(\tfrac{\Lambda_{max}}{2} + \phi_{max}^2\right) \|\theta - \theta^*\|_1^2. \quad (16)$$

Otherwise, the log loss converges linearly in the error:

$$\ell_L(\theta) \leq \ell_L(\theta^*) + \phi_{max} \|\theta - \theta^*\|_1. \quad (17)$$

This theorem describes two well-known convergence regimes: linear far from the optimum and quadratic close to the optimum. In the large sample limit, our results indicate that the log loss of the MLE and MCLE estimates converge at a rate approaching n^{-1}, matching the asymptotic results of [13]. To see this, let $\xi \longrightarrow 1$ in Theorem 4.1 and Theorem 4.3, and note that we enter the quadratic regime in Theorem 4.5.

Sample complexity bounds for MLE and MCLE w.r.t. log loss may be computed by combining Corollary 4.2 and Corollary 4.4 with Theorem 4.5. For lack of space, we relegate these bounds to the appendix.

4.4 Disjoint vs. Joint Optimization

In Sec. 4.2, we analyzed MCLE using *joint optimization*; i.e., we jointly minimized all likelihood components' contributions to the loss in Eq. (5), and parameters θ_t participating in multiple components were shared during optimization. In this section, we discuss *disjoint optimization*, in which each likelihood component is treated as a separate MLE regression problem over a subset of the variables.

With disjoint optimization, each component A_i produces an estimate of its parameter subvector; denote the estimate of each $\theta_t \in \theta_{A_i}$ by $\hat{\theta}_t^{(A_i)}$. These estimates obey the bounds for MLE in Sec. 4.1. We can obtain a global estimate $\hat{\theta}$ of the parameters by averaging these subvectors where they overlap:

$$\hat{\theta}_t = \operatorname{avg}_{i\,:\,\theta_t \in \theta_{A_i}} \hat{\theta}_t^{(A_i)}. \quad (18)$$

Disjoint optimization is simple to implement and is *data parallel*, permitting easy scaling via parallel computing (c.f., [5]). We now show how to bound the error in $\hat{\theta}$ using the bounds from each component's estimate.

Lemma 4.6. ***(Disjoint optimization)*** *Suppose we average the results of disjoint optimizations using likelihood components $\mathcal{A}$, as in Eq. (18). If each estimated subvector $\hat{\theta}^{(A_i)}$ has L_1 error at most ϵ, then the full estimate has error $\left\|\hat{\theta} - \theta^*\right\|_1 \leq |\mathcal{A}|\epsilon$.*

The factor of $|\mathcal{A}|$ appears since the estimation error in each $\hat{\theta}^{(A_i)}$ could be in elements of θ which participate only in likelihood component A_i.

Theorem 4.7. ***(Disjoint MCLE sample complexity)*** *Suppose we average the results of disjoint*

optimizations using likelihood components $\mathcal{A}$, as in Eq. (18). Given the assumptions from Theorem 4.1 for each component, with a single C_{min} denoting a bound for all components, in order to estimate the parameters θ within L_1 error ϵ with probability at least $1-\delta$, it suffices to have a training set of size

$$n \geq \frac{2^9 \phi_{max}^2}{C_{min}^2} \frac{|\mathcal{A}|^2}{(\epsilon/r)^2} \log \frac{2r(r+1)|\mathcal{A}|}{\delta}. \quad (19)$$

Since $M_{max} \leq |\mathcal{A}|$ and $\rho_{min} \geq C_{min}$, this sample complexity bound for disjoint MCLE is worse than that for joint MCLE in Corollary 4.4, as might be expected.

5 Canonical Parametrization of Abbeel et al. (2006)

To our knowledge, [1] is the only previous work with PAC bounds for learning parameters of high-treewidth discrete MRFs. They analyzed MRFs representable as bounded-degree factor graphs (i.e., for which each ϕ_t is a function of a bounded number of variables, and each variable Y_i is an argument of a bounded number of features ϕ_t). They proposed learning via a *canonical parametrization*, which re-parametrizes an MRF in terms of products of *canonical factors*. Each canonical factor is a conditional probability $P(Y_A|Y_B)$, where Y_A, Y_B are of polynomial size. Learning thus consists of estimating each canonical factor from data and then multiplying them together. We describe the canonical parametrization in the appendix.

[17] later improved upon this method, showing how to simplify the canonical factors s.t. $|Y_A| = 1$. We use this insight to prove the canonical parametrization is, in fact, an alternative method for computing MPLE.

Theorem 5.1. *If canonical factors are computed using the factorization of the target distribution P_{θ^*}, then the canonical parametrizations of [1] and [17] produce the same parameter estimate $\hat{\theta}$ as MPLE.*

Recall that MPLE for MRFs essentially involves computing a conditional probability $P(Y_i|Y_{\setminus i})$ for each $Y_i \in Y$. Even with the improvement from [17], the canonical parametrization requires computing, for each factor ϕ_t, a set of conditional probabilities (canonical factors) exponential in the number of arguments of ϕ_t. (Specifically, for each factor ϕ_t, for each subset A of arguments of ϕ_t, and for each subset of A, we compute a conditional probability.) Directly computing MPLE via traditional optimization involves much less computation.

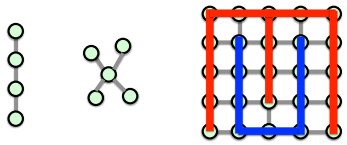

Figure 1: **Structures tested.** Left-to-right: chain, star, grid. The grid shows an example of two structured likelihood components ("combs"), as in Sec. 7.

6 Empirical Analysis of Bounds

We present an extensive study of our PAC bounds for MLE and MPLE on a variety of synthetic models. Our results show that our bounds accurately capture the learning methods' behaviors in terms of properties of the target distribution. We show how those properties vary across different model structures and factor types, indicating where MPLE may succeed or fail.

6.1 Setup

We tested synthetic models over binary variables, with features defined by edge factors $\phi(Y_i, Y_j)$ and $\phi(Y_i|X_i)$. We used three structures: chains, stars, and grids (Fig. 1). Our models had $|Y| = |X|$, and both $P(X)$ and $P(Y|X)$ shared the same structure. Our models used two factor types: associative (in which $\theta_t^* \phi_t(a,b) = s$ if $a = b$ and 0 otherwise) and random (in which each value $\theta_t^* \phi_t(\cdot,\cdot)$ was chosen uniformly at random from the range $[-s, s]$). We call s the factor strength, and we write associative(s) and random(s) for shorthand. Note the strength s is in log-space.

For optimization, we used conjugate gradient with exact inference for small models in this section and stochastic gradient with approximate inference (Gibbs sampling) for the large models in Sec. 7. We chose regularization λ according to each method's PAC bound, with $\xi = .5$ (though this choice is technically not valid for small training set sizes).[3] Our results are averaged over 10 runs on different random samples, and 10 models when using random factors.

6.2 Comparing Bounds

Our theoretical analysis included two types of bounds: a bound on the parameter estimation error $\|\hat{\theta} - \theta^*\|_1$ in terms of the training set size (Corollary 4.2, Corollary 4.4, and Theorem 4.7 for MLE, MPLE, and MPLE-disjoint, respectively) and a bound on the log

[3] We also ran experiments with regularization chosen via k-fold cross validation, which improved results but did not significantly change qualitative comparisons. We omit these results since they do not apply to our PAC analysis.

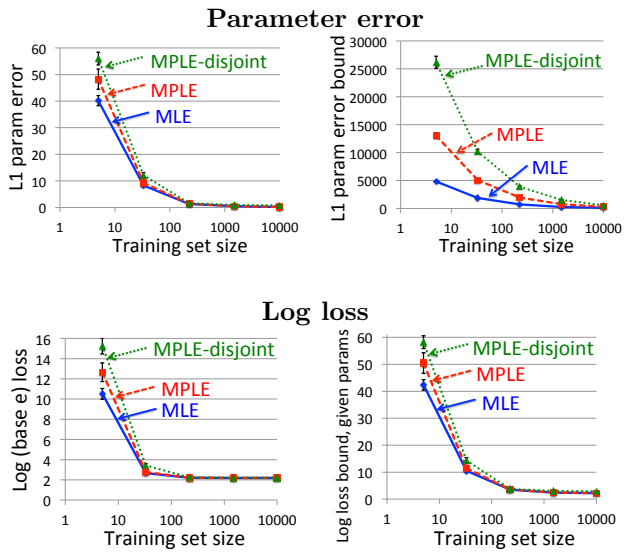

Figure 2: **Bounds: training set size.** *Top*: Actual (learned) parameter error (top-left) is much smaller than the bound on parameter error (top-right), but both decrease at similar rates w.r.t. training set size. *Bottom*: The actual log loss (bottom-left) is close to the loss bound given the actual parameter error (bottom-right). (Note scales on y-axes.)
Chains; $|Y| = 4$; random(1) factors. Averaged over 10 models $\times$ 10 datasets; error bars 1 stddev.

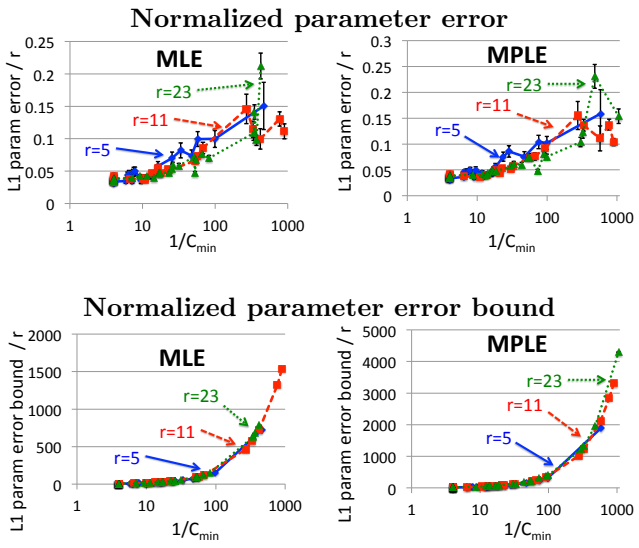

Figure 3: **Bounds: minimum eigenvalue C_{min}.** *Top*: Learned parameter error for MLE and MPLE (normalized by dimension r) increases with $1/C_{min}$. *Bottom*: Bound on parameter error for MLE and MPLE increases at a similar rate with $1/C_{min}$.
Chains; $|Y| = 2, 4, 8$ ($r = 5, 11, 23$); random(1) factors; 1495 training samples. Each point corresponds to 1 model, averaged 10 datasets; error bars 1 stddev.

loss $\ell_L(\hat{\theta})$ in terms of $\|\hat{\theta} - \theta^*\|_1$ (Theorem 4.5). Fig. 2 shows that the parameter error bound is much looser than the loss bound for MLE and MPLE with both joint and disjoint optimization. However, both bounds capture the convergence behavior w.r.t. training set size. As expected from our analysis, MPLE performs worse with disjoint optimization than with joint.

Our bounds for MLE and MPLE depend on a key property: C_{min}. Though C_{min} only needs to lower-bound the Hessian eigenvalues for our analysis, we simplify our discussion from here on by equating C_{min} with the minimum eigenvalue. Our bounds indicate that, for a fixed training set size, the parameter estimation error should be proportional to $1/C_{min}$. Fig. 3 plots the error for MLE and MPLE vs. their respective values $1/C_{min}$. Though the bound constants are loose, the $1/C_{min}$ dependence appears accurate.

Our bounds indicate that, for a fixed training set size, the normalized parameter error (normalized by the dimensionality r) should increase only logarithmically in r. Fig. 3 plots results for three values of r; increasing r does not significantly effect the normalized error.

This section's results were for chains with random factors, but other model types showed similar behavior. We next study a much wider variety of models.

6.3 Eigenspectra

As shown in the previous subsection, C_{min}, the minimum eigenvalue of the loss' Hessian, largely determines learning performance. When choosing a loss, we must trade off the goals of maximizing C_{min} (i.e., maximizing statistical efficiency) and of limiting computational complexity. In this section, we compare the C_{min} for MLE with the C_{min} for MPLE on a range of models, thus offering the reader guidance for when MPLE may replace MLE without sacrificing too much statistical efficiency.

Testing model diameter (chains): Fig. 4 plots the MLE/MPLE ratio of C_{min} values for chains with associative and random factors. Higher ratios imply lower statistical efficiency for MPLE, relative to MLE. For both, the ratio remains fairly constant as model size increases; i.e., model size does not significantly affect MPLE's relative statistical efficiency. However, increased factor strength decreases MPLE's efficiency, particularly for associative factors.

Testing node degree (stars): Fig. 5 compares MLE and MPLE for stars. For both associative and random factors, the C_{min} ratio increases as $|Y|$ increases, indicating that high-degree nodes decrease MPLE's relative statistical efficiency. As with chains, increased factor strength decreases MPLE's efficiency.

