A Computationally Efficient Method for Learning Exponential Family Distributions

Abhin Shah
MIT
abhin@mit.edu

Devavrat Shah
MIT
devavrat@mit.edu

Gregory W. Wornell
MIT
gww@mit.edu

Abstract

We consider the question of learning the natural parameters of a k-parameter *minimal* exponential family from i.i.d. samples in a computationally and statistically efficient manner. We focus on the setting where the support as well as the natural parameters are appropriately bounded. While the traditional maximum likelihood estimator for this class of exponential family is consistent, asymptotically normal, and asymptotically efficient, evaluating it is computationally hard. In this work, we propose a computationally efficient estimator that is consistent as well as asymptotically normal under mild conditions. We provide finite sample guarantees to achieve an (ℓ_2) error of α in the parameter estimation with sample complexity $O(\mathsf{poly}(k/\alpha))$ and computational complexity $O(\mathsf{poly}(k/\alpha))$. To establish these results, we show that, at the population level, our method can be viewed as the maximum likelihood estimation of a re-parameterized distribution belonging to the same class of exponential family. Further, we show that our estimator can be interpreted as a solution to minimizing a particular Bregman score as well as an instance of minimizing the *surrogate* likelihood.

1 Introduction

We are interested in the problem of learning the natural parameters of a *minimal* exponential family with bounded support. Consider a p-dimensional random vector $\mathbf{x} = (x_1, \cdots, x_p)$ with support $\mathcal{X} \subset \mathbb{R}^p$. An exponential family is a set of parametric probability distributions with probability densities of the following canonical form

$$f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) + \beta(\mathbf{x})\right), \tag{1}$$

where $\mathbf{x} \in \mathcal{X}$ is a realization of the underlying random variable $\mathbf{x}$, $\boldsymbol{\theta} \in \mathbb{R}^k$ is the natural parameter, $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^k$ is the natural statistic, k denotes the number of parameters, and β is the log base function. For representational convenience, we shall utilize the following equivalent representation of (1):

$$f_{\mathbf{x}}(\mathbf{x}; \Theta) \propto \exp\left(\left\langle\left\langle \Theta, \Phi(\mathbf{x}) \right\rangle\right\rangle\right) = \exp\left(\sum\nolimits_{i\in[k_1], j\in[k_2], l\in[k_3]} \Theta_{ijl} \times \Phi_{ijl}(\mathbf{x})\right) \tag{2}$$

where $\Theta = [\Theta_{ijl}] \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ is the natural parameter, $\Phi = [\Phi_{ijl}] : \mathcal{X} \to \mathbb{R}^{k_1 \times k_2 \times k_3}$ is the natural statistic, $k_1 \times k_2 \times k_3 - 1 = k$, and $\left\langle\left\langle \Theta, \Phi(\mathbf{x}) \right\rangle\right\rangle$ denotes the tensor inner product, i.e., the sum of product of entries of Θ and $\Phi(\mathbf{x})$. An exponential family is *minimal* if there does not exist a nonzero tensor $\mathbf{U} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ such that $\left\langle\left\langle \mathbf{U}, \Phi(\mathbf{x}) \right\rangle\right\rangle$ is equal to a constant for all $\mathbf{x} \in \mathcal{X}$.

Accepted for publication at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

The notion of exponential family was first introduced by Fisher [17] and was later generalized by Darmois [12], Koopman [30], and Pitman [40]. Exponential families play an important role in statistical inference and arise in many diverse applications for a variety of reasons: (a) they are analytically tractable, (b) they arise as the solutions to several natural optimization problems on the space of probability distributions, (c) they have robust generalization property (see [5, 2] for details).

Truncated (or bounded) exponential family, first introduced by Hogg and Craig [20], is a set of parametric probability distributions resulting from truncating the support of an exponential family. *Truncated* exponential families share the same parametric form with their non-truncated counterparts up to a normalizing constant. These distributions arise in many applications where we can observe only a truncated dataset (truncation is often imposed by during data acquisition) e.g., geolocation tracking data can only be observed up to the coverage of mobile signal, police department can often monitor crimes only within their city's boundary.

The natural parameter Θ specifies a particular distribution in the exponential family. If the natural statistic Φ and the support of $\mathbf{x}$ (i.e., $\mathcal{X}$) are known, then learning a distribution in the exponential family is equivalent to learning the corresponding natural parameter Θ. Despite having a long history, there has been limited progress on learning natural parameter Θ of a *minimal truncated* exponential family. More precisely, there is no known method (without any abstract condition) that is both computationally and statistically efficient for learning natural parameter of the *minimal truncated* exponential family considered in this work.

1.1 Contributions

As the primary contribution of this work, we provide a computationally tractable method with statistical guarantees for learning distributions in *truncated minimal* exponential families. Formally, the learning task of interest is estimating the true natural parameter Θ^* from i.i.d. samples of $\mathbf{x}$ obtained from $f_{\mathbf{x}}(\cdot;\Theta^*)$. We focus on the setting where Θ^* and Φ are appropriately bounded (see Section 2). We summarize our contributions in the following two categories.

1. Computationally Tractable Estimator: Consistency, Normality, Finite Sample Guarantees. Given n samples $\mathbf{x}^{(1)} \cdots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, we propose the following novel loss function to learn a distribution belonging to the exponential family in (2):

$$\mathcal{L}_n(\Theta) = \frac{1}{n}\sum_{t=1}^{n} \exp\big(-\big\langle\big\langle \Theta, \mathit{\Phi}(\mathbf{x}^{(t)})\big\rangle\big\rangle\big), \tag{3}$$

where $\mathit{\Phi}(\cdot) = \Phi(\cdot) - \mathbb{E}_{\mathcal{U}_\mathcal{X}}[\Phi(\cdot)]$ with $\mathcal{U}_\mathcal{X}$ being the uniform distribution over $\mathcal{X}$. We establish that the estimator $\hat{\Theta}_n$ obtained by minimizing $\mathcal{L}_n(\Theta)$ over all Θ in the constraint set Λ, i.e.,

$$\hat{\Theta}_n \in \underset{\Theta\in\Lambda}{\arg\min}\, \mathcal{L}_n(\Theta), \tag{4}$$

is consistent and (under mild further restrictions) asymptotically normal (see Theorem 4.2). We obtain an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ of the convex minimization problem in (4) (i.e., $\mathcal{L}_n(\hat{\Theta}_{\epsilon,n}) \leq \mathcal{L}_n(\hat{\Theta}_n) + \epsilon$) by implementing a projected gradient descent algorithm with $O(\text{poly}(k_1k_2/\epsilon))$[1] iterations (see Lemma 3.1). Finally, we provide rigorous finite sample guarantees for $\hat{\Theta}_{\epsilon,n}$ (with $\epsilon = O(\alpha^2)$) to achieve an error of α (in the tensor ℓ_2 norm) with respect to the true natural parameter Θ^* with $O(\text{poly}(k_1k_2/\alpha))$ samples

[1] We let $k_3 = O(1)$. See Section 2.

and $O(\text{poly}(k_1 k_2/\alpha))$ computations (see Theorem 4.3). By letting certain additional structure on the natural parameter, we allow our framework to capture various constraints on the natural parameter including sparse, low-rank, sparse-plus-low-rank (see Section 2.1).

2. Connections to maximum likelihood estimation (MLE) of a re-parameterized distribution. We establish connections between our method and the MLE of the distribution $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$. We show that the estimator that minimizes the population version of the loss function in (3) i.e.,

$$\mathcal{L}(\Theta) = \mathbb{E}\Big[\exp\big(-\big\langle\big\langle \Theta, \Phi(\mathbf{x})\big\rangle\big\rangle\big)\Big].$$

is equivalent to the estimator that minimizes the Kullback-Leibler (KL) divergence between $\mathcal{U}_{\mathcal{X}}$ (the uniform distribution on $\mathcal{X}$) and $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$ (see Theorem 4.1). Therefore, at the population level, our method can be viewed as the MLE of the parametric family $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$. We show that the KL divergence (and therefore $\mathcal{L}(\Theta)$) is minimized if and only if $\Theta = \Theta^*$, and this connection provides an intuitively pleasing justification of the estimator in (4).

1.2 Related Works

In this section, we look at the related works on learning exponential family. Broadly speaking, there are two line of approaches to overcome the computational hardness of the MLE : (a) approximating the MLE and (b) selecting a surrogate objective. Given the richness of both of approaches, we cannot do justice in providing a full overview. Instead, we look at a few examples from both. Next, we look at some of the related works that focus on learning a class of exponential family. More specifically, we look at works on (a) learning the Gaussian distribution and (b) learning exponential family Markov random fields (MRFs). Finally, we explore some works on the powerful technique of score matching. In Appendix A, we further review works on learning exponential family MRFs, score-based methods (including the related literature on Stein discrepancy) and latent variable graphical models (since these capture sparse-plus-low-rank constraints on the parameters similar to our framework).

Approximating the MLE. Most of the techniques falling in this category approximate the MLE by approximating the log-partition function. A few examples include : (a) approximating the gradient of log-likelihood with a stochastic estimator by minimizing the contrastive divergence [19]; (b) upper bounding the log-partition function by an iterative tree-reweighted belief propagation algorithm [57]; (c) using Monte Carlo methods like importance sampling for estimating the partition function [43]. Since these methods approximate the partition function, they come at the cost of an approximation error or result in a biased estimator.

Selecting surrogate objective. This line of approach selects an easier-to-compute surrogate objective that completely avoids the partition function. A few examples are as follows : (a) pseudo-likelihood estimators [4] approximate the joint distribution with the product of conditional distributions, each of which only represents the distribution of a single variable conditioned on the remaining variables; (b) score matching [22, 21] minimizes the Fisher divergence between the true log density and the model log density. Even though score matching does not require evaluating the partition function, it is computationally expensive as it requires computing third order derivatives for optimization; (c) kernel Stein discrepancy [32, 9] measures the kernel mean discrepancy between a data distribution and a model density using the Stein's identity. This measure is directly

characterized by the choice of the kernel and there is no clear objective for choosing the right kernel [61].

Learning the Gaussian distribution. Learning the Gaussian distribution is a special case of learning exponential family distributions. There has been a long history of learning Gaussian distributions in the form of learning Gaussian graphical models e.g. the neighborhood selection scheme [36], the graphical lasso [18], the CLIME [6], etc. However, finite sample analysis of these methods require various hard-to-verify conditions e.g. the restricted eigenvalue condition, the incoherence assumption ([59, 24]), bounded eigenvalues of the precision matrix, etc. A recent work [28] provided an algorithm whose sample complexity, for a specific subclass of Gaussian graphical models, match the information-theoretic lower bound of [60] without the aforementioned hard-to-verify conditions.

Learning Exponential Family Markov Random Fields (MRFs). MRFs can be naturally represented as exponential family distributions via the principle of maximum entropy (see [58]). A popular method for learning MRFs is estimating node-neighborhoods (fitting conditional distributions of each node conditioned on the rest of the nodes) because the natural parameter is assumed to be node-wise- sparse. A recent line of work has considered a subclass of node-wise-sparse pairwise continuous MRFs where the node-conditional distribution of $x_i \in \mathcal{X}_i$ for every i arise from an exponential family as follows:

$$f_{x_i|x_{-i}}(x_i|\mathbf{x}_{-i} = x_{-i}) \propto \exp\big(\big[\theta_i + \sum_{j\in[p],j\neq i} \theta_{ij}\phi(x_j)\big]\phi(x_i)\big), \tag{5}$$

where $\phi(x_i)$ is the natural statistics and $\theta_i + \sum_{j\in[p],j\neq i} \theta_{ij}\phi(x_j)$ is the natural parameter.[2] Yang et al. [62] showed that only the following joint distribution is consistent with the node-conditional distributions in (5) :

$$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\big(\sum_{i\in[p]} \theta_i\phi(x_i) + \sum_{j\neq i} \theta_{ij}\phi(x_i)\phi(x_j)\big). \tag{6}$$

To learn the node-conditional distribution in (5) for linear $\phi(\cdot)$ (i.e., $\phi(x) = x$), Yang et al. [62] proposed an ℓ_1 regularized node-conditional log-likelihood. However, their finite sample analysis required the following conditions: incoherence, dependency (see [59, 24]), bounded moments of the variables, and local smoothness of the log-partition function. Tansey et al. [51] extended the approach in [62] to vector-space MRFs (i.e., vector natural parameters and natural statistics) and non-linear $\phi(\cdot)$. They proposed a sparse group lasso (see [45]) regularized node-conditional log-likelihood and an alternating direction method of multipliers based approach to solving the resulting optimization problem. However, their analysis required same conditions as [62].