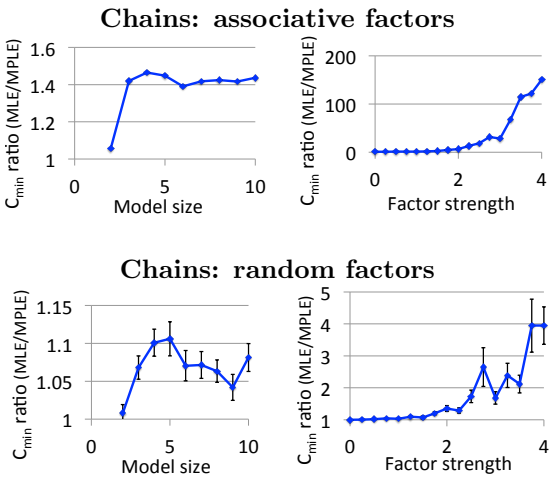

Figure 4: **Chains: Min eigval ratio MLE/MPLE.** Higher ratios imply MLE is superior.
Left: The ratio is about constant w.r.t. model size (for fixed factor strength 1). *Right*: The ratio increases with factor strength (for fixed model size $|Y| = 8$).
Note: Non-monotonicity in right-most parts of plots is from systematic numerical inaccuracy. Error bars 1 stddev, computed from 100 random models.

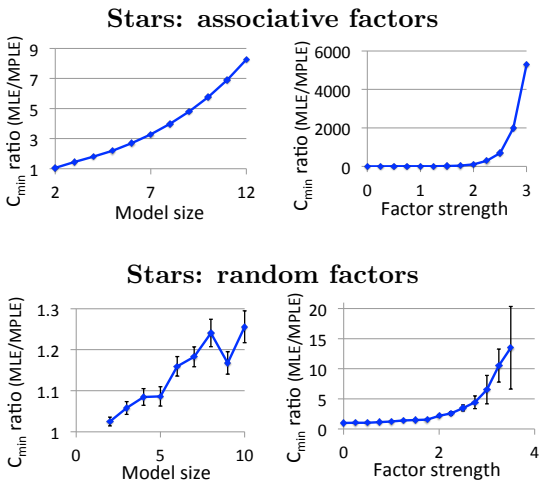

Figure 5: **Stars: Min eigval ratio MLE/MPLE.** Higher ratios imply MLE is superior.
Left: The ratio increases with model size (for fixed factor strength 1). *Right*: The ratio increases with factor strength (for fixed model size $|Y| = 8$). Error bars 1 stddev, computed from 100 random models.

Testing treewidth (grids): Fig. 6 compares MLE and MPLE for square grids (as in Fig. 1). For both factor types, the C_{min} ratio increases as $|Y|$ and factor strength increase. We estimated Hessians for grids with width > 3 by sampling from $P(X)$.

Overall, MPLE appears most statistically efficient for low-degree models with weak variable interactions. In the next section, we discuss how to overcome difficulties with high degree nodes and strong factors by using structured MCLE instead of MPLE.

7 Structured Composite Likelihood

MPLE sacrifices statistical efficiency for computational tractability. In this section, we show how to use MCLE to improve upon MPLE's statistical efficiency without much increase in computation. In particular, we demonstrate the benefits of careful selection of structured likelihood components for MCLE.

We state two propositions providing a simple method for choosing MCLE components. The first states how to choose a consistent MCLE estimator. The second states that consistent MCLE estimators may be combined to create new, consistent MCLE estimators.

Proposition 7.1. *Suppose a set of MCLE components $\mathcal{A}$ covers each Y variable exactly once; i.e., $\cup_i A_i = Y$, and $\sum_i |A_i| = |Y|$. Then the MCLE estimator defined by $\mathcal{A}$ is consistent.*

Proposition 7.2. *Suppose two MCLE estimators $\mathcal{A}'$ and $\mathcal{A}''$ are both consistent. Then their union $\mathcal{A} = \mathcal{A}' \cup \mathcal{A}''$ is also a consistent MCLE estimator.*

The simplest MCLE estimator buildable using Prop. 7.1 is MPLE. We advocate the use of *structured* likelihood components, i.e., components A_i containing multiple variables chosen according to the structure of the model. We give a simple example in Fig. 1, in which two *comb*-like components cover the entire model while maintaining low treewidth (i.e., tractable inference) within each component. In general, MCLE estimators with larger components are more statistically efficient. Fig. 6 demonstrates such behavior empirically, with combs (structured MCLE) having higher C_{min} values than MPLE (unstructured MCLE).

Our bound for MCLE indicates that we should choose MCLE estimators based on their components' minimum eigenvalues, but those eigenvalues are often expensive to compute. Corollary 4.4 and Prop. 7.2 offer a solution: use a mixture of MCLE estimators. Recall that our bound's dependence on ρ_{min} indicates that a mixture of MCLE estimators $\mathcal{A}' \cup \mathcal{A}''$ should have statistical efficiency somewhere in between that of $\mathcal{A}'$ and $\mathcal{A}''$. We present a toy example in Fig. 7. This example uses a grid with stronger vertical factors than horizontal ones. As might be expected, MCLE components which include these strong edges (`Combs-vert`) make better estimators than components which do not (`Combs-horiz`). The combination of the two MCLE estimators (`Combs-both`) lies in between.

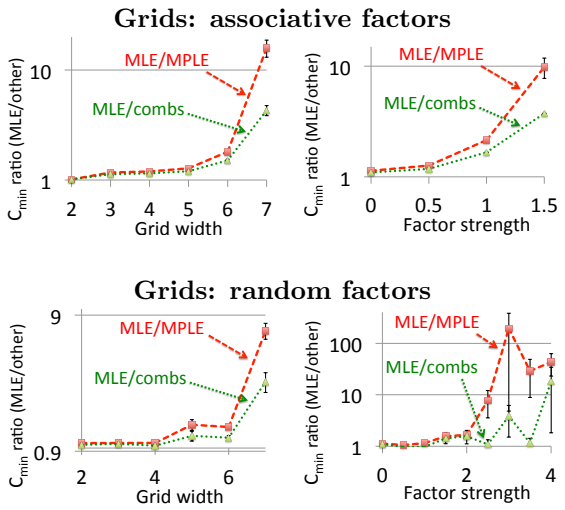

Figure 6: **Grids: Min eigval ratio MLE/MPLE.** Higher ratios imply MLE is superior. We include combs (MCLE) as described in Sec. 7; MCLE is strictly superior to MPLE. All y-axes are log-scale.
Left: The ratio increases with model size (for fixed factor strength 0.5). *Right*: The ratio increases with factor strength (for fixed model size $|Y| = 16$). Error bars 1 stddev, computed from 10 random models.

Our empirical results provide two rules of thumb for choosing a reasonable estimator when lacking expert knowledge: (A) use a small number of structured MCLE components which both cover Y and have low treewidth, and (B) combine multiple such estimators to average out the effects of "bad" components.

Fig. 8 gives empirical results on large grids, comparing MLE, MPLE, and comb-structured MCLE. Since we cannot easily compute eigenvalues for large models, we show results in terms of log loss on held-out test data. MCLE achieves much smaller log loss than MPLE, even though their training times are similar.

In related work, [2] discussed benefits of using of small tree-structured components for MCLE (though they used the components for sampling-based inference). [4] analyzed stochastic composite likelihood and found that MCLE components composed of two to four variables gave better empirical performance than MPLE.

8 Discussion

Using pseudolikelihood (MPLE) and composite likelihood (MCLE), we proved the first PAC bounds for learning parameters of general MRFs and CRFs. Our bounds are written in terms of problem-specific constants, and through empirical analysis, we showed that these constants accurately determine the relative statistical efficiency of MLE, MPLE, and MCLE.

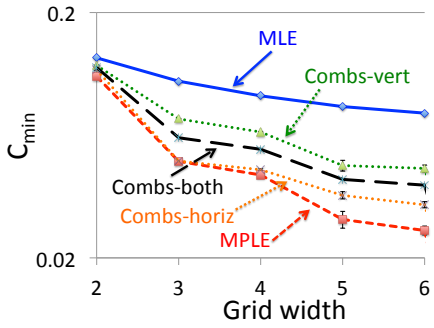

Figure 7: **Combining MCLE estimators.** As grid width increases, min eigenvalues C_{min} for estimators decrease (so learning becomes harder). Vertical factors are strongest: associative(1.5) vs. associative(.5). `Combs-vert` is two vertically oriented combs as in Fig. 1; `Combs-horiz` is the same combs rotated 90 degrees; and `Combs-both` is their combination.
Note C_{min} for `Combs-both` is the average of the C_{min} values for `Combs-vert` and `Combs-horiz`. C_{min} estimated via sampling for width ≥ 4; error bars 1 stddev.

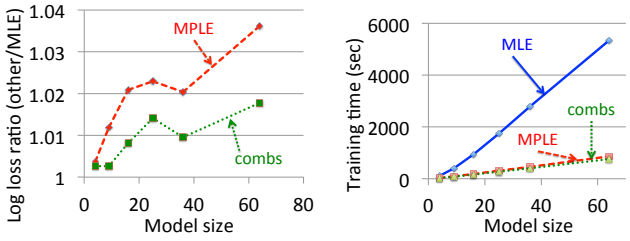

Figure 8: **Structured MCLE on grids.** Combs (MCLE) achieves smaller log loss than MPLE *(left)* but uses no more training time *(right)*. Associative(.5) factors; 10,000 training samples.

We demonstrated that structured MCLE components can provide better estimators with little additional computation. Our small-scale tests give guidance for choosing MCLE structure in practice, even when our bounds' constants may not be computed.

Future topics of interest: Generalizing our analysis to handle model misspecification would be useful; we postulate that MCLE should outperform MPLE in that setting since MCLE can be "closer" to MLE. Also, our analysis of graph properties was w.r.t. MLE, MPLE, and MCLE, and it could be augmented by similar analysis relating these estimators to other tractable learning methods, such as MLE with approximate inference. Finally, as parallel computing becomes more mainstream, more analysis of disjoint optimization could prove valuable, such as possibly improving statistical efficiency by using limited communication between separate optimizations.

Acknowledgements

Thanks to John Lafferty, Geoff Gordon, and our reviewers for helpful suggestions and feedback. Funded by NSF Career IIS-0644225, ONR YIP N00014-08-1-0752, and ARO MURI W911NF0810242.

References

[1] P. Abbeel, D. Koller, and A.Y. Ng. Learning factor graphs in polynomial time and sample complexity. *JMLR*, 7:1743–1788, 2006.

[2] A.U. Asuncion, Q. Liu, A.T. Ihler, and P. Smyth. Learning with blocks: Composite likelihood and contrastive divergence. In *AISTATS*, 2010.

[3] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[4] J. Dillon and G. Lebanon. Stochastic composite likelihood. *JMLR*, 11:2597–2633, 2010.

[5] J. Eidsvik, B.A. Shaby, B.J. Reich, M. Wheeler, and J. Niemi. Estimation and prediction in spatial models with block composite likelihoods using parallel computing. Technical report, NTNU, Duke, NCSU, UCSB, under submission.

[6] B. Gidas. Consistency of maximum likelihood and pseudolikelihood estimators for gibbs distributions. In *Proc. of Workshop on Stochastic Differential Systems with Applications in Electrical/Computer Engineering, Control Theory, and Operations Research*, 1986.

[7] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14:1771–1800, 2002.

[8] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[9] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge U. Press, Cambridge, 1990.

[10] A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.

[11] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[13] P. Liang and M. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *ICML*, 2008.

[14] B. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

[15] P. Ravikumar, M.J. Wainwright, and J. Lafferty. High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 38(3):1287, 2010.

[16] D. Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82:273–302, 1996.

[17] S. Roy, T. Lane, and M. Werner-Washburne. Learning structurally consistent undirected probabilistic graphical models. In *ICML*, 2009.

[18] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. In *CVPR*, 2008.

[19] C. Sutton and A. McCallum. Piecewise training for undirected models. In *UAI*, 2005.

[20] C. Sutton and A. McCallum. Piecewise pseudolikelihood for efficient training of conditional random fields. In *ICML*, 2007.

[21] L.G. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.

[22] M. Wainwright. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *JMLR*, 7:1829–1859, 2006.

9 APPENDIX–SUPPLEMENTARY MATERIAL

9.1 CRF Losses and Derivatives

We provide a list of losses and derivatives for general log-linear CRFs.

General log-linear CRF model:

$$\begin{aligned} P_\theta(Y|X) &= \frac{1}{Z(X;\theta)} \exp\left(\theta^T \phi(Y,X)\right) \\ &= \frac{1}{Z(X;\theta)} \exp\left(\sum_j \theta_j^T \phi_j(Y_{C_j}, X_{D_j})\right), \end{aligned} \tag{20}$$

where $\theta \in \mathbb{R}$ is a length-r vector of parameters and $\phi(Y,X) \in \mathbb{R}_+$ is a length-r non-negative feature vector. When discussing the factors making up the model, we express factor j with domain (Y_{C_j}, X_{D_j}) in terms of its corresponding parameters θ_j and features ϕ_j.