While node-conditional log-likelihood has been a natural choice for learning exponential family MRFs, M-estimation [56, 55, 44] and maximum pseudo-likelihood estimator [39, 63, 10] have recently gained popularity. The objective function in M-estimation is a sample average and the estimator is generally consistent and asymptotically normal. Shah et al. [44] proposed the following M-estimation (inspired from [56, 55]) for vector-space MRFs and non-linear $\phi(\cdot)$: with $\mathcal{U}_{\mathcal{X}_i}$ being the uniform distribution on $\mathcal{X}_i$ and $\tilde{\phi}(x_i) = \phi(x_i) - \int_{x_i'} \phi(x_i')\mathcal{U}_{\mathcal{X}_i}(x_i')dx_i'$

$$\arg\min \frac{1}{n}\sum_{i=1}^{n} \exp\Big(-\big[\theta_i\tilde{\phi}(x_i) + \sum_{j\in[p],j\neq i} \theta_{ij}\tilde{\phi}(x_i)\tilde{\phi}(x_j)\big]\Big). \tag{7}$$

[2] Under node-wise-sparsity, $\sum_{j\in[p],j\neq i} |\theta_{ij}|$ is bounded by a constant for every $i \in [p]$.

They provided an entropic descent algorithm (borrowing from [55]) to solve the optimization in (7) and their finite-sample bounds rely on bounded domain of the variables and a condition (naturally satisfied by linear $\phi(\cdot)$) that lower bounds the variance of a non-constant random variable.

Yuan et al. [64] considered a broader class of sparse pairwise exponential family MRFs compared to [62]. They studied the following joint distribution with natural statistics $\phi(\cdot)$ and $\psi(\cdot)$

$$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\Big(\sum_{i\in[p]} \theta_i \phi(x_i) + \sum_{j\neq i} \theta_{ij}\psi(x_i, x_j)\Big). \tag{8}$$

They proposed an $\ell_{2,1}$ regularized joint likelihood and an $\ell_{2,1}$ regularized node-conditional likelihood. They also presented a Monte-Carlo approximation to these estimators via proximal gradient descent. Their finite-sample analysis required restricted strong convexity (of the Hessian of the negative log-likelihood of the joint density) and bounded moment-generating function of the variables.

Building upon [55] and [44], Ren et al. [41] addressed learning continuous exponential family distributions through a series of numerical experiments. They considered unbounded distributions and allowed for terms corresponding to multi-wise interactions in the joint density. However, they considered only monomial natural statistics. Further, they assume node-wise-sparsity of the parameters as in MRFs and their estimator is defined as a series of node-wise optimization problems.

In summary, tremendous progress has been made on learning the sub-classes of exponential family in (6) and (8). However, this sub-classes are restricted by the assumption that the natural parameters are node-wise-sparse. For example, none of the existing methods for exponential family MRFs work in the setting where the natural parameters have a low-rank constraint.

Score-based method. A scoring rule $S(\mathbf{x}, Q)$ is a numerical score assigned to a realization $\mathbf{x}$ of a random variable $\mathbf{x}$ and it measures the quality of a predictive distribution Q (with probability density $q(\cdot)$). If P is the true distribution of $\mathbf{x}$, the divergence $D(P, Q)$ associated with a scoring rule is defined as $\mathbb{E}_P[S(\mathbf{x}, Q) - S(\mathbf{x}, P)]$. The MLE is an example of a scoring rule with $S(\cdot, Q) = -\log q(\cdot)$ and the resulting divergence is the KL-divergence.

To bypass the intractability of MLE, [22] proposed an alternative scoring rule with $S(\cdot, Q) = \Delta \log q(\cdot) + \frac{1}{2}\|\nabla \log q(\cdot)\|_2^2$ where Δ is the Laplacian operator, ∇ is the gradient and $\|\cdot\|_2$ is the ℓ_2 norm. This method is called *score matching* and the resulting divergence is the Fisher divergence. Score matching is widely used for estimating unnormalizable probability distributions because computing the scoring rule $S(\cdot, Q)$ does not require knowing the partition function. Despite the flexibility of this approach, it is computationally expensive in high dimensions since it requires computing the trace of the unnormalized density's Hessian (and its derivatives for optimization). Additionally, it breaks down for models in which the second derivative grows very rapidly.

In [34], the authors considered estimating truncated exponential family using the principle of score matching. They build on the framework of generalized score matching [21] and proposed a novel estimator that minimizes a weighted Fisher divergence. They showed that their estimator is a special case of minimizing a Stein Discrepancy. However, their finite sample analysis relies on certain hard-to-verify assumptions, for example, the assumption that the optimal parameter is well-separated from other neighboring parameters in terms of their population objective. Further, their estimator lacks the useful properties of asymptotic normality and asymptotic efficiency.

1.3 Useful notations and outline

Notations. For any positive integer t, let $[t] \coloneqq \{1, \cdots, t\}$. For a deterministic sequence $v_1, \cdots, v_t$, we let $\mathbf{v} \coloneqq (v_1, \cdots, v_t)$. For a random sequence $\mathsf{v}_1, \cdots, \mathsf{v}_t$, we let $\mathbf{v} \coloneqq (\mathsf{v}_1, \cdots, \mathsf{v}_t)$. For a matrix $\mathbf{M} \in \mathbb{R}^{u \times v}$, we denote the element in i^{th} row and j^{th} column by M_{ij}, the singular values of the matrix by $\sigma_i(\mathbf{M})$ for $i \in [\min\{u, v\}]$, the matrix maximum norm by $\|\mathbf{M}\|_{\max} \coloneqq \max_{i \in [u], j \in [v]} |M_{ij}|$, the entry-wise $L_{1,1}$ norm by $\|\mathbf{M}\|_{1,1} \coloneqq \sum_{i \in [u], j \in [v]} |M_{ij}|$, the nuclear norm by $\|\mathbf{M}\|_\star \coloneqq \sum_{i \in [\min\{u,v\}]} \sigma_i(\mathbf{M})$. We denote the Frobenius or Trace inner product of matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{u \times v}$ by $\langle \mathbf{M}, \mathbf{N} \rangle \coloneqq \sum_{i \in [u], j \in [v]} M_{ij} N_{ij}$. For a matrix $\mathbf{M} \in \mathbb{R}^{u \times v}$, we denote a generic norm on $\mathbb{R}^{u \times v}$ by $\mathcal{R}(\mathbf{M})$ and denote the associated dual norm by $\mathcal{R}^*(\mathbf{M}) \coloneqq \sup\{\langle \mathbf{M}, \mathbf{N} \rangle | \mathcal{R}(\mathbf{N}) \leq 1\}$ where $\mathbf{N} \in \mathbb{R}^{u \times v}$. For a tensor $\mathbf{U} \in \mathbb{R}^{u \times v \times w}$, we denote its (i, j, l) entry by U_{ijl}, its l^{th} slice (obtained by fixing the last index) by $U_{::l}$ or $U^{(l)}$, the tensor maximum norm (with a slight abuse of notation) by $\|\mathbf{U}\|_{\max} \coloneqq \max_{i \in [u], j \in [v], l \in [w]} |U_{ijl}|$, and the tensor norm by $\|\mathbf{U}\|_{\mathrm{T}} \coloneqq \sqrt{\sum_{i \in [u], j \in [v], l \in [w]} U_{ijl}^2}$. We denote the tensor inner product of tensors $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{u \times v \times w}$ by $\langle\langle \mathbf{U}, \mathbf{V} \rangle\rangle \coloneqq \sum_{i \in [u], j \in [v], l \in [w]} U_{ijl} V_{ijl}$. We denote the vectorization of the tensor $\mathbf{U} \in \mathbb{R}^{u \times v \times w}$ by $\mathrm{vec}(\mathbf{U}) \in \mathbb{R}^{uvw \times 1}$ (the ordering of the elements is not important as long as it is consistent). Let $\mathbf{0} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ denote the tensor with every entry zero. We denote a p-dimensional ball of radius b centered at 0 by $\mathcal{B}(0, b)$.

Outline. In Section 2, we formulate the problem of interest, state our assumptions, and provide examples. In Section 3, we provide our loss function and algorithm. In Section 4, we present our main results including the connections to the MLE of $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$, consistency, asymptotic normality, and finite sample guarantees. In Section 5, we conclude, provide some remarks, discuss limitations as well as some directions for future work. See supplementary for organization of the Appendix.

2 Problem Formulation

Let $\mathbf{x} = (\mathsf{x}_1, \cdots, \mathsf{x}_p)$ be a $p-$dimensional vector of continuous random variables.[3] For any $i \in [p]$, let the support of x_i be $\mathcal{X}_i \subset \mathbb{R}$. Define $\mathcal{X} \coloneqq \prod_{i=1}^p \mathcal{X}_i$. Let $\mathbf{x} = (x_1, \cdots, x_p) \in \mathcal{X}$ be a realization of $\mathbf{x}$. In this work, we assume that the random vector $\mathbf{x}$ belongs to an exponential family with bounded support (i.e., length of $\mathcal{X}_i$ is bounded) along with certain additional constraints. More specifically, we make certain assumptions on the natural parameter $\Theta \in \mathbb{R}^{k_1 \times k_2 \times k_3}$, and on the natural statistic $\Phi(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^{k_1 \times k_2 \times k_3}$ as follows.

Natural parameter Θ. We focus on natural parameters with bounded norms. However, instead of having such constraints on the natural parameter Θ as it is, we decompose Θ into k_3 slices (or matrices) and have slice specific constraints. The key motivation for this is to broaden the class of exponential family covered by our formulation. For example, this decomposability allows our formulation to en-capture the sparse-plus-low-rank decomposition of Θ in addition to only sparse or only low-rank decompositions of Θ (see Section 2.1). This is precisely the reason for considering tensor natural parameters instead of matrix natural parameters. Further, we assume $k_3 = O(1)$ i.e., it does not scale with p. We formally state this assumption below.

Assumption 2.1. *(Bounded norms of Θ.) For every $i \in [k_3]$, we let $\mathcal{R}_i(\Theta^{(i)}) \leq r_i$ where $\Theta^{(i)} \in \mathbb{R}^{k_1 \times k_2}$ is the i^{th} slice of Θ, $\mathcal{R}_i : \mathbb{R}^{k_1 \times k_2} \to \mathbb{R}_+$ is a norm and r_i is a known constant. This decomposition is represented compactly by $\boldsymbol{\mathcal{R}}(\Theta) \leq \boldsymbol{r}$ where $\boldsymbol{\mathcal{R}}(\Theta) = (\mathcal{R}_1(\Theta^{(1)}), \cdots, \mathcal{R}_{k_3}(\Theta^{(k_3)}))$ and $\boldsymbol{r} = (r_1, \cdots, r_{k_3})$.*

[3] Even though we focus on continuous variables, our framework applies equally to discrete variables.

We define Λ to be the set of all natural parameters satisfying Assumption 2.1 i.e., $\Lambda \coloneqq \{\Theta : \boldsymbol{\mathcal{R}}(\Theta) \leq \boldsymbol{r}\}$. For any $\tilde{\Theta}, \bar{\Theta} \in \Lambda$ and $t \in [0,1]$, we have $\boldsymbol{\mathcal{R}}(t\tilde{\Theta}+(1-t)\bar{\Theta}) \leq t\boldsymbol{\mathcal{R}}(\tilde{\Theta})+(1-t)\boldsymbol{\mathcal{R}}(\bar{\Theta}) \leq t\boldsymbol{r}+(1-t)\boldsymbol{r} = \boldsymbol{r}$. Therefore, $t\tilde{\Theta}+(1-t)\bar{\Theta} \in \Lambda$ and the constraint set Λ is a convex set.

Natural Statistic Φ. For mathematical simplicity, we center the natural statistic $\Phi(\cdot)$ such that their integral with respect to the uniform density on $\mathcal{X}$ (i.e., $\mathcal{U}_{\mathcal{X}}$) is zero. $\mathcal{U}_{\mathcal{X}}$ is well-defined because the support $\mathcal{X}$ is a strict subset of $\mathbb{R}^p$ i.e., $\mathcal{X} \subset \mathbb{R}^p$.

Definition 2.1. *(Centered natural statistics). The centered natural statistics are defined as follows:*

$$\Phi(\cdot) \coloneqq \Phi(\cdot) - \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\Phi(\mathbf{x})].$$

In this work, we focus on bounded natural statistics which may enforce certain restrictions on the length of support $\mathcal{X}$. See Section 2.1 for examples. We define two notions of boundedness. First, we make the following assumption to be able to bound the tensor inner product between the natural parameter Θ and the centered natural statistic $\Phi(\cdot)$ (see Appendix B.1).

Assumption 2.2. *(Bounded dual norms of Φ). For every $i \in [k_3]$ and norm $\mathcal{R}_i$, we assume that the dual norm $\mathcal{R}_i^*$ of the i^{th} slice of the centered natural statistic i.e., $\Phi^{(i)}$ is bounded by a constant d_i. Formally, for any $i \in [k_3]$ and $\mathbf{x} \in \mathcal{X}$, $\mathcal{R}_i^*(\Phi^{(i)}(\mathbf{x})) \leq d_i$. This is represented compactly by $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq \boldsymbol{d}$ where $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) = (\mathcal{R}_1^*(\Phi^{(1)}(\mathbf{x})), \cdots, \mathcal{R}_{k_3}^*(\Phi^{(k_3)}(\mathbf{x})))$ and $\boldsymbol{d} = (d_1, \cdots, d_{k_3})$.*

Next, we assume that the tensor maximum norm of the centered natural statistic $\Phi(\cdot)$ is bounded by a constant $\phi_{\max}$. This assumption is stated formally below.

Assumption 2.3. *(Bounded tensor maximum norm of Φ). For any $\mathbf{x} \in \mathcal{X}$, $\|\Phi(\mathbf{x})\|_{\max} \leq \phi_{\max}$.*

The Exponential Family. Summarizing, $\mathbf{x}$ belongs to a *minimal truncated* exponential family with probability density function as follows

$$f_{\mathbf{x}}(\mathbf{x}; \Theta) \propto \exp\Big(\Big\langle\Big\langle\Theta, \Phi(\mathbf{x})\Big\rangle\Big\rangle\Big). \tag{9}$$

where the natural parameter $\Theta \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ is such that $\boldsymbol{\mathcal{R}}(\Theta) \leq \boldsymbol{r}$ and the natural statistic $\Phi(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^{k_1 \times k_2 \times k_3}$ is such that for any $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq \boldsymbol{d}$ and $\|\Phi(\mathbf{x})\|_{\max} \leq \phi_{\max}$.

Let Θ^* denote the true natural parameter of interest and $f_{\mathbf{x}}(\mathbf{x}; \Theta^*)$ denote the true distribution of $\mathbf{x}$. Naturally, we assume $\boldsymbol{\mathcal{R}}(\Theta^*) \leq \boldsymbol{r}$. Formally, the learning task of interest is as follows:

Goal. (Natural Parameter Recovery). Given n independent samples of $\mathbf{x}$ i.e., $\mathbf{x}^{(1)} \cdots, \mathbf{x}^{(n)}$ obtained from $f_{\mathbf{x}}(\mathbf{x}; \Theta^*)$, compute an estimate $\hat{\Theta}$ of Θ^* in polynomial time such that $\|\Theta^* - \hat{\Theta}\|_{\mathrm{T}}$ is small.

2.1 Examples

We will first present examples of natural parameters that satisfy Assumption 2.1. Next, we will present examples of natural statistics along with the corresponding support that satisfy Assumptions 2.2, and 2.3. See Appendix H and I for more discussion on these examples.

Examples of natural parameter. We provide examples in Table 1 to illustrate the decomposability of Θ as in Assumption 2.1. We will revisit these examples briefly in

Section 4 and in-depth in Appendix H. Assumption 2.1 should be viewed as a potential flexibility in the problem specification i.e., a practitioner has the option to choose from a variety of constraints on the natural parameters (that could be handled by our framework). For example, in some real-world applications the parameters are sparse while in some other real-world applications the parameters have a low-rank and a practitioner could choose either depending on the application at hand. For the sparse-plus-low-rank decomposition,

Table 1: A few examples of natural parameter Θ.

Decomposition	k_3	Convex Relaxation
Sparse decomposition ($\Theta^* = (\Theta^{*(1)})$)	1	$\|\Theta^{*(1)}\|_{1,1} \leq r_1$
Low-rank decomposition ($\Theta^* = (\Theta^{*(1)})$)	1	$\|\Theta^{*(1)}\|_\star \leq r_1$
Sparse-plus-low-rank decomposition ($\Theta^* = (\Theta^{*(1)}, \Theta^{*(2)})$)	2	$\|\Theta^{*(1)}\|_{1,1} \leq r_1$ and $\|\Theta^{*(2)}\|_\star \leq r_2$

it is more natural to think about the *minimality* of the exponential family in terms of matrices as opposed to tensors. See Appendix I for details.

Examples of natural statistic. The following are a few example of natural statistics (along with the corresponding support) that fall in-line with Assumptions 2.2 and 2.3.

1. *Polynomial statistics*: Suppose the natural statistics are polynomials of $\mathbf{x}$ with maximum degree l, i.e., $\prod_{i\in[p]} x_i^{l_i}$ such that $l_i \geq 0\ \forall i \in [p]$ and $\sum_{i\in[p]} l_i \leq l$. If $\mathcal{X} = [0, b]$ for $b \in \mathbb{R}$, then $\phi_{\max} = 2b^l$. If Θ^* has a sparse decomposition and $\mathcal{X} = [0, b]$ for $b \in \mathbb{R}$, then $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq 2b^k$. Further, if Θ^* has a low-rank decomposition, $l = 2$, and $\mathcal{X} = \mathcal{B}(0, b)$ for $b \in \mathbb{R}$, then $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq 2(1 + b^2)$. Finally, if Θ^* has a sparse-plus-low-rank decomposition, $l = 2$, and $\mathcal{X} = \mathcal{B}(0, b)$ for $b \in \mathbb{R}$, then $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq (2b^2, 2 + 2b^2)$.
2. *Trigonometric statistics*: Suppose the natural statistics are sines and cosines of $\mathbf{x}$ with l different frequencies, i.e., $\sin(\sum_{i\in[p]} l_i x_i) \cup \cos(\sum_{i\in[p]} l_i x_i)$ such that $l_i \in [l] \cup \{0\}$. For any $\mathcal{X} \subset \mathbb{R}^p$, $\phi_{\max} = 2$. If Θ^* has a sparse decomposition, then $\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq 2$ for any $\mathcal{X} \subset \mathbb{R}^p$.

Our framework also allows combinations of polynomial and trigonometric statistics (see Appendix I).[4]

3 Algorithm

We propose a novel, computationally tractable loss function drawing inspiration from the recent advancements in exponential family Markov Random Fields [56, 55, 44].

The loss function and the estimator. The loss function, defined below, is an empirical average of the inverse of the function of $\mathbf{x}$ that the probability density $f_{\mathbf{x}}(\mathbf{x}; \Theta)$ is proportional to (see (9)).

Definition 3.1 (The loss function)**.** *Given n samples $\mathbf{x}^{(1)} \cdots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, the loss function maps $\Theta \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ to $\mathcal{L}_n(\Theta) \in \mathbb{R}$ defined as*

$$\mathcal{L}_n(\Theta) = \frac{1}{n} \sum_{t=1}^{n} \exp\big(- \big\langle\big\langle \Theta, \Phi(\mathbf{x}^{(t)}) \big\rangle\big\rangle\big). \tag{10}$$

[4] We believe that for polynomial and/or trigonometric natural statistics, Assumptions 2.2 and 2.3 would hold whenever the domain of $\mathcal{X}$ is appropriately bounded.

The proposed estimator $\hat{\Theta}_n$ produces an estimate of Θ^* by minimizing the loss function $\mathcal{L}_n(\Theta)$ over all natural parameters Θ satisfying Assumption 2.1 i.e.,

$$\hat{\Theta}_n \in \arg\min_{\Theta \in \Lambda} \mathcal{L}_n(\Theta). \tag{11}$$

For any $\epsilon > 0$, $\hat{\Theta}_{\epsilon,n}$ is an ϵ-optimal solution of $\hat{\Theta}_n$ if $\mathcal{L}_n(\hat{\Theta}_{\epsilon,n}) \leq \mathcal{L}_n(\hat{\Theta}_n) + \epsilon$. The optimization in (11) is a convex minimization problem (i.e., minimizing a convex function $\mathcal{L}_n$ over a convex set Λ) and has efficient implementations for finding an ϵ-optimal solution. Although alternative algorithms (including Frank-Wolfe) can be used, we provide a projected gradient descent algorithm below.

Algorithm 1: Projected Gradient Descent

Input: η, τ, Λ
Output: $\hat{\Theta}_{\epsilon,n}$
Initialization: $\Theta_{(0)} = \mathbf{0}$
for $t = 0, \cdots, \tau$ **do**
  $\Theta_{(t+1)} \leftarrow \arg\min_{\Theta \in \Lambda} \|\Theta_{(t)} - \eta \nabla \mathcal{L}_n(\Theta_{(t)}) - \Theta\|_{\mathrm{T}}$
$\hat{\Theta}_{\epsilon,n} \leftarrow \Theta_{(\tau+1)}$

The following Lemma shows that running sufficient iterations of the projected gradient descent in Algorithm 1 results in an ϵ-optimal solution of $\hat{\Theta}_n$.

Lemma 3.1. *Let Assumptions 2.1, 2.2 and 2.3 be satisfied. Let* $\eta = 1/k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})$. *Then, Algorithm 1 returns an* ϵ*-optimal solution* $\hat{\Theta}_{\epsilon,n}$ *as long as*

$$\tau \geq \frac{2 k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})}{\epsilon} \|\hat{\Theta}_n\|_{\mathrm{T}}^2. \tag{12}$$

Further, ignoring the dependence on k_3, $\phi_{\max}$, $\boldsymbol{r}$ *and* $\boldsymbol{d}$, τ *in* (12) *scales as* $O\big(\mathrm{poly}\big(\frac{k_1 k_2}{\epsilon}\big)\big)$.

The proof of Lemma 3.1 can be found in Appendix B. The proof outline is as follows : (a) First, we prove the smoothness property of $\mathcal{L}_n(\Theta)$. (b) Next, we complete the proof using a standard result from convex optimization for the projected gradient descent algorithm for smooth functions.

4 Analysis and Main results

In this section, we provide our analysis and main results. First, we focus on the connection between our method and the MLE of $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$. Then, we establish consistency and asymptotic normality of our estimator. Finally, we provide non-asymptotic finite sample guarantees to recover Θ^*.

1. Connection with MLE of $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$. First, we will establish a connection between the population version of the loss function in (10) (denoted by $\mathcal{L}(\Theta)$) and the KL-divergence of the uniform density on $\mathcal{X}$ with respect to $f_{\mathbf{x}}(\mathbf{x}; \Theta^* - \Theta)$. Then, using *minimality* of the exponential family, we will show that this KL-divergence and $\mathcal{L}(\Theta)$ are minimized if and only if $\Theta = \Theta^*$. This provides a justification for the estimator in (11) as well as helps us obtain consistency and asymptotic normality of $\hat{\Theta}_n$.

For any $\Theta \in \Lambda$, $\mathcal{L}(\Theta) = \mathbb{E}\Big[\exp\big(-\big\langle\big\langle \Theta, \Phi(\mathbf{x})\big\rangle\big\rangle\big)\Big]$. The following result shows that the population version of the estimator in (11) is equivalent to the maximum likelihood estimator of $f_{\mathbf{x}}(\mathbf{x}; \Theta^* - \Theta)$.

Theorem 4.1. *With $D(\cdot \parallel \cdot)$ representing the KL-divergence,*

$$\arg\min_{\Theta\in\Lambda} \mathcal{L}(\Theta) = \arg\min_{\Theta\in\Lambda} D(\mathcal{U}_{\mathcal{X}}(\cdot) \parallel f_{\mathbf{x}}(\cdot;\Theta^* - \Theta)).$$

Further, the true parameter Θ^* is the unique minimizer of $\mathcal{L}(\Theta)$.

The proof of Theorem 4.1 can be found in Appendix C. The proof outline is as follows : (a) First, we express $f_{\mathbf{x}}(\cdot;\Theta^* - \Theta)$ in terms of $\mathcal{L}(\Theta)$ (b) Next, we complete the proof by simplifying the KL-divergence between $\mathcal{U}_{\mathcal{X}}(\cdot)$ and $f_{\mathbf{x}}(\cdot;\Theta^* - \Theta)$.

2. Consistency and Normality. We establish consistency and asymptotic normality of the proposed estimator $\hat{\Theta}_n$ by invoking the asymptotic theory of M-estimation. We emphasize that, from Theorem 4.1, the population version of $\hat{\Theta}_n$ is equivalent to the maximum likelihood estimate of $f_{\mathbf{x}}(\cdot;\Theta^* - \Theta)$ and not $f_{\mathbf{x}}(\cdot;\Theta)$. Moreover, there is no clear connection between $\hat{\Theta}_n$ and the finite sample maximum likelihood estimate of $f_{\mathbf{x}}(\cdot;\Theta)$ or $f_{\mathbf{x}}(\cdot;\Theta^* - \Theta)$. Therefore, we cannot invoke the asymptotic theory of MLE to show consistency and asymptotic normality of $\hat{\Theta}_n$.