Log loss for model P_θ w.r.t. distribution P_{θ^*}:

$$\begin{aligned} \ell_L(\theta) &= \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[-\log P_\theta(Y|X)\right]\right] \\ &= \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[-\theta^T\phi(Y,X)\right] + \log Z(X;\theta)\right]. \end{aligned} \tag{21}$$

Gradient of log loss

$$\nabla \ell_L(\theta) = \mathbf{E}_{P(X)}\left[-\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi(Y,X)\right] + \mathbf{E}_{P_\theta(Y'|X)}\left[\phi(Y',X)\right]\right]. \tag{22}$$

Hessian of log loss:

$$\nabla^2 \ell_L(\theta) = \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_\theta(Y'|X)}\left[\phi(Y',X)^{\otimes}\right] - \left(\mathbf{E}_{P_\theta(Y'|X)}\left[\phi(Y',X)\right]\right)^{\otimes}\right] \tag{23}$$

$$= \mathbf{E}_{P(X)}\left[\mathbf{Var}_{P_\theta(Y'|X)}\left[\phi(Y',X)\right]\right] \tag{24}$$

Third derivative of log loss:

$$\begin{aligned} &\frac{\partial}{\partial \theta_i} \nabla^2 \ell_L(\theta) \\ &= \mathbf{E}_{P(X)}\left[\mathbf{E}\left[\phi_i \phi^{\otimes}\right] + 2\mathbf{E}\left[\phi_i\right]\mathbf{E}\left[\phi\right]^{\otimes} - \mathbf{E}\left[\phi_i\right]\mathbf{E}\left[\phi^{\otimes}\right] - \mathbf{E}\left[\phi\right]\mathbf{E}\left[\phi_i\phi\right]^T - \mathbf{E}\left[\phi_i\phi\right]\mathbf{E}\left[\phi\right]^T\right], \end{aligned} \tag{25}$$

where all unspecified expectations are w.r.t. $P_\theta(Y'|X)$ and $\phi = \phi(Y',X)$.

Composite likelihood loss for model P_θ w.r.t. distribution P_{θ^*}:

$$\begin{aligned} \ell_{CL}(\theta) &= \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[-\sum_i \log P_\theta(Y_{A_i}|Y_{\setminus A_i}, X)\right]\right] \\ &= \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\sum_i -\theta_{A_i}^T \phi_{A_i}(Y,X) + \log \sum_{y'_{A_i}} \theta_{A_i}^T \phi_{A_i}(y'_{A_i}, Y_{\setminus A_i}, X)\right]\right]. \end{aligned} \tag{26}$$

Above, the likelihood components are specified by subsets $Y_{A_i} \subseteq Y$. The parameter and feature subvectors associated with component i are specified as θ_{A_i} and ϕ_{A_i}; here, "associated with" means that at least one $Y_j \in Y_{A_i}$ is an argument of the function ϕ_k for each element k of ϕ_{A_i}.

Gradient of composite likelihood loss:

$$\nabla \ell_{CL}(\theta) = \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\sum_i -\phi_{A_i}(Y,X) + \mathbf{E}_{P_\theta(Y'_{A_i}|Y_{\setminus A_i},X)}\left[\phi_{A_i}(Y'_{A_i}, Y_{\setminus A_i}, X)\right]\right]\right]. \tag{27}$$

Above, we abuse notation in the vector summation: each element in the summation over i is a subvector of the full length-r gradient; these subvectors should be treated as length-r vectors (with zeros in the extra elements) in the summation.

Hessian of composite likelihood loss:

$$\begin{aligned}
&\nabla^2 \ell_{CL}(\theta) \\
&= \mathbf{E}_{P(X)P_{\theta^*}(Y|X)} \left[\sum_i \mathbf{E}_{P_\theta(Y'_{A_i}|Y_{\setminus A_i},X)} \left[\phi_{Y_{A_i}}(Y'_{A_i}, Y_{\setminus A_i}, X)^{\otimes} \right] - \left(\mathbf{E}_{P_\theta(Y'_{A_i}|Y_{\setminus A_i},X)} \left[\phi_{Y_{A_i}}(Y'_{A_i}, Y_{\setminus A_i}, X) \right] \right)^{\otimes} \right] \\
&= \mathbf{E}_{P(X)P_{\theta^*}(Y|X)} \left[\sum_i \mathbf{Var}_{P_\theta(Y'_{A_i}|Y_{\setminus A_i},X)} \left[\phi_{Y_{A_i}}(Y'_{A_i}, Y_{\setminus A_i}, X) \right] \right].
\end{aligned} \tag{28}$$

In the matrix summation above, we again abuse notation: each element in the sum over i is added to a submatrix of the full $r \times r$ Hessian. Each of these elements is the Hessian for a likelihood component; we denote the Hessian of the i^{th} component as $\nabla^2[\ell_{CL}(\theta)]_{A_i}$.

Third derivative of composite likelihood loss:

$$\begin{aligned}
&\frac{\partial}{\partial \theta_t} \nabla^2 \ell_{CL}(\theta) \\
&= \mathbf{E}_{P(X)P_{\theta^*}(Y|X)} \Bigg[\sum_{i\,:\,\theta_t \in \theta_{A_i}} \mathbf{E}\left[\phi_t \phi_{A_i}^{\otimes}\right] + 2\mathbf{E}\left[\phi_t\right]\mathbf{E}\left[\phi_{A_i}\right]^{\otimes} - \mathbf{E}\left[\phi_t\right]\mathbf{E}\left[\phi_{A_i}^{\otimes}\right] \\
&\qquad\qquad - \mathbf{E}\left[\phi_{A_i}\right]\mathbf{E}\left[\phi_t \phi_{A_i}\right]^T - \mathbf{E}\left[\phi_t \phi_{A_i}\right]\mathbf{E}\left[\phi_{A_i}\right]^T \Bigg],
\end{aligned} \tag{29}$$

where $\theta_t \in \theta_{A_i}$ indicates whether θ_t is an element of the parameter subvector θ_{A_i}. All unspecified expectations are w.r.t. $P_\theta(Y'_{A_i}|Y_{\setminus A_i}, X)$, and the feature function arguments are hidden: $\phi_{A_i} = \phi_{A_i}(Y'_{A_i}, Y_{\setminus A_i}, X)$ and $\phi_t = \phi_t(Y'_{A_i}, Y_{\setminus A_i}, X)$.

9.2 Parameter Estimation with MLE

We prove finite sample bounds for regression problems using log-linear models $P(Y|X)$, as defined in Eq. (20). This analysis applies both to learning CRF parameters via MLE and to learning parameters for each composite likelihood component (when using disjoint optimization). Our analysis in this section extends the analysis of Ravikumar et al. (2010) for Ising MRFs to the more general setting of log-linear CRFs. Also, while they were concerned with L_1-regularized regression in the high-dimensional setting (with the number of covariates increasing with the training set size), we limit our discussion to L_1- and L_2-regularized regression with a fixed number of covariates.

The log loss and its derivatives are defined in Eq. (21), Eq. (22), Eq. (23), and Eq. (25). We train our model by minimizing the regularized log loss over n training samples, as in eqnrefeqn:objective:crfs.

All of the results in this subsection are presented in terms of the parameters θ and corresponding features ϕ, each of which are assumed to be length-r vectors. If the feature space is overcomplete, then the constant C_{min}, the minimum eigenvalue of the Hessian of the log loss, will be zero, violating our assumption that $C_{min} > 0$. However, in all of these results, θ and ϕ (and quantities defined in terms of these vectors) may be replaced with $U^T\theta$ and $U^T\phi$, where U is a $r \times d$ matrix whose columns are eigenvectors of the Hessian corresponding to the non-zero eigenvalues. This transformation using U projects the parameters and features onto a minimal set of vectors spanning the feature space.

The following lemma (similar to Lemma 2 of Ravikumar et al. (2010)) lets us bound the max norm of the gradient of the empirical log loss at θ^* with high probability.

Lemma 9.1. *Given n samples, a bound ϕ_{max} on the magnitude of each of r features, and a constant $\delta > 0$, we have*

$$P\left[\|\nabla \hat{\ell}_L(\theta^*)\|_\infty > \delta \right] \leq 2r \exp\left(-\frac{\delta^2 n}{2\phi_{max}^2} \right). \tag{30}$$

Proof of Lemma 9.1: Consider one element j of the gradient at θ^*: $[\nabla \hat{\ell}_L(\theta^*)]_j$, as in Eq. (22). This element is a random variable (random w.r.t. the sample) with mean 0. The variable also has bounded range $[-\phi_{max}, \phi_{max}]$. We may therefore apply the Azuma-Hoeffding inequality [8], which states that

$$P\left[\left|[\nabla \hat{\ell}_L(\theta^*)]_j\right| > \delta\right] \leq 2\exp\left(-\tfrac{\delta^2 n}{2\phi_{max}^2}\right). \tag{31}$$

Applying a union bound over all r elements of the gradient gives the lemma's result. ■

We now give a lemma which shows that the minimum eigenvalue of the Hessian of the empirical log loss $\hat{\ell}_L$ (w.r.t. n samples from the target distribution) is close to that for the actual log loss ℓ_L (w.r.t. the target distribution itself).

Lemma 9.2. *Assume $\Lambda_{min}\left(\nabla^2 \ell_L(\theta^*)\right) \geq C_{min} > 0$ and $\phi_{max} = \max_{j,y,x} \phi_j(y,x)$. With n training samples, the minimum eigenvalue of the Hessian of the empirical log loss is not much smaller than C_{min} with high probability:*

$$P\left[\Lambda_{min}\left(\nabla^2 \hat{\ell}_L(\theta^*)\right) \leq C_{min} - \epsilon\right] \leq 2r^2 \exp\left(-\frac{n\epsilon^2}{8r^2\phi_{max}^4}\right). \tag{32}$$

Proof of Lemma 9.2: Our proof is similar to that of Lemma 5 from Ravikumar et al. (2010). Define shorthand for the Hessian of the log loss w.r.t. the target distribution: $\mathcal{Q} \doteq \nabla^2 \ell_L(\theta^*)$ and for the Hessian w.r.t. the empirical distribution: $\mathcal{Q}^n \doteq \nabla^2 \hat{\ell}_L(\theta^*)$. Using the Courant-Fischer variational representation [9], we can re-express the minimum eigenvalue of the Hessian:

$$\begin{aligned}
\Lambda_{min}(\mathcal{Q}) &= \min_{\|v\|_2=1} v^T \mathcal{Q} v && (33)\\
&= \min_{\|v\|_2=1} \left[v^T \mathcal{Q}^n v + v^T(\mathcal{Q} - \mathcal{Q}^n)v\right] && (34)\\
&\leq v_{min}^T \mathcal{Q}^n v_{min} + v_{min}^T(\mathcal{Q} - \mathcal{Q}^n)v_{min}, && (35)\\
& && (36)
\end{aligned}$$

where $v_{min} : \|v_{min}\|_2 = 1$ is an eigenvector corresponding to the minimum eigenvalue of $\mathcal{Q}^n$. Rearranging,

$$\begin{aligned}
\Lambda_{min}(\mathcal{Q}^n) &\geq \Lambda_{min}(\mathcal{Q}) - v_{min}^T(\mathcal{Q} - \mathcal{Q}^n)v_{min} && (37)\\
&\geq \Lambda_{min}(\mathcal{Q}) - \Lambda_{max}(\mathcal{Q} - \mathcal{Q}^n) && (38)\\
&\geq \Lambda_{min}(\mathcal{Q}) - \left(\sum_{s=1}^{r}\sum_{t=1}^{r}(\mathcal{Q}_{st} - \mathcal{Q}^n_{st})^2\right)^{1/2}, && (39)
\end{aligned}$$

where the last inequality upper-bounded the spectral norm with the Frobenius norm. Recall that

$$\nabla^2 \hat{\ell}_L(\theta^*) = \tfrac{1}{n}\sum_{i=1}^{n} \mathbf{E}_{P_{\theta^*}(Y'|x^{(i)})}\left[\phi(Y', x^{(i)})^{\otimes}\right] - \left(\mathbf{E}_{P_{\theta^*}(Y'|x^{(i)})}\left[\phi(Y', x^{(i)})\right]\right)^{\otimes} \tag{40}$$

We upper-bound the Frobenius norm term by noting that each element $(\mathcal{Q}_{st} - \mathcal{Q}^n_{st})$ may be written as an expectation over our n samples of zero-mean, bounded-range values. Abbreviating $\phi = \phi(Y, X)$ and $\phi^{(i)} = \phi(Y, x^{(i)})$, we can write:

$$\begin{aligned}
\mathcal{Q}_{st} - \mathcal{Q}^n_{st} &= \mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi_s\phi_t\right] - \mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi_s\right]\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi_t\right]\right] && (41)\\
&\quad -\tfrac{1}{n}\sum_{i=1}^{n}\left[\mathbf{E}_{P_{\theta^*}(Y|x^{(i)})}\left[\phi_s^{(i)}\phi_t^{(i)}\right] - \mathbf{E}_{P_{\theta^*}(Y|x^{(i)})}\left[\phi_s^{(i)}\right]\mathbf{E}_{P_{\theta^*}(Y|x^{(i)})}\left[\phi_t^{(i)}\right]\right] && (42)\\
&= \tfrac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{E}_{P(X)}\left[\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi_s\phi_t\right] - \mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi_s\right]\mathbf{E}_{P_{\theta^*}(Y|X)}\left[\phi_t\right]\right] && (43)\\
&\quad - \left[\mathbf{E}_{P_{\theta^*}(Y|x^{(i)})}\left[\phi_s^{(i)}\phi_t^{(i)}\right] - \mathbf{E}_{P_{\theta^*}(Y|x^{(i)})}\left[\phi_s^{(i)}\right]\mathbf{E}_{P_{\theta^*}(Y|x^{(i)})}\left[\phi_t^{(i)}\right]\right]\Big]. && (44)
\end{aligned}$$

Each of these n values has magnitude at most $2\phi_{max}^2$. The Azuma-Hoeffding inequality [8] tells us that, for any s, t,

$$P\left[(\mathcal{Q}_{st} - \mathcal{Q}^n_{st})^2 \geq \epsilon^2\right] = P\left[|\mathcal{Q}_{st} - \mathcal{Q}^n_{st})| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{8\phi_{max}^4}\right). \tag{45}$$

A union bound over all elements s, t shows:

$$P\left[\sum_{s=1}^{r}\sum_{t=1}^{r}(\mathcal{Q}_{st}-\mathcal{Q}_{st}^{n})^2 \geq r^2\epsilon^2\right] \leq 2r^2\exp\left(-\frac{n\epsilon^2}{8\phi_{max}^4}\right) \quad (46)$$

$$P\left[\left(\sum_{s=1}^{r}\sum_{t=1}^{r}(\mathcal{Q}_{st}-\mathcal{Q}_{st}^{n})^2\right)^{1/2} \geq \epsilon\right] \leq 2r^2\exp\left(-\frac{n\epsilon^2}{8r^2\phi_{max}^4}\right). \quad (47)$$

Using the above inequality with Eq. (39), we get:

$$P\left[\Lambda_{min}(\mathcal{Q}^n) \leq C_{min}-\epsilon\right] \leq 2r^2\exp\left(-\frac{n\epsilon^2}{8r^2\phi_{max}^4}\right). \quad \blacksquare \quad (48)$$

We next prove a lemma which lower-bounds the log loss (w.r.t. our training samples) in terms of the parameter estimation error (the distance between our estimated parameters $\hat{\theta}$ and the target parameters θ^*). The analysis resembles part of Lemma 3 from Ravikumar et al. (2010).

Lemma 9.3. *Let $\hat{\ell}_L(\theta)$ be the log loss w.r.t. n training samples. Assume bounds $\Lambda_{min}\left(\nabla^2\ell_L(\theta^*)\right) \geq C_{min} > 0$ and $\phi_{max} = \max_{j,y,x}\phi_j(y,x)$. Let $\delta > 0$. Let $B = \|\theta-\theta^*\|_1$. Then*

$$\hat{\ell}_L(\theta) - \hat{\ell}_L(\theta^*) \geq -\delta B + \tfrac{r^{-1}}{4}C_{min}B^2 - \tfrac{1}{2}\phi_{max}^3 B^3 \quad (49)$$

with probability at least

$$1 - 2r\exp\left(-\tfrac{\delta^2 n}{2\phi_{max}^2}\right) - 2r^2\exp\left(-\tfrac{nC_{min}^2}{2^5 r^2\phi_{max}^4}\right). \quad (50)$$

Proof of Lemma 9.3: Let $u = \theta-\theta^*$ and $\|u\|_1 = B$. Use a Taylor expansion of $\hat{\ell}_L$ around θ^*:

$$\hat{\ell}_L(\theta) = \hat{\ell}_L(\theta^*) + \left(\nabla\hat{\ell}_L(\theta^*)\right)^T u + \tfrac{1}{2}u^T\left(\nabla^2\hat{\ell}_L(\theta^*)\right)u + \tfrac{1}{6}\sum_i u_i u^T\left(\tfrac{\partial}{\partial\theta_i}\nabla^2\hat{\ell}_L(\bar{\theta})\big|_{\bar{\theta}=\theta^*+\alpha u}\right)u, \quad (51)$$

where $\alpha \in [0,1]$. We now lower-bound the first-, second-, and third-order terms.

First-order term in Eq. (51):

$$\left(\nabla\hat{\ell}_L(\theta^*)\right)^T u \geq -\left|\left(\nabla\hat{\ell}_L(\theta^*)\right)^T u\right| \quad (52)$$

$$\geq -\|\nabla\hat{\ell}_L(\theta^*)\|_\infty \|u\|_1 \quad (53)$$

$$= -\|\nabla\hat{\ell}_L(\theta^*)\|_\infty B \quad (54)$$

$$\geq -\delta B, \quad (55)$$

where the second inequality uses Holder's inequality, and where the last inequality uses Lemma 9.1 and holds with probability at least $1-2r\exp(-\frac{\delta^2 n}{2\phi_{max}^2})$.

Second-order term in Eq. (51):

$$\tfrac{1}{2}u^T\left(\nabla^2\hat{\ell}_L(\theta^*)\right)u \geq \tfrac{1}{2}\Lambda_{min}\left(\nabla^2\hat{\ell}_L(\theta^*)\right)\|u\|_2^2 \quad (56)$$

$$\geq \tfrac{r^{-1}}{2}\Lambda_{min}\left(\nabla^2\hat{\ell}_L(\theta^*)\right)\|u\|_1^2 \quad (57)$$

$$= \tfrac{r^{-1}}{2}\Lambda_{min}\left(\nabla^2\hat{\ell}_L(\theta^*)\right)B^2, \quad (58)$$

where we used the definition of $\Lambda_{min}(\nabla^2\hat{\ell}_L(\theta^*))$, the minimum eigenvalue of the Hessian at θ^*. Now use Lemma 9.2 with $\epsilon = \frac{C_{min}}{2}$ to show:

$$\tfrac{1}{2}u^T\left(\nabla^2\hat{\ell}_L(\theta^*)\right)u \geq \tfrac{r^{-1}}{4}C_{min}B^2, \quad (59)$$

which holds with probability at least $1 - 2r^2 \exp\left(-\frac{nC_{min}^2}{2^5 r^2 \phi_{max}^4}\right)$.

Third-order term in Eq. (51):

$$\frac{1}{6}\sum_i u_i u^T \left(\frac{\partial}{\partial\bar{\theta}_i}\nabla^2 \hat{\ell}_L(\bar{\theta})\big|_{\bar{\theta}=\theta^*+\alpha u}\right) u \tag{60}$$

$$= \frac{1}{6}\sum_i u_i u^T \Big(\mathbf{E}\left[\phi_i \phi^{\otimes}\right] + 2\mathbf{E}\left[\phi_i\right]\mathbf{E}\left[\phi\right]^{\otimes} - \mathbf{E}\left[\phi_i\right]\mathbf{E}\left[\phi^{\otimes}\right] \tag{61}$$

$$-\mathbf{E}\left[\phi\right]\mathbf{E}\left[\phi_i\phi\right]^T - \mathbf{E}\left[\phi_i\phi\right]\mathbf{E}\left[\phi\right]^T\Big) u,$$

where all expectations are w.r.t. $P_{\theta^*+\alpha u}(Y'|X)$ and $\phi = \phi(Y', X)$. Continuing,

$$\frac{1}{6}\sum_i u_i u^T \Big(\mathbf{E}\left[\phi_i \phi^{\otimes}\right] + 2\mathbf{E}\left[\phi_i\right]\mathbf{E}\left[\phi\right]^{\otimes} - \mathbf{E}\left[\phi_i\right]\mathbf{E}\left[\phi^{\otimes}\right] \tag{62}$$

$$-\mathbf{E}\left[\phi\right]\mathbf{E}\left[\phi_i\phi\right]^T - \mathbf{E}\left[\phi_i\phi\right]\mathbf{E}\left[\phi\right]^T\Big) u$$

$$= \frac{1}{6}\sum_i u_i \Big(\mathbf{E}\left[\phi_i (u^T\phi)^2\right] + 2\mathbf{E}\left[\phi_i\right]\mathbf{E}\left[u^T\phi\right]^2 - \mathbf{E}\left[\phi_i\right]\mathbf{E}\left[(u^T\phi)^2\right] \tag{63}$$

$$-2\mathbf{E}\left[u^T\phi\right]\mathbf{E}\left[\phi_i u^T\phi\right]\Big)$$

$$= \frac{1}{6}\Big(\mathbf{E}\left[(u^T\phi)^3\right] + 2\mathbf{E}\left[u^T\phi\right]^3 - 3\mathbf{E}\left[u^T\phi\right]\mathbf{E}\left[(u^T\phi)^2\right]\Big) \tag{64}$$

$$\geq \frac{1}{6}\Big(3\mathbf{E}\left[u^T\phi\right]^3 - 3\mathbf{E}\left[u^T\phi\right]\mathbf{E}\left[(u^T\phi)^2\right]\Big) \tag{65}$$

$$= -\frac{1}{2}\mathbf{E}\left[u^T\phi\right]\Big(\mathbf{E}\left[(u^T\phi)^2\right] - \mathbf{E}\left[u^T\phi\right]^2\Big) \tag{66}$$

$$\geq -\frac{1}{2}\left|\mathbf{E}\left[u^T\phi\right]\right| \cdot \left|\mathbf{E}\left[(u^T\phi)^2\right] - \mathbf{E}\left[u^T\phi\right]^2\right| \tag{67}$$

$$\geq -\frac{1}{2}\left|\mathbf{E}\left[u^T\phi\right]\right| \cdot \mathbf{E}\left[(u^T\phi)^2\right] \tag{68}$$

$$\geq -\frac{1}{2}\mathbf{E}\left[|u^T\phi|\right] \cdot \mathbf{E}\left[|u^T\phi|^2\right] \tag{69}$$

$$\geq -\frac{1}{2}\left\|u\right\|_1^3 \phi_{max}^3 \tag{70}$$

$$= -\frac{1}{2}B^3\phi_{max}^3. \tag{71}$$

Two of our bounds in this proof had small probabilities of failure. Using a union bound, we get the probability of at least one failing, finishing the proof. ■

We now prove a lemma which bounds our parameter estimation error in terms of our training sample size; it is similar to Lemma 3 from Ravikumar et al. (2010). Note that $\hat{\theta}$ is defined as the minimizer of Eq. (2) with $\hat{\ell} = \hat{\ell}_L$.

Proof of Theorem 4.1: Define a convex function $G : \mathbb{R}^r \to \mathbb{R}$ by

$$G(u) = \hat{\ell}_L(\theta^* + u) - \hat{\ell}_L(\theta^*) + \lambda\left(\|\theta^* + u\|_p - \|\theta^*\|_p\right). \tag{72}$$

By definition of $\hat{\theta}$, the function G is minimized at $\hat{u} = \hat{\theta} - \theta^*$. Since $G(0) = 0$, we know $G(\hat{u}) \leq 0$. Using the same argument as Ravikumar et al. (2010), if $G(u) > 0$ for all $u \in \mathbb{R}^r$ with $\|u\|_1 = B$ for some $B > 0$, then we know that $\|\hat{u}\|_1 \leq B$.