Let $A(\Theta^*)$ denote the covariance matrix of $\text{vec}\big(\Phi(\mathbf{x}) \exp\big(- \big\langle\big\langle \Theta^*, \Phi(\mathbf{x})\big\rangle\big\rangle\big)\big)$. Let $B(\Theta^*)$ denote the cross-covariance matrix of $\text{vec}(\Phi(\mathbf{x}))$ and $\text{vec}(\Phi(\mathbf{x}) \exp\big(-\big\langle\big\langle \Theta^*, \Phi(\mathbf{x})\big\rangle\big\rangle\big))$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the multi-variate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Theorem 4.2. *Let Assumptions 2.1, 2.2, and 2.3 be satisfied. Let $\hat{\Theta}_n$ be a solution of (11). Then, as $n \to \infty$, $\hat{\Theta}_n \xrightarrow{p} \Theta^*$. Further, assuming $\Theta^* \in$ interior(Λ) and $B(\Theta^*)$ is invertible, we have $\sqrt{n} \times vec(\hat{\Theta}_n - \Theta^*) \xrightarrow{d} \mathcal{N}(vec(\mathbf{0}), B(\Theta^*)^{-1} A(\Theta^*) B(\Theta^*)^{-1})$.*

The proof of Theorem 4.2 can be found in Appendix D. The proof is based on two key observations : (a) $\hat{\Theta}_n$ is an M-estimator and (b) $\mathcal{L}(\Theta)$ is uniquely minimized at Θ^*.

3. Finite Sample Guarantees. To provide the non-asymptotic guarantees for recovering Θ^*, we require the following assumption on the smallest eigenvalue of the autocorrelation matrix of $\text{vec}(\Phi(\mathbf{x}))$.

Assumption 4.1. *(Positive eigenvalue of the autocorrelation matrix of Φ.) Let $\lambda_{\min}$ denote the minimum eigenvalue of $\mathbb{E}_{\mathbf{x}}[vec(\Phi(\mathbf{x}))vec(\Phi(\mathbf{x}))^T]$. We assume $\lambda_{\min}$ is strictly positive i.e., $\lambda_{\min} > 0$.*

We also make use of the following property of the matrix norms.

Property 4.1. *For any norm $\tilde{\mathcal{R}} : \mathbb{R}^{k_1 \times k_2} \to \mathbb{R}_+$, and matrix $\mathbf{M} \in \mathbb{R}^{k_1 \times k_2}$, there exists g such that $\tilde{\mathcal{R}}(\mathbf{M}) \leq g k_1 k_2 \|\mathbf{M}\|_{\max}$.*

For most matrix norms of interest including entry-wise $L_{p,q}$ norm ($p, q \geq 1$), Schatten p-norm ($p \geq 1$), and operator $p-$norm ($p \geq 1$), we have $g = 1$ as shown in Appendix J.

Let $\boldsymbol{g} = (g_1, \cdots, g_{k_3})$ where $\forall i \in [k_3], g_i$ is such that $\mathcal{R}_i^*(\mathbf{M}) \leq g_i k_1 k_2 \|\mathbf{M}\|_{\max}$ with $\mathcal{R}_i^*$ being the dual norms from Assumption 2.2.

Theorem 4.3 below shows that, with enough samples, the ϵ-optimal solution of $\hat{\Theta}_n$ is close to the true natural parameter in the tensor norm with high probability.

Theorem 4.3. *Let $\hat{\Theta}_{\epsilon,n}$ be an ϵ-optimal solution of $\hat{\Theta}_n$ obtained from Algorithm 1 for ϵ of the order $O(\alpha^2 \lambda_{\min})$. Let Assumptions 2.1, 2.2, 2.3, and 4.1 be satisfied. Recall Property 4.1. Then, for any $\delta \in (0,1)$, we have $\|\hat{\Theta}_{\epsilon,n} - \Theta^*\|_{\mathrm{T}} \leq \alpha$ with probability at least $1 - \delta$ as long as*

$$n \geq O\bigg(\frac{k_1^2 k_2^2}{\alpha^4 \lambda_{\min}^2} \log\Big(\frac{k_1 k_2}{\delta}\Big)\bigg). \tag{13}$$

The computational cost scales as $O\Big(\frac{k_1k_2}{\alpha^2}\max\big(k_1k_2n, c(\Lambda)\big)\Big)$ *where* $c(\Lambda)$ *is the cost of projection onto* Λ*. Further, ignoring the dependence on* δ*,* $\lambda_{\min}$*, and* $c(\Lambda)$*,* n *in* (13) *(as well as the associated computational cost) scales as* $O\big(\text{poly}\big(\frac{k_1k_2}{\alpha}\big)\big)$*.*

The proof of Theorem 4.3 can be found in Appendix G. The proof is based on two key properties of the loss function $\mathcal{L}_n(\Theta)$: (a) with enough samples, the loss function $\mathcal{L}_n(\Theta)$ naturally obeys the restricted strong convexity with high probability and (b) with enough samples, $\|\nabla\mathcal{L}_n(\Theta^*)\|_{\max}$ is bounded with high probability. See the proof for the dependence of the sample complexity and the computational complexity on $k_3, \boldsymbol{r}, \boldsymbol{d}, \boldsymbol{g}$ and $\phi_{\max}$.

The computational cost of projection onto Λ i.e., $c(\Lambda)$ is typically polynomial in k_1k_2. In Appendix H, we provide the computational cost for the example constraints on the natural parameter Θ from Section 2.1 i.e., sparse decomposition, low-rank decomposition, and sparse-plus-low-rank decomposition.

4. Comparison with the traditional MLE. To contextualize our method, we compare it with the MLE of the parametric family $f_{\mathbf{x}}(\cdot;\Theta)$. The MLE of $f_{\mathbf{x}}(\cdot;\Theta)$ minimizes the following loss function

$$\min -\frac{1}{n}\sum_{t=1}^{n}\big\langle\big\langle\Theta, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle + \log\int_{\mathbf{x}\in\mathcal{X}}\exp\big(\big\langle\big\langle\Theta, \Phi(\mathbf{x}\big\rangle\big\rangle\big)d\mathbf{x}. \tag{14}$$

The maximum likelihood estimator has many attractive asymptotic properties : (a) consistency (see [16, Theorem 17]), i.e., as the sample size goes to infinity, the bias in the estimated parameters goes to zero, (b) asymptotic normality (see [16, Theorem 18]), i.e., as the sample size goes to infinity, normalized estimation error coverges to a Gaussian distribution and (c) asymptotic efficiency (see [16, Theorem 20]), i.e., as the sample size goes to infinity, the variance in the estimation error attains the minimum possible value among all consistent estimators. Despite having these useful asymptotic properties of consistency, normality, and efficiency, computing the maximum likelihood estimator is computationally hard [52, 26].

Our method can be viewed as a computationally efficient proxy for the MLE. More precisely, our method is computationally tractable as opposed to the MLE while retaining the useful properties of consistency and asymptotic normality. However, our method misses out on asymptotic efficiency. This raises an important question for future work — *can computational and asymptotic efficiency be achieved by a single estimator for this class of exponential family?*

5 Conclusion, Remarks, Limitations, Future Work

In this section, we conclude, provide a few remarks, discuss the limitations of our work as well as some interesting future directions.

Conclusion. In this work, we provide a computationally and statistically efficient method to learn distributions in a *minimal truncated* k-parameter exponential family from i.i.d. samples. We propose a novel estimator via minimizing a convex loss function and obtain consistency and asymptotic normality of the same. We provide rigorous finite sample analysis to achieve an α-approximation to the true natural parameters with $O(\text{poly}(k/\alpha))$ samples and $O(\text{poly}(k/\alpha))$ computations. We also provide an interpretation of our estimator in terms of a maximum likelihood estimation.

Node-wise-sparse exponential family MRFs vs general exponential family. We highlight that the focus of our work is beyond the exponential families associated with node-wise-sparse MRFs and towards general exponential families. The former focuses on local assumptions on the parameters such as node-wise-sparsity and the sample complexity depends logarithmically on the parameter dimension i.e., $O(\log(k))$. In contrast, our work can handle global structures on the parameters (e.g., a low-rank constraint) and there are no prior work that can handle such global structures with sample complexity $O(\log(k))$. Similarly, for node-wise-sparse MRFs there has been a lot of work to relax the assumptions required for learning (see the discussion on Assumption 4.1 below). Since our work focuses on global structures associated with the parameters, we leave the question of relaxing the assumptions required for learning as an open question. Likewise, the interaction screening objective [56] and generalized interaction screening objective [55, 44] were designed for node-wise parameter estimation i.e., they require the parameters to be node-wise-sparse and are less useful when the parameters have a global structure. On the contrary, our loss function is designed to accommodate global structures on the parameters.

Assumption 4.1. For node-wise-sparse pairwise exponential family MRFs (e.g., Ising models), which is a special case of the setting considered in our work, Assumption 4.1 is proven (e.g., Appendix T.1 of [44] provides one such analysis for a condition that is equivalent to Assumption 4.1 for sparse continuous graphical model). However, such analysis typically requires (a) a bound on the infinity norm of the parameters and a bound on the degree of each node or (b) a bound on the ℓ_1 norm of the parameters associated with each node. Since the focus of our work is beyond the exponential families associated with node-wise-sparse MRFs, we view Assumption 4.1 as an adequate condition to rule out certain singular distributions (as evident in the proof of Proposition E.1 where this condition is used to effectively lower bounds the variance of a non-constant random variable) and expect it to hold for most real-world applications. Further, we highlight that the MLE in (14) remains computationally intractable even under Assumption 4.1. To see this, one could again focus on node-wise-sparse pairwise exponential family MRFs where Assumption 4.1 is proven and the MLE is still known to be computationally intractable.

Sample Complexity. We do not assume p (the dimension of $\mathbf{x}$) to be a constant and think of k_1 and k_2 as implicit functions of p. Typically, for an exponential family, the quantity of interest is the number of parameters i.e., k and this quantity scales polynomially in p e.g., $k = O(p^2)$ for Ising model, $k = O(p^t)$ for t-wise MRFs over binary alphabets. Therefore, in this scenario, the dependence of the sample complexity on p would also be $O(\text{poly}(p))$. Further, the $1/\alpha^4$ dependence of the sample complexity seems fundamental to our loss function. For learning node-wise-sparse MRFs, this dependence is in-line with some prior works that use a similar loss function [44, 55] as well as that do not use a similar loss function [29]. While it is known that for learning node-wise-sparse MRFs [56] and truncated Gaussian [13] one could achieve a better dependence of $1/\alpha^2$, it is not yet clear how the lower bound on the sample complexity would depend on α for the general class of exponential families considered in this work (which may not be sparse or Gaussian).

Practicality of Algorithm 1. While the optimization associated with Algorithm 1 is a convex minimization problem (i.e., (11)) and the computational complexity of Algorithm 1 is polynomial in the parameter dimension and the error tolerance, computing the gradient of the loss function requires centering of the natural statistics (see (26)). If the natural statistics are polynomials or trigonometric, centering them should be relatively straightforward (since the integrals would have closed-form expressions). In other cases, centering them may not be polynomial-time and one might require an assumption of

computationally efficient sampling or that obtaining approximately random samples of $\mathbf{x}$ is computationally efficient [14].

Limitations and Future Work. First, in our current framework, we assume boundedness of the support. While, conceptually, most non-compact distributions could be truncated by introducing a controlled amount of error, we believe this assumption could be lifted as for exponential families: $\mathbb{P}(|x_i| \geq \delta \log \gamma) \leq c\gamma^{-\delta}$ where $c > 0$ is a constant and $\gamma > 0$. Alternatively, the notion of multiplicative regularizing distribution from [41] could also be used. Second, while the population version of our estimator has a nice interpretation in terms of maximum likelihood estimation, the finite sample version of our estimator does not have a similar interpretation. We believe there could be connections with the Bregman score and this is an important direction for immediate future work. Third, while our estimator is computationally efficient, consistent, and asymptotically normal, it is not asymptotically efficient. Investigating the possibility of a single estimator that achieves computational and asymptotic efficiency for this class of exponential family could be an interesting future direction. Lastly, building on our framework, empirical study is an important direction for future work.

Acknowledgements

This work was supported, in part, by NSR under Grant No. CCF-1816209, ONR under Grant No. N00014-19-1-2665, the NSF TRIPODS Phase II grant towards Foundations of Data Science Institute, the MIT-IBM project on time series anomaly detection, and the KACST project on Towards Foundations of Reinforcement Learning.

References

[1] T. Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[2] O. Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

[3] A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. *arXiv preprint arXiv:1906.08283*, 2019.

[4] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

[5] L. D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

[6] T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[8] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.

[9] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR, 2016.

[10] Y. Dagan, C. Daskalakis, N. Dikkala, and A. V. Kandiros. Learning ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 161–168, 2021.

[11] B. Dai, Z. Liu, H. Dai, N. He, A. Gretton, L. Song, and D. Schuurmans. Exponential family estimation via adversarial dynamics embedding. *arXiv preprint arXiv:1904.12083*, 2019.

[12] G. Darmois. Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85, 1935.

[13] C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE, 2018.

[14] I. Diakonikolas, D. M. Kane, A. Stewart, and Y. Sun. Outlier-robust learning of ising models under dobrushin's condition. *arXiv preprint arXiv:2102.02171*, 2021.

[15] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

[16] T. S. Ferguson. *A course in large sample theory*. Routledge, 2017.

[17] R. A. Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 144(852):285–307, 1934.

[18] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[19] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[20] R. V. Hogg and A. T. Craig. Sufficient statistics in elementary distribution theory. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 17(3):209–216, 1956.

[21] A. Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

[22] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[23] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[24] A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 378–387, 2011.

[25] R. I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643, 04 1969.

[26] M. Jerrum and A. Sinclair. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

[27] S. Kakade, O. Shamir, K. Sindharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388. JMLR Workshop and Conference Proceedings, 2010.

[28] J. Kelner, F. Koehler, R. Meka, and A. Moitra. Learning some popular gaussian graphical models without condition number bounds. 2019.

[29] A. R. Klivans and R. Meka. Learning graphical models using multiplicative weights. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 343–354, 2017.

[30] B. O. Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

[31] L. Lin, M. Drton, and A. Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electronic journal of statistics*, 10(1):806, 2016.

[32] Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[33] S. Liu, T. Kanamori, W. Jitkrittum, and Y. Chen. Fisher efficient inference of intractable models. *Advances in Neural Information Processing Systems*, 32:8793–8803, 2019.

[34] S. Liu, T. Kanamori, and D. J. Williams. Estimating density models with truncation boundaries. *arXiv preprint arXiv:1910.03834*, 2019.

[35] B. Meghana and N. He. Lower bounds & projected gradient descent.

[36] N. Meinshausen, P. Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2006.

[37] Z. Meng, B. Eriksson, and A. Hero. Learning latent variable gaussian graphical models. In *International Conference on Machine Learning*, pages 1269–1277. PMLR, 2014.

[38] S. Na, M. Kolar, and O. Koyejo. Estimating differential latent variable graphical models with applications to brain connectivity. *arXiv preprint arXiv:1909.05892*, 2019.

[39] Y. Ning, T. Zhao, H. Liu, et al. A likelihood ratio framework for high-dimensional semiparametric regression. *Annals of Statistics*, 45(6):2299–2327, 2017.

[40] E. J. G. Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press, 1936.

[41] C. X. Ren, S. Misra, M. Vuffray, and A. Y. Lokhov. Learning continuous exponential families beyond gaussian, 2021.

[42] B. Rhodes, K. Xu, and M. U. Gutmann. Telescoping density-ratio estimation. *arXiv preprint arXiv:2006.12204*, 2020.

[43] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[44] A. Shah, D. Shah, and G. Wornell. On learning continuous pairwise markov random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 1153–1161. PMLR, 2021.

[45] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[46] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18, 2017.

[47] H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton. Gradient-free hamiltonian monte carlo with efficient kernel exponential families. *arXiv preprint arXiv:1506.02564*, 2015.

[48] A. S. Suggala, M. Kolar, and P. Ravikumar. The expxorcist: Nonparametric graphical models via conditional exponential densities. In *Advances in Neural Information Processing Systems*, pages 4446–4456, 2017.

[49] S. Sun, M. Kolar, and J. Xu. Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems*, pages 2287–2295, 2015.

[50] D. Sutherland, H. Strathmann, M. Arbel, and A. Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 652–660. PMLR, 2018.

[51] W. Tansey, O. H. M. Padilla, A. S. Suggala, and P. Ravikumar. Vector-space markov random fields via exponential families. In *International Conference on Machine Learning*, pages 684–692, 2015.

[52] L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

[53] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[54] G. Vinci, V. Ventura, M. A. Smith, and R. E. Kass. Adjusted regularization in latent graphical models: Application to multiple-neuron spike count data. *The annals of applied statistics*, 12(2):1068, 2018.

[55] M. Vuffray, S. Misra, and A. Y. Lokhov. Efficient learning of discrete graphical models. *CoRR*, abs/1902.00600, 2019.

[56] M. Vuffray, S. Misra, A. Y. Lokhov, and M. Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[57] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching. In *AISTATS*, volume 3, page 3, 2003.

[58] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[59] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems*, pages 1465–1472, 2006.

[60] W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE, 2010.

[61] L. Wenliang, D. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746. PMLR, 2019.

[62] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, 16:3813–3847, 2015.

[63] Z. Yang, Y. Ning, and H. Liu. On semiparametric exponential family graphical models. *The Journal of Machine Learning Research*, 19(1):2314–2372, 2018.

[64] X. Yuan, P. Li, T. Zhang, Q. Liu, and G. Liu. Learning additive exponential family graphical models via $\ell_{2,1}$-norm regularized m-estimation. In *Advances in Neural Information Processing Systems*, pages 4367–4375, 2016.

Appendix

Organization. In Appendix A, we provide additional discussion on exponential family Markov random fields, score-based methods, as well as review the related literature on Stein discrepancy and latent variable graphical models. In Appendix B, we state and prove the smoothness property of the loss function as well as provide the proof of Lemma 3.1. In Appendix C, we provide the proof of Theorem 4.1. In Appendix D, we provide the proof of Theorem 4.2. In Appendix E, we provide the restricted strong convexity property of the loss function. In Appendix F, we provide bounds on the tensor maximum norm of the gradient of the loss function evaluated at the true natural parameter. In Appendix G, we provide the proof of Theorem 4.3. In Appendix H, we provide the computational cost for the example constraints on the natural parameter Θ. In Appendix I, we provide a discussion on the examples of natural parameter and natural statistics from Section 2.1. In Appendix J, we provide a discussion on Property 4.1.

Additional Notations. We denote the ℓ_p norm ($p \geq 1$) of a vector $\mathbf{v} \in \mathbb{R}^t$ by $\|\mathbf{v}\|_p := (\sum_{i=1}^t |v_i|^p)^{1/p}$ and its ℓ_∞ norm by $\|\mathbf{v}\|_\infty := \max_{i\in[t]} |v_i|$. For a matrix $\mathbf{M} \in \mathbb{R}^{u\times v}$, we denote the spectral norm by $|||\mathbf{M}||| := \max_{i\in[\min\{u,v\}]} \sigma_i(\mathbf{M})$ and the Frobenius norm by $\|\mathbf{M}\|_{\mathrm{F}} := \sqrt{\sum_{i\in[u],j\in[v]} M_{ij}^2}$. For a tensor $\mathbf{U} \in \mathbb{R}^{u\times v\times w}$, we let $\|\mathbf{U}\|_{1,1,1} := \sum_{i\in[u],j\in[v],l\in[w]} |U_{ijl}|$.

A Related Works

In this Section, we review additional works on exponential family Markov random fields, score-based methods, as well as the related literature on Stein discrepancy and latent variable graphical models.

A.1 Exponential Family Markov Random Fields

Having reviewed some of the works on sparse exponential family MRFs in Section 1.2, we present here a brief overview of a few other works on the same.

Following the lines of [62], the authors in [48] proposed an ℓ_1 regularized node-conditional log-likelihood to learn the node-conditional density in (5) for non-linear $\phi(\cdot)$. They used an alternating minimization technique and proximal gradient descent to solve the resulting optimization problem. However, their analysis required restricted strong convexity, bounded domain of the variables, non-negative node parameters, and hard-to-verify assumptions on gradient of the population loss.

In [63], the authors introduced a non-parametric component to the node-conditional density in (5) while focusing on linear $\phi(\cdot)$. More specifically, they focused on the following joint density:

$$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\Big(\sum_{i\in[p]} \eta_i(x_i) + \sum_{j\neq i} \theta_{ij} x_i x_j\Big),$$

where $\eta_i(\cdot)$ is the non-parametric node-wise term. They proposed a node-conditional pseudo-likelihood (introduced in [39]) regularized by a non-convex penalty and an adaptive multi-stage convex relaxation method to solve the resulting optimization problem. However, their finite-sample bounds require bounded moments of the variables, sparse eigenvalue condition on their loss function, and local smoothness of the log-partition function. In [49], the authors investigated infinite dimensional sparse pairwise exponential family MRFs where they assumed that the node and edge potentials lie in a Reproducing Kernel Hilbert space (RKHS). They used a penalized version of the score matching

objective of [22]. However, their finite-sample analysis required incoherence and dependency conditions (see [59, 24]). In [31], the authors considered the joint distribution in (8) restricting the variables to be non-negative. They proposed a group lasso regularized generalized score matching objective [21] which is a generalization of the score matching objective [22] to non-negative data. However, their finite-sample analysis required the incoherence condition.

A.2 Score-based and Stein discrepancy methods

Having mentioned the principle behind and an example for the score-based method in Section 1.2, we briefly review a few other score-based methods in relation to the Stein discrepancy.

Stein discrepancy is a quantitative measure of how well a predictive density $q(\cdot)$ fits the density of interest $p(\cdot)$ based on the classical Stein's identity. Stein's identity defines an infinite number of identities indexed by a critic function f and does not require evaluation of the partition function like the score matching method. By focusing on Stein discrepancy constructed from a RKHS, the authors in [32] and [9] independently proposed the kernel Stein discrepancy as a test statistic to access the goodness-of-fit for unnormalized densities. The authors in [32] and [3] showed that the Fisher divergence, which was the minimization criterion used by the score matching method, can be viewed a special case of the kernel Stein discrepancy with a specific, fixed critic function f. In [3], the authors showed that a few other methods (including the contrastive divergence by [19]) can also be viewed as a kernel Stein discrepancy with respect to a different class of critics. Despite the kernel Stein discrepancy being a natural criterion for fitting computationally hard models, there is no clear objective for choosing the right kernel and the kernels typically chosen (e.g. [49, 47, 46, 50]) are insufficient for complex datasets as pointed out by [61].

In [11], the authors exploited the primal-dual view of the MLE to avoid estimating the normalizing constant at the price of introducing dual variables to be jointly estimated. They showed that many other methods including the contrastive divergence by [19], pseudo-likelihood by [4], score matching by [22] and minimum Stein discrepancy estimator by [32], [9], and [3] are special cases of their estimator. However, this method results in expensive optimization problems since they rely on adversarial optimization (see [42] for details). In [33], the authors proposed an inference method for unnormalized models known as discriminative likelihood estimator. This estimator follows the KL divergence minimization criterion and is implemented via density ratio estimation and a Stein operator. However, this method requires certain hard-to-verify conditions.

A.3 Literature on Latent Variable Graphical Models

In recent years, sparse-plus-low-rank matrix recovery has received considerable attention in machine learning and statistical inference, e.g., robust PCA [7], latent variable graphical models [8]. Latent variable graphical models has a variety of applications including assessing the functional interactions between neurons recorded from two brain areas [54, 38]. In latent variable graphical models, there are variables not present in observations. The presence of such variables leads to a challenge in learning the graphical model. The graphical model corresponding to the conditional distribution of the observed variables conditioned on the latent variables is in general different from the graphical model corresponding to the marginal distribution of the observed variables. The marginal graphical model consists of dependencies that are induced due to marginalization over the latent variables and typically consists of many more edges than the conditional graphical model. In [8], authors considered latent variable Gaussian graphical models and exploited the observation that the precision matrix of the marginal graphical model

can be decomposed into the superposition of a sparse matrix and a low-rank matrix. They provided a tractable convex program based on regularized maximum-likelihood to estimate the precision matrix. While the authors in [8] focused on simultaneous model selection consistency of both the sparse and low-rank components, the authors in [37] focused on estimating the precision matrix of latent variable Gaussian graphical model. They consider a regularized MLE estimator and utilize the *almost strong convexity* [27] of the log-likelihood to derive non-asymptotic error bounds under the restricted Fisher eigenvalue and Structural Fisher Incoherence assumptions. Compared to [37], our tensor norm error bounds are derived under mild condition. Additionally, our framework captures various constraints on the natural parameters in addition to the sparse-plus-low-rank constraint.

B Smoothness of the loss function and proof of Lemma 3.1

In this Section, we will prove the smoothness of $\mathcal{L}_n(\cdot)$ as well as prove Lemma 3.1. However, before either of this, we provide bounds on the absolute tensor inner product between Θ and Φ i.e., $\left|\left\langle\left\langle \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle\right|$ for $\Theta \in \Lambda$ and $\mathbf{x} \in \mathcal{X}$.

B.1 Bounds on the absolute tensor inner product between Θ and Φ.