Let $u \in \mathbb{R}^r$ with $\|u\|_1 = B$. Using Lemma 9.3, we can lower-bound G:

$$\begin{aligned} G(u) \quad \geq \quad & -\delta B + \tfrac{r^{-1}}{4} C_{min} B^2 - \tfrac{1}{2}\phi_{max}^3 B^3 \\ & +\lambda\left(\|\theta^* + u\|_p - \|\theta^*\|_p\right), \end{aligned} \tag{73}$$

which holds with the probability given in Lemma 9.3. We can lower-bound the regularization term (for both L_1 and L_2 regularization):

$$\lambda\left(\|\theta^* + u\|_p - \|\theta^*\|_p\right) \quad \geq \quad -\lambda\|u\|_p \tag{74}$$

$$(\text{If } p = 1) \qquad = -\lambda\|u\|_1 = -\lambda B. \tag{75}$$

$$(\text{If } p = 2) \qquad = -\lambda\|u\|_2 \geq -\lambda\|u\|_1 = -\lambda B. \tag{76}$$

Combine the above bound into Eq. (73):

$$\begin{aligned} G(u) &\geq -\delta B + \tfrac{r^{-1}}{4} C_{min} B^2 - \tfrac{1}{2}\phi_{max}^3 B^3 - \lambda B & (77)\\ &= B\left[-\delta + \tfrac{r^{-1}}{4} C_{min} B - \tfrac{1}{2}\phi_{max}^3 B^2 - \lambda\right]. & (78)\end{aligned}$$

Note that we need $\lambda > 0, B > 0, \delta > 0$. Eq. (78) will be strictly greater than 0 if $B > 0$ and

$$\lambda < -\delta + \tfrac{r^{-1}}{4} C_{min} B - \tfrac{1}{2}\phi_{max}^3 B^2. \quad (79)$$

Maximizing this bound w.r.t. B gives $B = \frac{C_{min}}{4r\phi_{max}^3}$. However, we would like for B to shrink as $n^{-1/2}$, the asymptotic rate of convergence, so instead let

$$B = \frac{C_{min}}{4r\phi_{max}^3} n^{-\xi/2}, \quad (80)$$

where $\xi \in (0, 1)$. Plugging in this value for B gives

$$\lambda < -\delta + \frac{C_{min}^2}{2^4 r^2 \phi_{max}^3} n^{-\xi/2} - \frac{C_{min}^2}{2^5 r^2 \phi_{max}^3} n^{-\xi}. \quad (81)$$

We want to choose δ to be large but still keep $\lambda > 0$, so choose

$$\lambda = \delta = \frac{C_{min}^2}{2^6 r^2 \phi_{max}^3} n^{-\xi/2}, \quad (82)$$

which makes Eq. (81) hold if $n > 1$. Now that we have chosen δ, we can simplify the probability of failure from Lemma 9.3:

$$\begin{aligned} & 2r \exp\left(-\tfrac{\delta^2 n}{2\phi_{max}^2}\right) + 2r^2 \exp\left(-\tfrac{n C_{min}^2}{2^5 r^2 \phi_{max}^4}\right) & (83)\\ &= 2r \exp\left(-\tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi}\right) + 2r^2 \exp\left(-\tfrac{C_{min}^2}{2^5 r^2 \phi_{max}^4} n\right) & (84)\\ &\leq 2r \exp\left(-\tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi}\right) + 2r^2 \exp\left(-\tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi}\right) & (85)\\ &= 2r(r+1) \exp\left(-\tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi}\right). & (86)\end{aligned}$$

Above, we upper-bounded the spectral norm with the Frobenius norm to show that $C_{min} \leq \Lambda_{max}\left(\nabla^2 \hat{\ell}_L(\theta^*)\right) \leq \phi_{max}^2 r$. ■

We can convert the previous result into a sample complexity bound for achieving a given parameter estimation error.

Proof of Corollary 4.2: If we wish to have a probability of failure of at most δ when we have n samples, we may choose ξ accordingly:

$$\begin{aligned} 2r(r+1) \exp\left(-\tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi}\right) &\leq \delta & (87)\\ \log(2r(r+1)) - \tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi} &\leq \log \delta & (88)\\ \tfrac{C_{min}^4}{2^{13} r^4 \phi_{max}^8} n^{1-\xi} &\geq \log \tfrac{2r(r+1)}{\delta} & (89)\\ n^{1-\xi} &\geq \tfrac{2^{13} r^4 \phi_{max}^8}{C_{min}^4} \log \tfrac{2r(r+1)}{\delta} & (90)\\ 1 - \xi &\geq \tfrac{1}{\log n}\left(\log \tfrac{2^{13} r^4 \phi_{max}^8}{C_{min}^4} + \log\log \tfrac{2r(r+1)}{\delta}\right) & (91)\\ \xi &\leq 1 - \tfrac{1}{\log n}\left(\log \tfrac{2^{13} r^4 \phi_{max}^8}{C_{min}^4} + \log\log \tfrac{2r(r+1)}{\delta}\right). & (92)\end{aligned}$$

We will set ξ equal to this upper bound in the next part. Likewise, if we wish to have parameter estimation error at most ϵ, then we need:

$$\begin{aligned}
\tfrac{C_{min}}{4r\phi_{max}^3} n^{-\xi/2} &\leq \epsilon && (93)\\
\log \tfrac{C_{min}}{4r\phi_{max}^3} - \tfrac{\xi}{2}\log n &\leq \log\epsilon && (94)\\
\tfrac{\xi}{2}\log n &\geq \log \tfrac{C_{min}}{4r\phi_{max}^3} - \log\epsilon && (95)\\
\tfrac{1}{2}\left(1 - \tfrac{1}{\log n}\left(\log \tfrac{2^{13}r^4\phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(r+1)}{\delta}\right)\right)\log n &\geq \log \tfrac{C_{min}}{4r\phi_{max}^3} - \log\epsilon && (96)\\
\log n - \left(\log \tfrac{2^{13}r^4\phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(r+1)}{\delta}\right) &\geq 2\log\tfrac{C_{min}}{4r\phi_{max}^3\epsilon}. && (97)
\end{aligned}$$

$$\begin{aligned}
\log n &\geq 2\log\tfrac{C_{min}}{4r\phi_{max}^3\epsilon} + \log\tfrac{2^{13}r^4\phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(r+1)}{\delta} && (98)\\
&= \log\tfrac{C_{min}^2}{2^4r^2\phi_{max}^6} + \log\tfrac{2^{13}r^4\phi_{max}^8}{C_{min}^4} + 2\log\tfrac{1}{\epsilon} + \log\log\tfrac{2r(r+1)}{\delta} && (99)\\
&= \log\tfrac{2^9r^2\phi_{max}^2}{C_{min}^2} + 2\log\tfrac{1}{\epsilon} + \log\log\tfrac{2r(r+1)}{\delta} && (100)\\
n &\geq \tfrac{2^9r^2\phi_{max}^2}{C_{min}^2}\tfrac{1}{\epsilon^2}\log\tfrac{2r(r+1)}{\delta}. \quad \blacksquare && (101)
\end{aligned}$$

9.3 Parameter Estimation with MCLE

We train our model by minimizing the regularized composite likelihood loss over n training samples:

$$\min_\theta \hat{\ell}_{CL}(\theta) + \lambda\|\theta\|_p, \tag{102}$$

where $\hat{\ell}_{CL}(\theta)$ is the composite likelihood loss w.r.t. the n training samples, $\lambda \geq 0$ is a regularization parameter, and $p \in \{1, 2\}$ specifies the L_1 or L_2 norm. Note that $\hat{\ell}_{CL}(\theta)$ is Eq. (26) with $P(X)P_{\theta^*}(Y|X)$ replaced with the empirical distribution.

Lemma 9.4. *Given n samples, a bound ϕ_{max} on the magnitude of each of r features, a bound M_{max} on the number of likelihood components in which any feature participates, and a constant $\delta > 0$, we have*

$$P\left[\|\nabla\ell_{CL}(\theta^*)\|_\infty > \delta\right] \leq 2r\exp\left(-\frac{\delta^2 n}{2M_{max}^2\phi_{max}^2}\right). \tag{103}$$

Proof of Lemma 9.4: Consider one element j of the expected gradient at θ^*: $\mathbf{E}_{P(X)P_{\theta^*}(Y|X)}\left[[\nabla\ell_{CL}(\theta^*)]_j\right]$. This element is a random variable (random w.r.t. the sample) with mean 0. The variable also has bounded range $[-M_j\phi_{max}, M_j\phi_{max}]$, where M_j is the number of likelihood components in which θ_j participates. We may therefore apply the Azuma-Hoeffding inequality [8], which states that

$$P\left[|[\nabla\ell_{CL}(\theta^*)]_j| > \delta\right] \leq 2\exp\left(-\tfrac{\delta^2 n}{2M_j^2\phi_{max}^2}\right). \tag{104}$$

Applying a union bound over all r elements of the gradient and using $M_{max} = \max_j M_j$ gives the lemma's result. $\blacksquare$

Lemma 9.5. *Define $\rho_t = \sum_{i:\theta_t\in\theta_{A_i}} \Lambda_{min}(\nabla^2[\ell_{CL}(\theta^*)]_{A_i})$, i.e., the sum of minimum eigenvalues for all likelihood components in which parameter θ_t participates (w.r.t. the target distribution). Define $\hat{\rho}_t$ analogously w.r.t. $\hat{\ell}_{CL}$ (i.e., w.r.t. the empirical distribution). Assume $\phi_{max} = \max_{j,y,x}\phi_j(y,x)$. Let $\mathcal{A}$ denote the set of likelihood components. With n training samples, the empirical quantities $\hat{\rho}_t$ are not much smaller than ρ_t with high probability:*

$$P\left[\exists t\,:\,\hat{\rho}_t \leq \tfrac{\rho_t}{2}\right] \leq 2|\mathcal{A}|r^2\exp\left(-\frac{nC_{min}^2}{2^5r^2\phi_{max}^4}\right). \tag{105}$$

Proof of Lemma 9.5: Let C_i be the minimum eigenvalue of the Hessian of likelihood component A_i w.r.t the target distribution. The Hessian of each likelihood component may be analyzed using Lemma 9.2, with $\epsilon = \frac{C_i}{2}$, giving the result:

$$P\left[\Lambda_{min}\left(\nabla^2[\hat{\ell}_{CL}(\theta^*)]_{A_i}\right) \leq \tfrac{C_i}{2}\right] \leq 2r_{A_i}^2\exp\left(-\frac{nC_i^2}{2^5r_{A_i}^2\phi_{max}^4}\right), \tag{106}$$

where r_{A_i} denotes the length of the parameter vector corresponding to likelihood component A_i. A union bound over all $|\mathcal{A}|$ likelihood components in the min operation gives:

$$P\left[\exists i \,:\, \Lambda_{min}\left(\nabla^2[\hat{\ell}_{CL}(\theta^*)]_{A_i}\right) \leq \tfrac{C_i}{2}\right] \leq 2|\mathcal{A}|r^2 \exp\left(-\frac{nC_{min}^2}{2^5 r^2 \phi_{max}^4}\right). \tag{107}$$

I.e., all components' minimum eigenvalues (w.r.t. the training data) are within a factor of $\frac{1}{2}$ of their true eigenvalues (w.r.t. the target distribution) with high probability, which implies that sums of sets of eigenvalues are likewise estimated within a factor of $\frac{1}{2}$, giving the lemma's result. ■

Lemma 9.6. *Let $\hat{\ell}_{CL}(\theta)$ be the composite likelihood loss w.r.t. n training samples. Assume bounds $\min_i \Lambda_{min}\left(\nabla^2[\ell_{CL}(\theta^*)]_{A_i}\right) \geq C_{min} > 0$ and $\phi_{max} = \max_{j,y,x} \phi_j(y,x)$. Let $\rho_{min} = \min_t \rho_t$. Let M_{max}, M_{min} be the maximum and minimum numbers of components any feature participates in, respectively. Let $\delta > 0$. Let $B = \|\theta - \theta^*\|_1$. Then*

$$\hat{\ell}_{CL}(\theta) - \hat{\ell}_{CL}(\theta^*) \geq -\delta B + \tfrac{r^{-1}}{4}\rho_{min} B^2 - \tfrac{1}{2} M_{max} \phi_{max}^3 B^3 \tag{108}$$

with probability at least

$$1 - 2r \exp\left(-\tfrac{\delta^2 n}{2M_{max}^2 \phi_{max}^2}\right) - 2|\mathcal{A}|r^2 \exp\left(-\tfrac{nC_{min}^2}{2^5 r^2 \phi_{max}^4}\right). \tag{109}$$

Proof of Lemma 9.6: We abbreviate this proof where it is similar to that of Lemma 9.3.

Let $u = \theta - \theta^*$ and $\|u\|_1 = B$. Use a Taylor expansion of $\hat{\ell}_{CL}$ around θ^*:

$$\hat{\ell}_{CL}(\theta) = \hat{\ell}_{CL}(\theta^*) + \left(\nabla \hat{\ell}_{CL}(\theta^*)\right)^T u + \tfrac{1}{2} u^T \left(\nabla^2 \hat{\ell}_{CL}(\theta^*)\right) u + \tfrac{1}{6} \sum_i u_i u^T \left(\tfrac{\partial}{\partial \bar{\theta}_i} \nabla^2 \hat{\ell}_{CL}(\bar{\theta})|_{\bar{\theta} = \theta^* + \alpha u}\right) u, \tag{110}$$

where $\alpha \in [0,1]$. We now lower-bound the first-, second-, and third-order terms.

First-order term in Eq. (110): We can use Lemma 9.4 to bound the first term with $(\nabla \hat{\ell}_{CL}(\theta^*))^T u \geq -\delta B$ with probability at least $1 - 2r\exp(-\frac{\delta^2 n}{2M_{max}^2 \phi_{max}^2})$.