We have

$$\left|\left\langle\left\langle \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle\right| \overset{(a)}{=} \left|\sum_{i=1}^{k_3} \left\langle \Theta^{(i)}, \Phi^{(i)}(\mathbf{x})\right\rangle\right| \overset{(b)}{\leq} \sum_{i=1}^{k_3} \left|\left\langle \Theta^{(i)}, \Phi^{(i)}(\mathbf{x})\right\rangle\right| \overset{(c)}{\leq} \sum_{i=1}^{k_3} \mathcal{R}_i(\Theta^{(i)}) \times \mathcal{R}_i^*(\Phi^{(i)}(\mathbf{x}))$$

$$\overset{(d)}{\leq} \boldsymbol{r}^T \boldsymbol{d}, \tag{15}$$

where (a) follows from the definitions of a slice of a tensor, tensor inner product, and Frobenius inner product, (b) follows from the triangle inequality, (c) follows from the definition of a dual norm, and (d) follows from Assumptions 2.1 and 2.2.

B.2 Smoothness of the loss function

Now, we will state and prove our result for smoothness of $\mathcal{L}_n(\Theta)$.

Proposition B.1. *Under Assumptions 2.1, 2.2 and 2.3, $\mathcal{L}_n(\Theta)$ is a $k_1k_2k_3\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})$ smooth function of Θ.*

Proof of Proposition B.1. To show $k_1k_2k_3\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})$ smoothness of $\mathcal{L}_n(\Theta)$, we will show that the largest eigenvalue of the Hessian[5] of $\mathcal{L}_n(\Theta)$ is upper bounded by $k_1k_2k_3\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})$.

First, we simplify the Hessian of $\mathcal{L}_n(\Theta)$ i.e., $\nabla^2\mathcal{L}_n(\Theta)$. The component of the Hessian of $\mathcal{L}_n(\Theta)$ corresponding to $\Theta_{u_1v_1w_1}$ and $\Theta_{u_2v_2w_2}$ for $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$ and $w_1, w_2 \in [k_3]$ is given by

$$\frac{\partial^2\mathcal{L}_n(\Theta)}{\partial\Theta_{u_1v_1w_1}\partial\Theta_{u_2v_2w_2}} = \frac{1}{n}\sum_{t=1}^{n} \Phi_{u_1v_1w_1}(\mathbf{x}^{(t)})\Phi_{u_2v_2w_2}(\mathbf{x}^{(t)}) \exp\Big(-\Big\langle\Big\langle \Theta, \Phi(\mathbf{x}^{(t)})\Big\rangle\Big\rangle\Big). \tag{16}$$

From the Gershgorin circle theorem, we know that the largest eigenvalue of any matrix is upper bounded by the largest absolute row sum or column sum. Let $\lambda_{\max}(\nabla^2\mathcal{L}_n(\Theta))$

[5]Ideally, one would consider the Hessian of $\mathcal{L}_n(\text{vec}(\Theta))$. However, for the ease of the exposition we abuse the terminology.

denote the largest eigenvalue of $\nabla^2 \mathcal{L}_n(\Theta)$. We have the following

$$\lambda_{\max}(\nabla^2 \mathcal{L}_n(\Theta)) \leq \max_{u_2,v_2,w_2} \sum_{u_1,v_1,w_1} \Big| \frac{\partial^2 \mathcal{L}_n(\Theta)}{\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}} \Big| \overset{(a)}{\leq} \max_{u_2,v_2,w_2} \sum_{u_1,v_1,w_1} \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d}) \\ \leq k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d}),$$

where (a) follows from (16), (15), and Assumption 2.3. Therefore, $\mathcal{L}_n(\Theta)$ is a $k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})$ smooth function of Θ. □

B.3 Proof of Lemma 3.1

Next, we restate the Lemma 3.1 and provide the proof.

Lemma 3.1. *Let Assumptions 2.1, 2.2 and 2.3 be satisfied. Let* $\eta = 1/k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})$. *Then, Algorithm 1 returns an* ϵ*-optimal solution* $\hat{\Theta}_{\epsilon,n}$ *as long as*

$$\tau \geq \frac{2 k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})}{\epsilon} \|\hat{\Theta}_n\|_{\mathrm{T}}^2. \tag{12}$$

Further, ignoring the dependence on k_3, $\phi_{\max}$, $\boldsymbol{r}$ *and* $\boldsymbol{d}$, τ *in* (12) *scales as* $O\big(\mathrm{poly}\big(\frac{k_1 k_2}{\epsilon}\big)\big)$.

Proof of Lemma 3.1. Let us recall Theorem 10.6 from [35].

[35, Theorem 10.6]: Let L be a c-smooth convex function of a parameter vector $\theta \in \Lambda$. Consider the following constrained optimization problem

$$\min_{\theta \in \Lambda} L(\theta). \tag{17}$$

Let θ^* be an optimal solution of (17). Let $\theta^{(1)}, \cdots, \theta^{(t)}$ denote the iterates of the projected gradient descent algorithm with step size $\eta = 1/c$. Let $\theta^{(0)}$ denote the initialization of θ in the projected gradient descent algorithm. Then,

$$L(\theta^{(t)}) - L(\theta^*) \leq \frac{2c}{t} \|\theta^{(0)} - \theta^*\|_2^2. \tag{18}$$

We will make direct use of this theorem in our proof. From Proposition B.1, $\mathcal{L}_n(\Theta)$ is $c_1 \coloneqq k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})$ smooth. Using (18), we have

$$\mathcal{L}_n(\Theta_{(\tau)}) - \mathcal{L}_n(\hat{\Theta}_n) \leq \frac{2c_1}{\tau} \|\Theta_{(0)} - \hat{\Theta}_n\|_{\mathrm{T}}^2.$$

Plugging in $c_1 = k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})$, $\tau = \dfrac{2 k_1 k_2 k_3 \phi_{\max}^2 \exp(\boldsymbol{r}^T \boldsymbol{d})}{\epsilon} \|\hat{\Theta}_n\|_{\mathrm{T}}^2$, and $\Theta_{(0)} = \mathbf{0}$ we have

$$\mathcal{L}_n(\Theta_{(\tau)}) - \mathcal{L}_n(\hat{\Theta}_n) \leq \epsilon.$$

Therefore, $\Theta_{(\tau)}$ is an ϵ-optimal solution.

We will now upper bound $\|\hat{\Theta}_n\|_{\mathrm{T}}^2$. First let us upper bound this tensor norm in terms of tensor maximum norm and therefore the matrix maximum norms. We have

$$\|\hat{\Theta}_n\|_{\mathrm{T}}^2 \leq k_1 k_2 k_3 \|\hat{\Theta}_n\|_{\max}^2 = k_1 k_2 k_3 \max_{i \in [k_3]} \|\hat{\Theta}_n^{(i)}\|_{\max}^2.$$

Now, observe that most matrix norms of interest including the entry-wise $L_{p,q}$ norm ($p, q \geq 1$), the Schatten p-norm ($p \geq 1$), and the operator p-norm ($p \geq 1$) are bounded from below

by the matrix maximum norm i.e., the matrix maximum norm is upper bounded if either of these matrix norms are upper bounded. Suppose $\forall i \in [k_3]$, $\mathcal{R}_i$ is either the entry-wise $L_{p,q}$ norm ($p, q \geq 1$), the Schatten p-norm ($p \geq 1$), or the operator p-norm ($p \geq 1$). Then, $\forall i \in [k_3]$, $\|\hat{\Theta}_n^{(i)}\|_{\max} \leq \mathcal{R}_i(\hat{\Theta}_n^{(i)})$. We have $\mathcal{R}_i(\hat{\Theta}_n^{(i)}) \leq r_i$ from Assumption 2.1 because $\hat{\Theta}_n^{(i)} \in \Lambda$. Therefore, we have

$$\|\hat{\Theta}_n\|_{\mathrm{T}}^2 \leq k_1 k_2 k_3 \max_{i \in [k_3]} r_i^2.$$

Summarizing and using the fact that $\phi_{\max}, \boldsymbol{r}, \boldsymbol{d}, k_3$ are $O(1)$, we have

$$\frac{2k_1k_2k_3\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})}{\epsilon}\|\hat{\Theta}_n\|_{\mathrm{T}}^2 \leq \frac{2k_1^2k_2^2k_3^2\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})}{\epsilon} \max_{i \in [k_3]} r_i^2 = O\left(\frac{k_1^2k_2^2}{\epsilon}\right).$$

□

C Proof of Theorem 4.1

In this Section, we prove Theorem 4.1. We restate the Theorem below and then provide the proof.

Theorem 4.1. *With $D(\cdot \parallel \cdot)$ representing the KL-divergence,*

$$\arg\min_{\Theta \in \Lambda} \mathcal{L}(\Theta) = \arg\min_{\Theta \in \Lambda} D(\mathcal{U}_{\mathcal{X}}(\cdot) \parallel f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)).$$

Further, the true parameter Θ^* is the unique minimizer of $\mathcal{L}(\Theta)$.

Proof of Theorem 4.1. We will first express $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$ in terms of $\mathcal{L}(\Theta)$. We have

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}; \Theta^* - \Theta) &= \frac{\exp\left(\left\langle\left\langle \Theta^* - \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle\right)}{\int_{\mathbf{y} \in \mathcal{X}} \exp\left(\left\langle\left\langle \Theta^* - \Theta, \Phi(\mathbf{y})\right\rangle\right\rangle\right) d\mathbf{y}} \overset{(a)}{=} \frac{\exp\left(\left\langle\left\langle \Theta^* - \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle\right)}{\int_{\mathbf{y} \in \mathcal{X}} \exp\left(\left\langle\left\langle \Theta^* - \Theta, \Phi(\mathbf{y})\right\rangle\right\rangle\right) d\mathbf{y}} \\ &\overset{(b)}{=} \frac{f_{\mathbf{x}}(\mathbf{x}; \Theta^*) \exp\left(-\left\langle\left\langle \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle\right)}{\int_{\mathbf{y} \in \mathcal{X}} f_{\mathbf{x}}(\mathbf{x}; \Theta^*) \exp\left(-\left\langle\left\langle \Theta, \Phi(\mathbf{y})\right\rangle\right\rangle\right) d\mathbf{y}} \\ &\overset{(c)}{=} \frac{f_{\mathbf{x}}(\mathbf{x}; \Theta^*) \exp\left(-\left\langle\left\langle \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle\right)}{\mathcal{L}(\Theta)}, \end{aligned} \tag{19}$$

where (a) follows because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\Phi(\mathbf{x})]$ is a constant, (b) follows by dividing the numerator and the denominator by the constant $\int_{\mathbf{y} \in \mathcal{X}} \exp\left(\left\langle\left\langle \Theta^*, \Phi(\mathbf{y})\right\rangle\right\rangle\right) d\mathbf{y}$ and using the definition of $f_{\mathbf{x}}(\mathbf{x}; \Theta^*)$, and (c) follows from definition of $\mathcal{L}(\Theta)$. We will now simplify the KL-divergence between $\mathcal{U}_{\mathcal{X}}(\cdot)$ and $f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)$.

$$\begin{aligned} D(\mathcal{U}_{\mathcal{X}}(\cdot) \parallel f_{\mathbf{x}}(\cdot; \Theta^* - \Theta)) &\overset{(a)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)\mathcal{L}(\Theta)}{f_{\mathbf{x}}(\cdot; \Theta^*) \exp\left(-\left\langle\left\langle \Theta, \Phi(\cdot)\right\rangle\right\rangle\right)}\right)\right] \\ &\overset{(b)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)}{f_{\mathbf{x}}(\cdot; \Theta^*)}\right)\right] + \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\left\langle\left\langle \Theta, \Phi(\cdot)\right\rangle\right\rangle\right] + \log \mathcal{L}(\Theta) \\ &\overset{(c)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)}{f_{\mathbf{x}}(\cdot; \Theta^*)}\right)\right] + \left\langle\left\langle \Theta, \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\Phi(\cdot)]\right\rangle\right\rangle + \log \mathcal{L}(\Theta) \\ &\overset{(d)}{=} \mathbb{E}_{\mathcal{U}_{\mathcal{X}}}\left[\log\left(\frac{\mathcal{U}_{\mathcal{X}}(\cdot)}{f_{\mathbf{x}}(\cdot; \Theta^*)}\right)\right] + \log \mathcal{L}(\Theta), \end{aligned}$$

where (a) follows from (19) and the definition of KL-divergence, (b) follows because $\log(abc) = \log a + \log b + \log c$ and $\mathcal{L}(\Theta)$ is a constant, (c) follows from the linearity

of the expectation and (d) follows because $\mathbb{E}_{\mathcal{U}_\mathcal{X}}[\Phi(\mathbf{x})] = 0$ from Definition 2.1. Observing that the first term in the above equation is not dependent on Θ, we can write

$$\arg\min_{\Theta\in\Lambda} D(\mathcal{U}_\mathcal{X}(\cdot) \parallel f_\mathbf{x}(\cdot;\Theta^* - \Theta)) = \arg\min_{\Theta\in\Lambda} \log \mathcal{L}(\Theta) \overset{(a)}{=} \arg\min_{\Theta\in\Lambda} \mathcal{L}(\Theta),$$

where (a) follows because log is a monotonic function. Further, the KL-divergence between $\mathcal{U}_\mathcal{X}(\cdot)$ and $f_\mathbf{x}(\cdot;\Theta^* - \Theta)$ is minimized when $\mathcal{U}_\mathcal{X}(\cdot) = f_\mathbf{x}(\cdot;\Theta^* - \Theta)$. Recall that the natural statistic are such that the exponential family is minimal. Therefore, $\mathcal{U}_\mathcal{X}(\cdot) = f_\mathbf{x}(\cdot;\Theta^* - \Theta)$ if and only if $\Theta = \Theta^*$. Thus, $\Theta^* \in \arg\min_{\Theta\in\Lambda} \mathcal{L}(\Theta)$, and it is a unique minimizer of $\mathcal{L}(\Theta)$. □

D Proof of Theorem 4.2

In this Section, we prove Theorem 4.2 by using the theory of M-estimation. In particular, observe that $\hat{\Theta}_n$ is an M-estimator i.e., $\hat{\Theta}_n$ is a sample average. Therefore, we invoke Theorem 4.1.1 and Theorem 4.1.3 of [1] to prove the consistency and normality of $\hat{\Theta}_n$. We restate the Theorem below and then provide the proof.