Second-order term in Eq. (110): Let u_{A_i} denote the elements of u corresponding to the component of the pseudolikelihood loss for Y_{A_i}; let r_{A_i} denote the length of u_{A_i}; and let M_{min} denote the minimum number of likelihood components in which any parameter participates.

$$\begin{aligned}
\tfrac{1}{2} u^T \left(\nabla^2 \hat{\ell}_{CL}(\theta^*)\right) u &= \tfrac{1}{2} \sum_i u_{A_i}^T \left(\nabla^2 [\hat{\ell}_{CL}(\theta^*)]_{A_i}\right) u_{A_i} && (111)\\
&\geq \tfrac{1}{2} \sum_i \Lambda_{min}\left(\nabla^2 [\hat{\ell}_{CL}(\theta^*)]_{A_i}\right) \|u_{A_i}\|_2^2 && (112)
\end{aligned}$$

Continuing,

$$\begin{aligned}
\tfrac{1}{2} u^T \left(\nabla^2 \hat{\ell}_{CL}(\theta^*)\right) u &\geq \tfrac{1}{2} \sum_i \Lambda_{min}\left(\nabla^2 [\hat{\ell}_{CL}(\theta^*)]_{A_i}\right) \|u_{A_i}\|_2^2 && (113)\\
&= \tfrac{1}{2} \sum_t \left(\sum_{i\,:\, u_t \in u_{A_i}} \Lambda_{min}\left(\nabla^2 [\hat{\ell}_{CL}(\theta^*)]_{A_i}\right)\right) u_t^2 && (114)\\
&= \tfrac{1}{2} \sum_t \hat{\rho}_t u_t^2 && (115)\\
&\geq \tfrac{1}{4} \sum_t \rho_t u_t^2 && (116)\\
&\geq \tfrac{1}{4} \rho_{min} \sum_t u_t^2 && (117)\\
&= \tfrac{1}{4} \rho_{min} \|u\|_2^2 && (118)\\
&\geq \tfrac{r^{-1}}{4} \rho_{min} \|u\|_1^2 && (119)\\
&= \tfrac{r^{-1}}{4} \rho_{min} B^2, && (120)
\end{aligned}$$

where we used Lemma 9.5 to lower-bound $\hat{\rho}_t$ with $\rho_t/2$ with high probability.

Third-order term in Eq. (110):

$$\frac{1}{6}\sum_t u_t u^T \left(\frac{\partial}{\partial\bar{\theta}_t}\nabla^2\hat{\ell}_{CL}(\bar{\theta})\big|_{\bar{\theta}=\theta^*+\alpha u}\right) u \tag{121}$$

$$= \frac{1}{6}\sum_t u_t u^T \Bigg(\sum_{i\,:\,\theta_t\in\theta_{A_i}} \mathbf{E}\left[\phi_t \phi_{A_i}^{\otimes}\right] + 2\mathbf{E}\left[\phi_t\right]\mathbf{E}\left[\phi_{A_i}\right]^{\otimes} - \mathbf{E}\left[\phi_t\right]\mathbf{E}\left[\phi_{A_i}^{\otimes}\right] \tag{122}$$

$$-\mathbf{E}\left[\phi_{A_i}\right]\mathbf{E}\left[\phi_t\phi_{A_i}\right]^T - \mathbf{E}\left[\phi_t\phi_{A_i}\right]\mathbf{E}\left[\phi_{A_i}\right]^T \Bigg) u, \tag{123}$$

where all expectations are w.r.t. $P_{\theta^*+\alpha u}(Y'_{A_i}|Y_{\setminus A_i}, X)$, and $\phi_{A_i} = \phi_{A_i}(Y'_{A_i}, Y_{\setminus A_i}, X)$ and $\phi_t = \phi_t(Y'_{A_i}, Y_{\setminus A_i}, X)$. We can lower-bound and collapse the various terms on the right-hand side, just as in the proof of Theorem 4.1:

$$\frac{1}{6}\sum_t u_t u^T \left(\frac{\partial}{\partial\bar{\theta}_t}\nabla^2\hat{\ell}_{CL}(\bar{\theta})\big|_{\bar{\theta}=\theta^*+\alpha u}\right) u \tag{124}$$

$$= \frac{1}{6}\sum_t u_t \sum_{i\,:\,\theta_t\in\theta_{A_i}} u_{A_i}^T \Bigg(\mathbf{E}\left[\phi_t \phi_{A_i}^{\otimes}\right] + 2\mathbf{E}\left[\phi_t\right]\mathbf{E}\left[\phi_{A_i}\right]^{\otimes} - \mathbf{E}\left[\phi_t\right]\mathbf{E}\left[\phi_{A_i}^{\otimes}\right] \tag{125}$$

$$-\mathbf{E}\left[\phi_{A_i}\right]\mathbf{E}\left[\phi_t\phi_{A_i}\right]^T - \mathbf{E}\left[\phi_t\phi_{A_i}\right]\mathbf{E}\left[\phi_{A_i}\right]^T \Bigg) u_{A_i} \tag{126}$$

$$= \frac{1}{6}\sum_t u_t \sum_{i\,:\,\theta_t\in\theta_{A_i}} \Bigg(\mathbf{E}\left[\phi_t (u_{A_i}^T\phi_{A_i})^2\right] + 2\mathbf{E}\left[\phi_t\right]\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]^2 - \mathbf{E}\left[\phi_t\right]\mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right] \tag{127}$$

$$-2\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[\phi_t u_{A_i}^T\phi_{A_i}\right]^T \Bigg) u_{A_i} \tag{128}$$

$$= \frac{1}{6}\sum_i \mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^3\right] + 2\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]^2 - 3\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right]. \tag{129}$$

Using Jensen's inequality multiple times, we can continue:

$$\frac{1}{6}\sum_i \mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^3\right] + 2\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]^2 - 3\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right] \tag{130}$$

$$\geq \frac{1}{6}\sum_i 3\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]^3 - 3\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right] \tag{131}$$

$$= \frac{1}{2}\sum_i \mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]^3 - \mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right] \tag{132}$$

$$= -\frac{1}{2}\sum_i \left|\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\right| \cdot \left(\mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right] - \mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]^2\right) \tag{133}$$

$$\geq -\frac{1}{2}\sum_i \left|\mathbf{E}\left[u_{A_i}^T\phi_{A_i}\right]\right| \mathbf{E}\left[(u_{A_i}^T\phi_{A_i})^2\right] \tag{134}$$

$$\geq -\frac{1}{2}\sum_i \mathbf{E}\left[\left|u_{A_i}^T\phi_{A_i}\right|^3\right]. \tag{135}$$

Applying Holder's inequality, we can continue:

$$-\frac{1}{2}\sum_i \mathbf{E}\left[\left|u_{A_i}^T\phi_{A_i}\right|^3\right] \;\geq\; -\frac{1}{2}\sum_i \|u_{A_i}\|_1^3\,\phi_{max}^3 \;\geq\; -\frac{1}{2}M_{max}\,\|u\|_1^3\,\phi_{max}^3 \;=\; -\frac{1}{2}M_{max}B^3\phi_{max}^3. \tag{136}$$

Two of our bounds in this proof had small probabilities of failure. Using a union bound, we get the probability of at least one failing, finishing the proof. ■

Proof of Theorem 4.3: As in the proof of Theorem 4.1, we define $G : \mathbb{R}^r \to \mathbb{R}$ by

$$G(u) = \hat{\ell}_{CL}(\theta^* + u) - \hat{\ell}_{CL}(\theta^*) + \lambda\left(\|\theta^* + u\|_p - \|\theta^*\|_p\right), \tag{137}$$

with the difference that we now use the composite likelihood loss. As in Theorem 4.1, we wish to show that $G(u) > 0$ for all $u \in \mathbb{R}^r$ with $\|u\|_1 = B$ for some $B > 0$, which will imply that $\|\hat{u}\|_1 \leq B$.

Let $u \in \mathbb{R}^r$ with $\|u\|_1 = B$. Using Lemma 9.6, we can lower-bound G:

$$\begin{aligned} G(u) &\geq -\delta B + \tfrac{r^{-1}}{4}\rho_{min}B^2 - \tfrac{1}{2}M_{max}\phi_{max}^3 B^3 \\ &\quad +\lambda\left(\|\theta^* + u\|_p - \|\theta^*\|_p\right), \end{aligned} \tag{138}$$

which holds with the probability given in Lemma 9.6. As in the proof of Theorem 4.1, we can lower-bound the regularization term (for both L_1 and L_2 regularization): $\lambda\left(\|\theta^* + u\|_p - \|\theta^*\|_p\right) \geq -\lambda B$. Combining these bounds, we get:

$$G(u) \geq -\delta B + \tfrac{r^{-1}}{4}\rho_{min}B^2 - \tfrac{1}{2}M_{max}B^3\phi_{max}^3 - \lambda B \tag{139}$$

$$= B\left[-\delta + \tfrac{r^{-1}}{4}\rho_{min}B - \tfrac{1}{2}M_{max}B^2\phi_{max}^3 - \lambda\right]. \tag{140}$$

Note that we need $\lambda > 0, B > 0, \delta > 0$. Eq. (140) will be strictly greater than 0 if $B > 0$ and

$$\lambda < -\delta + \tfrac{r^{-1}}{4}\rho_{min}B - \tfrac{1}{2}M_{max}B^2\phi_{max}^3. \tag{141}$$

Maximizing this bound w.r.t. B gives $B = \frac{\rho_{min}}{4rM_{max}\phi_{max}^3}$. However, we would like for B to shrink as $n^{-1/2}$, the asymptotic rate of convergence, so instead let

$$B = \frac{\rho_{min}}{4rM_{max}\phi_{max}^3} n^{-\xi/2}, \tag{142}$$

where $\xi \in (0, 1)$. Plugging in this value for B gives

$$\lambda < -\delta + \frac{\rho_{min}^2}{2^4 r^2 M_{max}\phi_{max}^3} n^{-\xi/2} - \frac{\rho_{min}^2}{2^5 r^2 M_{max}\phi_{max}^3} n^{-\xi}. \tag{143}$$

We want to choose δ to be large but still keep $\lambda > 0$, so choose

$$\lambda = \delta = \frac{\rho_{min}^2}{2^6 r^2 M_{max}\phi_{max}^3} n^{-\xi/2}, \tag{144}$$

which makes Eq. (143) hold if $n > 1$. Now that we have chosen δ, we can simplify the probability of failure from Lemma 9.6:

$$2r\exp\left(-\tfrac{\delta^2 n}{2M_{max}^2\phi_{max}^2}\right) + 2|\mathcal{A}|r^2\exp\left(-\tfrac{nC_{min}^2}{2^5 r^2\phi_{max}^4}\right) \tag{145}$$

$$= 2r\exp\left(-\tfrac{\rho_{min}^4 n^{1-\xi}}{2^{13}r^4 M_{max}^4\phi_{max}^8}\right) + 2|\mathcal{A}|r^2\exp\left(-\tfrac{nC_{min}^2}{2^5 r^2\phi_{max}^4}\right) \tag{146}$$

$$\leq 2r\exp\left(-\tfrac{C_{min}^4 n^{1-\xi}}{2^{13}r^4 M_{max}^4\phi_{max}^8}\right) + 2|\mathcal{A}|r^2\exp\left(-\tfrac{nC_{min}^2}{2^5 r^2\phi_{max}^4}\right) \tag{147}$$

$$\leq 2r(|\mathcal{A}|r + 1)\exp\left(-\tfrac{C_{min}^4}{2^{13}r^4 M_{max}^4\phi_{max}^8} n^{1-\xi}\right). \tag{148}$$

The above bound is very loose in combining the two exponential terms; in particular, we would like ρ_{min} to remain in the bound. To derive a sufficient condition for this tighter combination, we first require that the left-hand term in the probability of failure be meaningful, i.e., at most 1:

$$2r\exp\left(-\tfrac{\rho_{min}^4 n^{1-\xi}}{2^{13}r^4 M_{max}^4\phi_{max}^8}\right) \leq 1 \tag{149}$$

$$\log(2r) - \tfrac{\rho_{min}^4 n^{1-\xi}}{2^{13}r^4 M_{max}^4\phi_{max}^8} \leq 0 \tag{150}$$

$$n^{1-\xi} \geq \tfrac{2^{13}r^4 M_{max}^4\phi_{max}^8}{\rho_{min}^4}\log(2r). \tag{151}$$

To keep ρ_{min} in the bound, we want to show that the left-hand exponential term in Eq. (146) dominates the right-hand term. Thus, we must use Eq. (151) to show a sufficient condition for:

$$2r\exp\left(-\tfrac{\rho_{min}^4 n^{1-\xi}}{2^{13}r^4 M_{max}^4\phi_{max}^8}\right) \geq 2|\mathcal{A}|r^2\exp\left(-\tfrac{nC_{min}^2}{2^5 r^2\phi_{max}^4}\right). \tag{152}$$

We may replace n with $n^{1-\xi}$ on the right-hand side. Since the left-hand term decreases more slowly in n than the right-hand term, we may replace n with the value from Eq. (151):

$$\begin{aligned} 1 &\geq 2|\mathcal{A}|r^2 \exp\left(-\tfrac{2^8 C_{min}^2 r^2 M_{max}^4 \phi_{max}^4}{\rho^4} \log(2r)\right) && (153)\\ |\mathcal{A}| &\leq \tfrac{1}{2r^2} \exp\left(\tfrac{2^8 C_{min}^2 r^2 M_{max}^4 \phi_{max}^4}{\rho^4} \log(2r)\right). && (154)\end{aligned}$$

Since ρ_{min} is a sum of at most M_{max} eigenvalues, and since any eigenvalue is at most $\phi_{max}^2 r$ (as shown in Theorem 4.1), we know $\rho \leq M_{max}\phi_{max}^2 r$. Plugging this in, it suffices to show that:

$$\begin{aligned} |\mathcal{A}| &\leq \tfrac{1}{2r^2} \exp\left(\tfrac{2^8 C_{min}^2 M_{max}^2}{\rho^2} \log(2r)\right) && (155)\\ &= \tfrac{1}{2r^2} (2r)^{\left[\frac{2^8 C_{min}^2 M_{max}^2}{\rho^2}\right]}. && (156)\end{aligned}$$

Therefore, if we have

$$|\mathcal{A}| \leq \tfrac{1}{2r^2} (2r)^{\left[\frac{2^8 C_{min}^2 M_{max}^2}{\rho^2}\right]}, \quad (157)$$

then we know that the probability of failure is at most:

$$4r \exp\left(-\tfrac{\rho_{min}^4 n^{1-\xi}}{2^{13} r^4 M_{max}^4 \phi_{max}^8}\right). \quad \blacksquare \quad (158)$$

Note that this bound is better than that from separate regressions, though both bounds are loose in terms of their treatments of shared parameters.

Proof of Corollary 4.4: If we wish to have a probability of failure of at most δ when we have n samples, we may choose ξ accordingly:

$$\begin{aligned} 2r(|\mathcal{A}|r+1) \exp\left(-\tfrac{C_{min}^4}{2^{13} r^4 M_{max}^4 \phi_{max}^8} n^{1-\xi}\right) &\leq \delta && (159)\\ \log(2r(|\mathcal{A}|r+1)) - \tfrac{C_{min}^4}{2^{13} r^4 M_{max}^4 \phi_{max}^8} n^{1-\xi} &\leq \log\delta && (160)\\ \tfrac{C_{min}^4}{2^{13} r^4 M_{max}^4 \phi_{max}^8} n^{1-\xi} &\geq \log\tfrac{2r(|\mathcal{A}|r+1)}{\delta} && (161)\end{aligned}$$

$$\begin{aligned} n^{1-\xi} &\geq \tfrac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{C_{min}^4} \log\tfrac{2r(|\mathcal{A}|r+1)}{\delta} && (162)\\ (1-\xi)\log n &\geq \log\tfrac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(|\mathcal{A}|r+1)}{\delta} && (163)\\ (1-\xi) &\geq \tfrac{1}{\log n}\left(\log\tfrac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(|\mathcal{A}|r+1)}{\delta}\right) && (164)\\ \xi &\leq 1 - \tfrac{1}{\log n}\left(\log\tfrac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(|\mathcal{A}|r+1)}{\delta}\right) && (165)\\ & && (166)\end{aligned}$$

We will set ξ equal to this upper bound in the next part. Likewise, if we wish to have parameter estimation error at most ϵ, then we need:

$$\begin{aligned} \tfrac{\rho_{min}}{4r M_{max}\phi_{max}^3} n^{-\xi/2} &\leq \epsilon && (167)\\ \log\tfrac{\rho_{min}}{4r M_{max}\phi_{max}^3} - \tfrac{\xi}{2}\log n &\leq \log\epsilon. && (168)\end{aligned}$$

Rewrite the left-hand side:

$$\begin{aligned} &\log\tfrac{\rho_{min}}{4r M_{max}\phi_{max}^3} - \tfrac{\xi}{2}\log n && (169)\\ &= \log\tfrac{\rho_{min}}{4r M_{max}\phi_{max}^3} - \tfrac{1}{2}\left(\log n - \left(\log\tfrac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(|\mathcal{A}|r+1)}{\delta}\right)\right) && (170)\\ &= \tfrac{1}{2}\left[\log\tfrac{\rho_{min}^2}{2^4 r^2 M_{max}^2 \phi_{max}^6} - \log n + \log\tfrac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{C_{min}^4} + \log\log\tfrac{2r(|\mathcal{A}|r+1)}{\delta}\right] && (171)\\ &= \tfrac{1}{2}\left[-\log n + \log\tfrac{2^9 r^2 M_{max}^2 \phi_{max}^2 \rho_{min}^2}{C_{min}^4} + \log\log\tfrac{2r(|\mathcal{A}|r+1)}{\delta}\right]. && (172)\end{aligned}$$

Recombine this left-hand side with Eq. (168):

$$-\log n + \log \frac{2^9 r^2 M_{max}^2 \phi_{max}^2 \rho_{min}^2}{C_{min}^4} + \log\log \frac{2r(|\mathcal{A}|r+1)}{\delta} \quad \leq \quad 2\log\epsilon \tag{173}$$

$$\log n \quad \geq \quad \log \frac{2^9 r^2 M_{max}^2 \phi_{max}^2 \rho_{min}^2}{C_{min}^4} + \log\log \frac{2r(|\mathcal{A}|r+1)}{\delta} - 2\log\epsilon \tag{174}$$

$$n \quad \geq \quad \frac{2^9 r^2 M_{max}^2 \phi_{max}^2 \rho_{min}^2}{C_{min}^4} \frac{1}{\epsilon^2} \log \frac{2r(|\mathcal{A}|r+1)}{\delta} \tag{175}$$

If, however, we assume that $|\mathcal{A}| \leq \frac{1}{2r^2}(2r)^{\left[\frac{2^8 C_{min}^2 M_{max}^2}{\rho^2}\right]}$, then our probability of failure changes, so we require:

$$4r\exp\left(-\frac{\rho_{min}^4 n^{1-\xi}}{2^{13} r^4 M_{max}^4 \phi_{max}^8}\right) \quad \leq \quad \delta \tag{176}$$

$$\log(4r) - \frac{\rho_{min}^4 n^{1-\xi}}{2^{13} r^4 M_{max}^4 \phi_{max}^8} \quad \leq \quad \log\delta \tag{177}$$

$$\frac{\rho_{min}^4 n^{1-\xi}}{2^{13} r^4 M_{max}^4 \phi_{max}^8} \quad \geq \quad \log\frac{4r}{\delta} \tag{178}$$

$$n^{1-\xi} \quad \geq \quad \frac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{\rho_{min}^4} \log\frac{4r}{\delta} \tag{179}$$

$$\xi \quad \leq \quad 1 - \frac{1}{\log n}\left(\log \frac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{\rho_{min}^4} + \log\log\frac{4r}{\delta}\right). \tag{180}$$

Plugging this value for ξ into Eq. (168), we get:

$$\log\frac{\rho_{min}}{4rM_{max}\phi_{max}^3} - \frac{1}{2}\left[1 - \frac{1}{\log n}\left(\log \frac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{\rho_{min}^4} + \log\log\frac{4r}{\delta}\right)\right]\log n \quad \leq \quad \log\epsilon \tag{181}$$

$$\log\frac{\rho_{min}^2}{2^4 r^2 M_{max}^2 \phi_{max}^6} - \log n + \log\frac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{\rho_{min}^4} + \log\log\frac{4r}{\delta} \quad \leq \quad 2\log\epsilon \tag{182}$$

$$\log n \quad \geq \quad \log\frac{\rho_{min}^2}{2^4 r^2 M_{max}^2 \phi_{max}^6} + \log\frac{2^{13} r^4 M_{max}^4 \phi_{max}^8}{\rho_{min}^4} + \log\log\frac{4r}{\delta} + \log\frac{1}{\epsilon^2} \tag{183}$$

$$= \quad \log\frac{2^9 r^2 M_{max}^2 \phi_{max}^2}{\rho_{min}^2} + \log\log\frac{4r}{\delta} + \log\frac{1}{\epsilon^2} \tag{184}$$

$$n \quad \geq \quad \frac{2^9 r^2 M_{max}^2 \phi_{max}^2}{\rho_{min}^2}\frac{1}{\epsilon^2}\log\frac{4r}{\delta}. \quad \blacksquare \tag{185}$$

9.4 Disjoint Optimization

Proof of Lemma 4.6: Let $M_t = |\{i : \theta_t \in \theta_{A_i}\}|$.

$$\left\|\hat{\theta} - \theta^*\right\|_1 \quad = \quad \sum_t \left|\hat{\theta}_t - \theta_t^*\right| \tag{186}$$

$$= \quad \sum_t \left| \frac{1}{M_t} \sum_{i\,:\,\theta_t \in \theta_{A_i}} \hat{\theta}_t^{(A_i)} - {\theta^*}_t \right| \tag{187}$$

$$= \quad \sum_t \frac{1}{M_t} \left| \sum_{i\,:\,\theta_t \in \theta_{A_i}} \hat{\theta}_t^{(A_i)} - {\theta^*}_t \right| \tag{188}$$

$$\leq \quad \sum_t \frac{1}{M_t} \sum_{i\,:\,\theta_t \in \theta_{A_i}} \left|\hat{\theta}_t^{(A_i)} - {\theta^*}_t\right| \tag{189}$$

$$= \quad \sum_i \sum_{t\,:\,\theta_t \in \theta_{A_i}} \frac{1}{M_t} \left|\hat{\theta}_t^{(A_i)} - {\theta^*}_t\right| \tag{190}$$

$$\leq \quad \sum_i \sum_{t\,:\,\theta_t \in \theta_{A_i}} \left|\hat{\theta}_t^{(A_i)} - {\theta^*}_t\right| \tag{191}$$

$$\leq \quad \sum_i \epsilon \tag{192}$$

$$= \quad |\mathcal{A}|\epsilon \quad \blacksquare \tag{193}$$

Proof of Theorem 4.7: Lemma 4.6 shows we can use Corollary 4.2 by shrinking the desired error ϵ and the probability of failure δ by factors of $1/|\mathcal{A}|$. A union bound combines the probabilities of failure. ■

9.5 Bounding the KL with Bounds on Parameter Estimation Error

This subsection uses bounds on the parameter estimation error to bound the log loss of our estimated distribution w.r.t. the target distribution.

The previous theorem demonstrates that there are two convergence regimes. Far from the optimum parameters, the log loss is approximately linear in the parameter estimation error. Close to the optimum, the log loss converges quadratically w.r.t. the parameter estimation error.

We prove two lemmas (for the two regimes) before proving the theorem.

Lemma 9.7. ***(Third-Order Taylor Expansion)*** *Given a CRF factorizing as in Eq. (20) with parameters θ^* and maximum feature magnitude ϕ_{max}, assume that the maximum eigenvalue of the Hessian of the log loss at θ^* is Λ_{max}. Then the expected loss using a vector of parameters θ obeys the following bounds:*

$$\ell_L(\theta) \leq \ell_L(\theta^*) + \frac{\Lambda_{max}}{2}\|\theta-\theta^*\|_1^2 + \phi_{max}^3\|\theta-\theta^*\|_1^3 \tag{194}$$

$$\ell_L(\theta) \leq \ell_L(\theta^*) + \frac{\Lambda_{max}}{2}\|\theta-\theta^*\|_2^2 + \phi_{max}^3 r^{3/2}\|\theta-\theta^*\|_2^3. \tag{195}$$

The second-order term dominates when, respectively,

$$\|\theta-\theta^*\|_1 \leq \frac{\Lambda_{max}}{2\phi_{max}^3} \tag{196}$$

$$\|\theta-\theta^*\|_2 \leq \frac{\Lambda_{max}}{2r^{3/2}\phi_{max}^3}. \tag{197}$$

Proof: Write out the third-order Taylor expansion of the log loss in Eq. (21) w.r.t. θ around θ^*:

$$\begin{aligned}\ell_L(\theta) = {} & \ell_L(\theta^*) \\ & + \tfrac{1}{2}(\theta-\theta^*)^T\left(\nabla^2\ell_L(\theta^*)\right)(\theta-\theta^*) \\ & + \tfrac{1}{6}\sum_i(\theta_i-\theta_i^*)(\theta-\theta^*)^T\left(\tfrac{\partial}{\partial\overline{\theta}_i}\nabla^2\ell_L(\overline{\theta})\big|_{\overline{\theta}=\alpha\theta+(1-\alpha)\theta^*}\right)(\theta-\theta^*),\end{aligned} \tag{198}$$

where $\alpha\in[0,1]$. Note that the first-order term is 0. Let $u=\theta-\theta^*$. The second-order term may be upper-bounded using the maximum eigenvalue of the Hessian:

$$\tfrac{1}{2}u^T\left(\nabla^2\ell_L(\theta^*)\right)u \leq \tfrac{1}{2}\Lambda_{max}\|u\|_2^2 \leq \tfrac{1}{2}\Lambda_{max}\|u\|_1^2. \tag{199}$$