Theorem 4.2. *Let Assumptions 2.1, 2.2, and 2.3 be satisfied. Let $\hat{\Theta}_n$ be a solution of* (11)*. Then, as $n \to \infty$, $\hat{\Theta}_n \xrightarrow{p} \Theta^*$. Further, assuming $\Theta^* \in$ interior(Λ) and $B(\Theta^*)$ is invertible, we have $\sqrt{n} \times vec(\hat{\Theta}_n - \Theta^*) \xrightarrow{d} \mathcal{N}(vec(\mathbf{0}), B(\Theta^*)^{-1}A(\Theta^*)B(\Theta^*)^{-1})$.*

Proof of Theorem 4.2. We divide the proof in two parts.

Consistency. We will first show that $\hat{\Theta}_n$ is asymptotically consistent. In order to show this, let us recall Theorem 4.1.1 of [1].

[1, Theorem 4.1.1]: Let $z_1, \cdots, z_n$ be i.i.d. samples of a random variable z. Let $q(z;\theta)$ be some function of z parameterized by $\theta \in \Upsilon$. Let θ^* be the true underlying parameter. Define

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} q(z_i;\theta) \qquad \text{and} \qquad \hat{\theta}_n \in \arg\min_{\theta\in\Upsilon} Q_n(\theta).$$

Let the following be true.

(a) Υ is compact,
(b) $Q_n(\theta)$ converges uniformly in probability to a non-stochastic function $Q(\theta)$,
(c) $Q(\theta)$ is continuous, and
(d) $Q(\theta)$ is uniquely minimized at θ^*.

Then, $\hat{\theta}_n$ is consistent for θ^* i.e., $\hat{\theta}_n \xrightarrow{p} \theta^*$ as $n \to \infty$.

Letting $z \coloneqq \mathbf{x}$, $\theta \coloneqq \Theta$, $\hat{\theta}_n \coloneqq \hat{\Theta}_n$, $\theta^* \coloneqq \Theta^*$, $\Upsilon = \Lambda$, $q(z;\theta) \coloneqq \exp\big(-\big\langle\big\langle \Theta, \Phi(\mathbf{x})\big\rangle\big\rangle\big)$, and $Q_n(\theta) \coloneqq \mathcal{L}_n(\Theta)$, it is sufficient to show the following:

(a) Λ is compact,
(b) $\mathcal{L}_n(\Theta)$ converges uniformly in probability to a non-stochastic function $\mathcal{L}(\Theta)$,
(c) $\mathcal{L}(\Theta)$ is continuous, and
(d) $\mathcal{L}(\Theta)$ is uniquely minimized at Θ^*.

Let us show these one by one.

(a) We have $\Lambda = \{\Theta : \boldsymbol{\mathcal{R}}(\Theta) \leq \boldsymbol{r}\}$ which is bounded and closed. Therefore, Λ is compact.

(b) Recall [25, Theorem 2]: Let $z_1, \cdots, z_n$ be i.i.d. samples of a random variable z. Let $g(\mathsf{z}; \theta)$ be a function of θ parameterized by $\theta \in \Upsilon$. Then, $n^{-1} \sum_t g(z_t, \theta)$ converges uniformly in probability to $\mathbb{E}[g(\mathsf{z}, \theta)]$ if

(i) Υ is compact,

(ii) $g(\mathsf{z}, \theta)$ is continuous at each $\theta \in \Upsilon$ with probability one,

(iii) $g(\mathsf{z}, \theta)$ is dominated by a function $G(\mathsf{z})$ i.e., $|g(\mathsf{z}, \theta)| \leq G(\mathsf{z})$, and

(iv) $\mathbb{E}[G(\mathsf{z})] < \infty$.

Using this theorem with $\mathsf{z} \coloneqq \mathbf{x}$, $\theta \coloneqq \Theta$, $\Upsilon \coloneqq \Lambda$, $g(\mathsf{z}, \theta) \coloneqq \exp\big(-\big\langle\big\langle \Theta, \Phi(\mathbf{x})\big\rangle\big\rangle\big)$, $G(\mathsf{z}) \coloneqq \exp(\boldsymbol{r}^T \boldsymbol{d})$ and (15), we conclude that $\mathcal{L}_n(\Theta)$ converges to $\mathcal{L}(\Theta)$ uniformly in probability.

(c) $\exp\big(-\big\langle\big\langle \Theta, \Phi(\mathbf{x})\big\rangle\big\rangle\big)$ is a continuous function of $\Theta \in \Lambda$. Further, $f_{\mathbf{x}}(\mathbf{x}; \Theta^*)$ does not functionally depend on Θ. Therefore, we have continuity of $\mathcal{L}(\Theta)$ for all $\Theta \in \Lambda$.

(d) From Theorem 4.1, $\mathcal{L}(\Theta)$ is uniquely minimized at Θ^*.

Therefore, we have asymptotic consistency of $\hat{\Theta}_n$.

Normality. We will now show that $\hat{\Theta}_n$ is asymptotically normal. In order to show this, let us recall Theorem 4.1.3 of [1].

[1, Theorem 4.1.3]: Let $z_1, \cdots, z_n$ be i.i.d. samples of a random variable z. Let $q(\mathsf{z}; \theta)$ be some function of z parameterized by $\theta \in \Upsilon$. Let θ^* be the true underlying parameter. Define

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} q(z_i; \theta) \qquad \text{and} \qquad \hat{\theta}_n \in \arg\min_{\theta \in \Upsilon} Q_n(\theta).$$

Let the following be true.

(a) $\hat{\theta}_n$ is consistent for θ^*,

(b) θ^* lies in the interior of the parameter space Υ,

(c) Q_n is twice continuously differentiable in an open and convex neighborhood of θ^*,

(d) $\sqrt{n} \nabla Q_n(\theta)|_{\theta=\theta^*} \xrightarrow{d} \mathcal{N}(\mathbf{0}, A(\theta^*))$, and

(e) $\nabla^2 Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{p} B(\theta^*)$ with $B(\theta)$ finite, non-singular, and continuous at θ^*,

Then, $\hat{\theta}_n$ is normal for θ^* i.e., $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, B^{-1}(\theta^*) A(\theta^*) B^{-1}(\theta^*))$.

Letting $z \coloneqq \mathbf{x}$, $\theta \coloneqq \Theta$, $\hat{\theta}_n \coloneqq \hat{\Theta}_n$, $\theta^* \coloneqq \Theta^*$, $\Upsilon = \Lambda$, $q(z; \theta) \coloneqq \exp\big(-\big\langle\big\langle \Theta, \Phi(\mathbf{x})\big\rangle\big\rangle\big)$, and $Q_n(\theta) \coloneqq \mathcal{L}_n(\Theta)$, it is sufficient to show the following:

(a) $\hat{\Theta}_n$ is consistent for Θ^*,

(b) Θ^* lies in the interior of the parameter space Λ,

(c) $\mathcal{L}_n$ is twice continuously differentiable in an open and convex neighborhood of Θ^*,

(d) $\sqrt{n} \nabla \mathcal{L}_n(\mathrm{vec}(\Theta))|_{\Theta=\Theta^*} \xrightarrow{d} \mathcal{N}(\mathbf{0}, A(\Theta^*))$, and

(e) $\nabla^2 \mathcal{L}_n(\mathrm{vec}(\Theta))|_{\Theta=\hat{\Theta}_n} \xrightarrow{p} B(\Theta^*)$ with $B(\Theta)$ finite, non-singular, and continuous at Θ^*,

Let us show these one by one.

(a) We have established that $\hat{\Theta}_n$ is consistent for Θ^* in the first half of the proof.

(b) The assumption that $\Theta^* \in \text{interior}(\Lambda)$ is equivalent to Θ^* belonging to the interior of Λ.

(c) Fix $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$, and $w_1, w_2 \in [k_3]$. We have

$$\frac{\partial^2 \mathcal{L}_n(\Theta)}{\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}} = \frac{1}{n} \sum_{t=1}^{n} \Phi_{u_1 v_1 w_1}(\mathbf{x}^{(t)}) \Phi_{u_2 v_2 w_2}(\mathbf{x}^{(t)}) \exp\big(- \big\langle\big\langle \Theta, \Phi(\mathbf{x}^{(t)}) \big\rangle\big\rangle\big).$$

Thus, $\partial^2 \mathcal{L}_n(\Theta)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}$ exists. Using the continuity of $\Phi(\cdot)$ and $\exp\big(- \big\langle\big\langle \Theta, \Phi(\cdot) \big\rangle\big\rangle\big)$, we see that $\partial^2 \mathcal{L}_n(\Theta)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}$ is continuous in an open and convex neighborhood of Θ^*.

(d) For any $u \in [k_1]$, $v \in [k_2]$ and $w \in [k_3]$, define the random variable

$$x_{uvw} = -\Phi_{uvw}(\mathbf{x}) \exp\big(- \big\langle\big\langle \Theta^*, \Phi(\mathbf{x}) \big\rangle\big\rangle\big).$$

The component of the gradient of $\mathcal{L}_n(\text{vec}(\Theta))$ corresponding to Θ_{uvw} evaluated at Θ^* is given by

$$\frac{\partial \mathcal{L}_n(\Theta^*)}{\partial \Theta_{uvw}} = -\frac{1}{n} \sum_{t=1}^{n} \Phi_{uvw}(\mathbf{x}^{(t)}) \exp\big(- \big\langle\big\langle \Theta^*, \Phi(\mathbf{x}^{(t)}) \big\rangle\big\rangle\big).$$

Each term in the above summation is distributed as the random variable x_{uvw}. The random variable x_{uvw} has zero mean (see Lemma F.1). Using this and the multivariate central limit theorem [53], we have

$$\sqrt{n} \nabla \mathcal{L}_n(\text{vec}(\Theta))|_{\Theta=\Theta^*} \xrightarrow{d} \mathcal{N}(\mathbf{0}, A(\Theta^*)),$$

where $A(\Theta^*)$ is the covariance matrix of $\text{vec}\big(\Phi(\mathbf{x}) \exp\big(- \big\langle\big\langle \Theta^*, \Phi(\mathbf{x}) \big\rangle\big\rangle\big)\big)$.