The third-order term may be upper-bounded as well:

$$\tfrac{1}{6}\sum_i u_i u^T\left(\tfrac{\partial}{\partial\overline{\theta}_i}\nabla^2\ell_L(\overline{\theta})\big|_{\overline{\theta}=\alpha\theta+(1-\alpha)\theta^*}\right)u \tag{200}$$

$$= \tfrac{1}{6}\sum_i u_i\Big(\mathbf{E}\left[\phi_i(u^T\phi)^2\right] + 2\mathbf{E}\left[\phi_i\right](\mathbf{E}\left[u^T\phi\right])^2 \tag{201}$$

$$-\mathbf{E}\left[\phi_i\right]\mathbf{E}\left[(u^T\phi)^2\right] - 2\mathbf{E}\left[u^T\phi\right]\mathbf{E}\left[\phi_i(u^T\phi)\right]\Big) \tag{202}$$

$$= \tfrac{1}{6}\left(\mathbf{E}\left[(u^T\phi)^3\right] + 2\mathbf{E}\left[u^T\phi\right](\mathbf{E}\left[u^T\phi\right])^2 - 3\mathbf{E}\left[u^T\phi\right]\mathbf{E}\left[(u^T\phi)^2\right]\right) \tag{203}$$

$$\leq \left(\mathbf{E}\left[|u^T\phi|\right]\right)^3 \tag{204}$$

$$\leq \left(\phi_{max}\|u\|_1\right)^3 \leq \left(\phi_{max}\sqrt{r}\|u\|_2\right)^3. \quad ■ \tag{205}$$

Lemma 9.8. ***(First-Order Taylor Expansion)*** *Given a CRF factorizing as in Eq. (20) with parameters θ^* and maximum feature magnitude ϕ_{max}, the expected loss using a vector of parameters θ obeys the following bounds:*

$$\ell_L(\theta) \leq \ell_L(\theta^*) + \phi_{max}\|\theta-\theta^*\|_1 \tag{206}$$

$$\ell_L(\theta) \leq \ell_L(\theta^*) + \phi_{max}\sqrt{r}\|\theta-\theta^*\|_2. \tag{207}$$

Proof: Write out the first-order Taylor expansion of the log loss in Eq. (21) w.r.t. θ around θ^*:

$$\ell_L(\theta) \quad = \quad \ell_L(\theta^*) + \left(\nabla \ell_L(\overline{\theta})|_{\overline{\theta}=\alpha\theta+(1-\alpha)\theta^*}\right)(\theta - \theta^*), \tag{208}$$

where $\alpha \in [0,1]$. We can upper-bound the first-order term using Holder's inequality:

$$\begin{aligned}
&\left(\nabla \ell_L(\overline{\theta})|_{\overline{\theta}=\alpha\theta+(1-\alpha)\theta^*}\right)(\theta - \theta^*), \\
&= \left(\mathbf{E}_{P(X)}\left[-\mathbf{E}_{P_{\overline{\theta}}(Y|X)}\left[\phi(Y,X)\right] + \mathbf{E}_{P_\theta(Y'|X)}\left[\phi(Y',X)\right]\right]\right)(\theta - \theta^*) && (209)\\
&\leq \left\|\mathbf{E}_{P(X)}\left[-\mathbf{E}_{P_{\overline{\theta}}(Y|X)}\left[\phi(Y,X)\right] + \mathbf{E}_{P_\theta(Y'|X)}\left[\phi(Y',X)\right]\right]\right\|_\infty \left\|\theta - \theta^*\right\|_1 && (210)\\
&\leq \phi_{max}\left\|\theta - \theta^*\right\|_1 \;\leq\; \phi_{max}\sqrt{r}\left\|\theta - \theta^*\right\|_2. \quad \blacksquare && (211)
\end{aligned}$$

Proof of Theorem 4.5: Let $\delta = \left\|\theta - \theta^*\right\|_1$. Suppose the third-order bound is tighter; i.e.,

$$\begin{aligned}
\tfrac{\Lambda_{max}}{2}\delta^2 + \phi_{max}^3\delta^3 \quad &\leq \quad \phi\delta && (212)\\
\phi_{max}^3\delta^2 + \tfrac{\Lambda_{max}}{2}\delta - \phi \quad &\leq \quad 0. && (213)
\end{aligned}$$

Solving, we get

$$\delta \leq \frac{-\frac{\Lambda_{max}}{2} + \sqrt{\frac{\Lambda_{max}^2}{4} + 4\phi_{max}^4}}{2\phi_{max}^3}. \tag{214}$$

Plugging this into the third-order bound, we can rewrite the third-order bound as:

$$\begin{aligned}
\tfrac{\Lambda_{max}}{2}\delta^2 + \phi_{max}^3\delta^3 \quad &\leq \quad \tfrac{\Lambda_{max}}{2}\delta^2 + \frac{-\frac{\Lambda_{max}}{2} + \sqrt{\frac{\Lambda_{max}^2}{4} + 4\phi_{max}^4}}{2}\delta^2 && (215)\\
&= \quad \tfrac{1}{2}\left(\tfrac{\Lambda_{max}}{2} + \sqrt{\tfrac{\Lambda_{max}^2}{4} + 4\phi_{max}^4}\right)\delta^2 && (216)\\
&\leq \quad \tfrac{1}{2}\left(\tfrac{\Lambda_{max}}{2} + \tfrac{\Lambda_{max}}{2} + 2\phi_{max}^2\right)\delta^2 && (217)\\
&\leq \quad \left(\tfrac{\Lambda_{max}}{2} + \phi_{max}^2\right)\delta^2. \quad \blacksquare && (218)
\end{aligned}$$

We discuss sample complexity bounds. The key is to establish a bound on the number of samples required to be in the quadratic convergence regime, after which the proof is trivial. To guarantee that Eq. (15) holds, we can use Corollary 4.4 to show it suffices to have:

$$\begin{aligned}
\tfrac{2^9 r^2 M_{max}^2 \phi_{max}^2}{\rho_{min}^2}\tfrac{1}{n}\log\tfrac{4r}{\delta} \quad &\leq \quad \left[\frac{-\frac{\Lambda_{max}}{2} + \sqrt{\frac{\Lambda_{max}^2}{4} + r\phi_{max}^4}}{2\phi_{max}^3}\right]^2 && (219)\\
&= \quad \frac{\frac{\Lambda_{max}^2}{4} + \frac{\Lambda_{max}^2}{4} + r\phi_{max}^4 - \Lambda_{max}\sqrt{\frac{\Lambda_{max}^2}{4} + r\phi_{max}^4}}{4\phi_{max}^6} && (220)\\
&= \quad \frac{\Lambda_{max}^2 + 2r\phi_{max}^4 - \Lambda_{max}\sqrt{\Lambda_{max}^2 + 4r\phi_{max}^4}}{8\phi_{max}^6} && (221)\\
n \quad &\geq \quad \frac{2^{12} r^2 M_{max}^2 \phi_{max}^8}{\rho_{min}^2\left(\Lambda_{max}^2 + 2r\phi_{max}^4 - \Lambda_{max}\sqrt{\Lambda_{max}^2 + 4r\phi_{max}^4}\right)}\log\tfrac{4r}{\delta} && (222)
\end{aligned}$$

In this quadratic regime, to achieve log loss $\ell_L(\theta^*) + \epsilon$, we need:

$$\left(\tfrac{\Lambda_{max}}{2} + \phi_{max}^2\right)\tfrac{2^9 r^2 M_{max}^2 \phi_{max}^2}{\rho_{min}^2}\tfrac{1}{n}\log\tfrac{4r}{\delta} \leq \epsilon \tag{223}$$

$$n \geq \tfrac{2^8 r^2 M_{max}^2 \phi_{max}^2(\Lambda_{max} + 2\phi_{max}^2)}{\rho_{min}^2}\tfrac{1}{\epsilon}\log\tfrac{4r}{\delta}. \tag{224}$$

Likewise, in the linear regime, we need:

$$n \geq \tfrac{2^9 r^2 M_{max}^2 \phi_{max}^3}{\rho_{min}^2}\tfrac{1}{\epsilon}\log\tfrac{4r}{\delta}. \tag{225}$$

9.6 Canonical Parametrization of Abbeel et al. (2006)

We present the canonical parametrization in its improved version from [17]. We omit proofs already provided by [1, 17]. These two previous works discussed the method in terms of MRFs, but it is trivially generalizable to CRFs. Like our work, the canonical parametrization requires that we know the structure of the target distribution (and that it lies within our model class used for learning).

The canonical parametrization method is based on re-expressing a distribution $P(Y|X)$ as a product of *canonical factors*. Each canonical factor is a ratio of many local conditional probabilities of the form $P(Y_i|Y_{\setminus i}, X)$. In expressing $P(y|x)$, each of these local conditional probabilities is instantiated partly using the values of interest in y and partly using values from a *reference assignment* $\overline{y}$.

Before explicitly stating the canonical parametrization, we define some new notation. We use an asterisk $*$ to mark domains and indices corresponding to canonical factors. 2^C is the powerset of C. For each canonical factor domain $Y_{C_j^*} \subseteq Y$, we define Y_{i_j} to be an arbitrary element in $Y_{C_j^*}$. If A is a variable or set of variables and u, v are assignments to disjoint sets of variables U, V s.t. $A \subseteq U \cup V$, then we write $A[u, v]$ to denote an assignment to variables in A taking values from u, v. We write MB_{Y_i} to denote the Markov blanket of Y_i in Y and X.

The parametrization is based on the following equality (partly proven in [1], completed for MRFs in [17], and extended to CRFs here). The proof for the extension to CRFs simply requires conditioning every probability on X.

Theorem 9.9. *Suppose a CRF factorizes according to factors ϕ_j:*

$$P(Y|X) \propto \prod_{j=1}^{J} \phi_j(Y_{C_j}, X_{D_j}). \tag{226}$$

Let $\overline{y}$ be an arbitrary assignment of values to Y. Define a set of canonical factor domains $\{C_j^*\}_{j=1}^{J^*} \doteq \cup_{j=1}^{J} 2^{C_j} \setminus \emptyset$, We may exactly represent the distribution via the following equality:

$$P(Y|X) = P(\overline{y}|X) \prod_{j=1}^{J^*} \exp\left(\sum_{U \subseteq C_j^*} (-1)^{|C_j^* \setminus U|} \log P\left(Y_{i_j}[Y_U, \overline{y}_{\setminus U}] \,\middle|\, MB_{Y_{i_j}}[Y_U, \overline{y}_{\setminus U}, X] \right) \right). \tag{227}$$

Given this equality, the canonical parametrization method is simple: choose an arbitrary reference assignment $\overline{y}$, use data to compute each local conditional probability $P(Y_i|MB_{Y_i})$, and plug the results into equation Eq. (227).

Note that, in general, each local conditional probability $P\left(Y_{i_j}[Y_U, \overline{y}_{\setminus U}] \,\middle|\, MB_{Y_{i_j}}[Y_U, \overline{y}_{\setminus U}, X] \right)$ may be difficult to compute directly, especially if the Markov blanket is large. For discrete data, the required instantiation $MB_{Y_{i_j}}[\overline{y}_{\setminus U}]$ might not even appear in the dataset. Therefore, we assume that these conditional probabilities are computed via regression (MLE), using the factorization of the target distribution: $\max_P \mathbf{E}_{data}\left[\log P(Y_i|Y_{\setminus i}, X)\right]$, where $P(Y_i|Y_{\setminus i}, X) \propto \prod_{j\,:\,i \in C_j} \phi_j(Y_{C_j}, X_{D_j})$. (We may assume we know the factorization of the target distribution in Eq. (226), for the canonical parametrization PAC bounds require knowing this factorization.) As discussed in Sec. 4.4, we can estimate these probabilities via joint or disjoint optimization. With joint optimization, we compute single estimates of each factor ϕ_j; with disjoint optimization, we estimate ϕ_j once for each $Y_i \in Y_{C_j}$, and we assume that we use the average of these $|C_j|$ estimates.

Note that, like us, [1] compute local conditional probabilities via maximum-likelihood estimates (using disjoint optimization). Since they work with binary variables and low-degree factor graphs, they can compute MLE by computing each $P(Y_i|Y_{\setminus i})$ (in tabular form) via simple counts. [17] do not suggest a method for computing the probabilities, for they only use the canonical parametrization as theoretical motivation for a different algorithm.

To summarize, we use two (very reasonable) assumptions to prove Theorem 5.1, which shows the equivalence of the canonical parametrization and MPLE:

1. In computing the local conditional probabilities in the canonical parametrization, we take advantage of the factorization of the target distribution.
2. We compute local conditional probabilities using either joint optimization (joint MPLE) or disjoint optimization with factor averaging (disjoint MPLE).

Proof of Theorem 5.1: Given our two assumptions, the proof is simple. We need to show that plugging our estimates of local conditional probabilities into the canonical parametrization in Eq. (227) produces the same model as plugging our estimates of factors ϕ_j into the model in Eq. (226). This equivalence is exactly what Theorem 9.9 proves. ■

Clarification: [17] uses probabilities $P(Y_i|\cdot)$ while [1] uses probabilities $P(Y_A|\cdot)$ with $|A| > 1$; it is thus tempting to state that [17] is similar to MPLE while [1] is more similar to MCLE. However, this argument is misleading since the parametrization of [1] can be further simplified into the parametrization of [17]. The two algorithms compute the same set of canonical factors (where each is defined by its domain); the algorithms differ in their parametrizations of these canonical factors. However, both types of canonical factors for a given domain are exactly equal if computed from the same data (Thm. 2.1 from [17]). Thus, the two algorithms are optimizing the same objective. By showing that the method of [17] is equivalent to MPLE, we show that both methods are.