(e) We will start by showing that the following is true.

$$\nabla^2 \mathcal{L}_n(\text{vec}(\Theta))|_{\Theta=\hat{\Theta}_n} \xrightarrow{p} \nabla^2 \mathcal{L}(\text{vec}(\Theta))|_{\Theta=\Theta^*}. \tag{20}$$

To begin with, using the uniform law of large numbers [25, Theorem 2] for any $\Theta \in \Lambda$ results in

$$\nabla^2 \mathcal{L}_n(\text{vec}(\Theta)) \xrightarrow{p} \nabla^2 \mathcal{L}(\text{vec}(\Theta)). \tag{21}$$

Using the consistency of $\hat{\Theta}_n$ and the continuous mapping theorem, we have

$$\nabla^2 \mathcal{L}(\text{vec}(\Theta))|_{\Theta=\hat{\Theta}_n} \xrightarrow{p} \nabla^2 \mathcal{L}(\text{vec}(\Theta))|_{\Theta=\Theta^*}. \tag{22}$$

Let $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$, and $w_1, w_2 \in [k_3]$. From (21) and (22), for any $\epsilon > 0$, for any $\delta > 0$, there exists integers n_1, n_2 such that for $n \geq \max\{n_1, n_2\}$ we have,

$$\mathbb{P}(|\partial^2 \mathcal{L}_n(\hat{\Theta}_n)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2} - \partial^2 \mathcal{L}(\hat{\Theta}_n)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}| > \epsilon/2) \leq \delta/2$$

and

$$\mathbb{P}(|\partial^2 \mathcal{L}(\hat{\Theta}_n)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2} - \partial^2 \mathcal{L}(\Theta^*)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}| > \epsilon/2) \leq \delta/2.$$

Now for $n \geq \max\{n_1, n_2\}$, using the triangle inequality we have

$$\mathbb{P}(|\partial^2 \mathcal{L}_n(\hat{\Theta}_n)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2} - \partial^2 \mathcal{L}(\Theta^*)/\partial \Theta_{u_1 v_1 w_1} \partial \Theta_{u_2 v_2 w_2}| > \epsilon) \leq \delta/2 + \delta/2 = \delta.$$

Thus, we have (20). Using the definition of $\mathcal{L}(\Theta)$, we have

$$\begin{aligned}\partial^2\mathcal{L}(\Theta^*)/\partial\Theta_{u_1v_1w_1}\partial\Theta_{u_2v_2w_2} &= \mathbb{E}\Big[\Phi_{u_1v_1w_1}(\mathbf{x})\Phi_{u_2v_2w_2}(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\Big] \\ &\overset{(b)}{=} \mathbb{E}\Big[\Phi_{u_1v_1w_1}(\mathbf{x})\Phi_{u_2v_2w_2}(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\Big] \\ &\quad - \mathbb{E}\Big[\Phi_{u_1v_1w_1}(\mathbf{x})\Big]\mathbb{E}\Big[\Phi_{u_2v_2w_2}(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\Big] \\ &= \mathrm{cov}\Big(\Phi_{u_1v_1w_1}(\mathbf{x}),\Phi_{u_2v_2w_2}(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\Big),\end{aligned}$$

where (b) follows because $\mathbb{E}\big[\Phi_{u_2v_2w_2}(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\big] = 0$ for any $u_2 \in [k_1]$, $v_2 \in [k_2]$, and $w_2 \in [k_3]$ from Lemma F.1. Therefore, we have

$$\nabla^2\mathcal{L}_n(\mathrm{vec}(\Theta))|_{\Theta=\hat{\Theta}_n} \xrightarrow{p} B(\Theta^*),$$

where $B(\Theta^*)$ is the cross-covariance matrix of $\mathrm{vec}(\Phi(\mathbf{x}))$ and $\mathrm{vec}\big(\Phi(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\big)$. Finiteness and continuity of $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)$ implies the finiteness and continuity of $B(\Theta^*)$. By assumption, the cross-covariance matrix of $\mathrm{vec}(\Phi(\mathbf{x}))$ and $\mathrm{vec}\big(\Phi(\mathbf{x})\exp\big(-\big\langle\big\langle\Theta^*,\Phi(\mathbf{x})\big\rangle\big\rangle\big)\big)$ is invertible.

Therefore, we have the asymptotic normality of $\hat{\Theta}_n$. □

E Restricted strong convexity of the loss function

In this Section, we will show that, with enough samples, the loss function obeys the restricted strong convexity property with high probability. This result will in turn allow us to prove Theorem 4.3 in Appendix G

We will first state the main result of this Section (Proposition E.1). Next, we will introduce the notion of correlation for the centered natural statistics and provide a supporting Lemma wherein we will bound the deviation between the true correlation and the empirical correlation. Finally, we will prove Proposition E.1.

Consider any $\Theta \in \Lambda$. Let $\Delta = \Theta - \Theta^*$. Define the residual of the first-order Taylor expansion as

$$\delta\mathcal{L}_n(\Delta,\Theta^*) = \mathcal{L}_n(\Theta^*+\Delta) - \mathcal{L}_n(\Theta^*) - \langle\langle\nabla\mathcal{L}_n(\Theta^*),\Delta\rangle\rangle. \tag{23}$$

Proposition E.1. *Let Assumptions 2.1, 2.2, 2.3 and 4.1 be satisfied. For any $\delta_3 \in (0,1)$, the residual defined in* (23) *satisfies*

$$\delta\mathcal{L}_n(\Delta,\Theta^*) \geq \frac{\lambda_{\min}\exp(-\boldsymbol{r}^T\boldsymbol{d})}{4(1+\boldsymbol{r}^T\boldsymbol{d})}\|\Delta\|_{\mathrm{T}}^2,$$

with probability at least $1-\delta_3$ as long as

$$n > \frac{8\phi_{\max}^4 k_1^2k_2^2k_3^3}{\lambda_{\min}^2}\log\Big(\frac{2k_1^2k_2^2k_3^3}{\delta_3}\Big).$$

E.1 Correlation between centered natural statistics

For any $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$, and $w_1, w_2 \in [k_3]$, let $H_{u_1v_1w_1u_2v_2w_2}$ denote the correlation between $\Phi_{u_1v_1w_1}(\mathbf{x})$ and $\Phi_{u_2v_2w_2}(\mathbf{x})$ defined as

$$H_{u_1v_1w_1u_2v_2w_2} = \mathbb{E}\big[\Phi_{u_1v_1w_1}(\mathbf{x})\Phi_{u_2v_2w_2}(\mathbf{x})\big], \tag{24}$$

and let $\mathbf{H} = [H_{u_1v_1w_1u_2v_2w_2}] \in \mathbb{R}^{[k_1]\times[k_2]\times[k_3]\times[k_1]\times[k_2]\times[k_3]}$ be the corresponding correlation tensor. Similarly, we define $\hat{\mathbf{H}}$ based on the empirical estimates of the correlation

$$\hat{H}_{u_1v_1w_1u_2v_2w_2} = \frac{1}{n}\sum_{t=1}^{n} \Phi_{u_1v_1w_1}(\mathbf{x}^{(t)})\Phi_{u_2v_2w_2}(\mathbf{x}^{(t)}). \tag{25}$$

The following lemma bounds the deviation between the true correlation and the empirical correlation.

Lemma E.1. *Consider any $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$, and $w_1, w_2 \in [k_3]$. Let Assumption 2.3 be satisfied. Then, we have for any $\epsilon_2 > 0$,*

$$|\hat{H}_{u_1v_1w_1u_2v_2w_2} - H_{u_1v_1w_1u_2v_2w_2}| < \epsilon_2,$$

with probability at least $1-\delta_2$ as long as

$$n > \frac{2\phi_{\max}^4}{\epsilon_2^2}\log\Big(\frac{2k_1^2k_2^2k_3^2}{\delta_2}\Big).$$

Proof of Lemma E.1. Fix $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$, and $w_1, w_2 \in [k_3]$. The random variable defined as $Y_{u_1v_1w_1u_2v_2w_2} \coloneqq \Phi_{u_1v_1w_1}(\mathbf{x})\Phi_{u_2v_2w_2}(\mathbf{x})$ satisfies $|Y_{u_1v_1w_1u_2v_2w_2}| \le \phi_{\max}^2$ (from Assumption 2.3). Using the Hoeffding inequality we get

$$\mathbb{P}\left(|\hat{H}_{u_1v_1w_1u_2v_2w_2} - H_{u_1v_1w_1u_2v_2w_2}| > \epsilon_2\right) < 2\exp\left(-\frac{n\epsilon_2^2}{2\phi_{\max}^4}\right).$$

The proof follows by using the union bound over all $u_1, u_2 \in [k_1]$, $v_1, v_2 \in [k_2]$, and $w_1, w_2 \in [k_3]$. ∎

E.2 Proof of Proposition E.1

Proof of Proposition E.1. First, we will simplify the gradient of $\mathcal{L}_n(\Theta)$[6] evaluated at Θ^*. For any $u \in [k_1]$, $v \in [k_2]$ and $w \in [k_3]$, the component of the gradient of $\mathcal{L}_n(\Theta)$ corresponding to Θ_{uvw} evaluated at Θ^* is given by

$$\frac{\partial \mathcal{L}_n(\Theta^*)}{\partial \Theta_{uvw}} = -\frac{1}{n}\sum_{t=1}^{n} \Phi_{uvw}(\mathbf{x}^{(t)}) \exp\big(-\big\langle\big\langle \Theta^*, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle\big). \tag{26}$$

We will now provide the desired lower bound on the residual. Substituting (10) and (26) in (23), we have

$$\delta\mathcal{L}_n(\Delta, \Theta^*) = \frac{1}{n}\sum_{t=1}^{n} \exp\big(-\big\langle\big\langle \Theta^*, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle\big) \times \Big[\exp\big(-\big\langle\big\langle \Delta, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle\big) - 1 + \big\langle\big\langle \Delta, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle\Big]$$

$$\overset{(a)}{\ge} \exp(-\boldsymbol{r}^T\boldsymbol{d}) \times \frac{1}{n}\sum_{t=1}^{n} \Big[\exp\big(-\big\langle\big\langle \Delta, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle\big) - 1 + \big\langle\big\langle \Delta, \Phi(\mathbf{x}^{(t)})\big\rangle\big\rangle\Big]$$

[6] Ideally, one would consider the gradient of $\mathcal{L}_n(\mathrm{vec}(\Theta))$. However, for the ease of the exposition we abuse the terminology.

$$
\begin{aligned}
&\overset{(b)}{\geq} \exp(-\boldsymbol{r}^T\boldsymbol{d}) \times \frac{1}{n}\sum_{t=1}^{n} \frac{\left|\left\langle\left\langle \Delta, \Phi(\mathbf{x}^{(t)})\right\rangle\right\rangle\right|^2}{2+\left|\left\langle\left\langle \Delta, \Phi(\mathbf{x}^{(t)})\right\rangle\right\rangle\right|} \\
&\overset{(c)}{\geq} \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \frac{1}{n}\sum_{t=1}^{n} \left|\left\langle\left\langle \Delta, \Phi(\mathbf{x}^{(t)})\right\rangle\right\rangle\right|^2 \\
&\overset{(d)}{=} \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{w_1=1}^{k_3}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2}\sum_{w_2=1}^{k_3} \Delta_{u_1v_1w_1}\hat{H}_{u_1v_1w_1u_2v_2w_2}\Delta_{u_2v_2w_2} \\
&= \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{w_1=1}^{k_3}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2}\sum_{w_2=1}^{k_3} \Delta_{u_1v_1w_1} \times \\
&\qquad\qquad [H_{u_1v_1w_1u_2v_2w_2} + \hat{H}_{u_1v_1w_1u_2v_2w_2} - H_{u_1v_1w_1u_2v_2w_2}]\Delta_{u_2v_2w_2},
\end{aligned}
$$

where (a) follows because $-\left\langle\left\langle \Theta, \Phi(\mathbf{x})\right\rangle\right\rangle \geq -\boldsymbol{r}^T\boldsymbol{d}$ from (15), (b) follows because $e^{-z}-1+z \geq \frac{z^2}{2+|z|}$ for any $z \in \mathbb{R}$, (c) follows from (15), and (d) follows from (25).

Let the number of samples satisfy

$$
n > \frac{8\phi_{\max}^4 k_1^2k_2^2k_3^2}{\lambda_{\min}^2}\log\Big(\frac{2k_1^2k_2^2k_3^2}{\delta_3}\Big).
$$

Using Lemma E.1 with $\epsilon_2 = \frac{\lambda_{\min}}{2k_1k_2k_3}$ and $\delta_2 = \delta_3$, and the triangle inequality, we have the following with probability at least $1-\delta_3$

$$
\begin{aligned}
\delta\mathcal{L}_n(\Delta, \Theta^*) &\geq \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \Bigg[\sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{w_1=1}^{k_3}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2}\sum_{w_2=1}^{k_3} \Delta_{u_1v_1w_1}H_{u_1v_1w_1u_2v_2w_2}\Delta_{u_2v_2w_2} \\
&\qquad\qquad - \frac{\lambda_{\min}}{2k_1k_2k_3}\|\Delta\|_{1,1,1}^2\Bigg] \\
&\overset{(a)}{\geq} \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \Bigg[\sum_{u_1=1}^{k_1}\sum_{v_1=1}^{k_2}\sum_{w_1=1}^{k_3}\sum_{u_2=1}^{k_1}\sum_{v_2=1}^{k_2}\sum_{w_2=1}^{k_3} \Delta_{u_1v_1w_1}H_{u_1v_1w_1u_2v_2w_2}\Delta_{u_2v_2w_2} \\
&\qquad\qquad - \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{T}}^2\Bigg] \\
&\overset{(b)}{=} \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \Big[\mathrm{vec}(\Delta)\mathbb{E}[\mathrm{vec}(\Phi(\mathbf{x}))\mathrm{vec}(\Phi(\mathbf{x}))^T]\mathrm{vec}(\Delta)^T - \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{T}}^2\Big] \\
&\overset{(c)}{\geq} \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \Big[\lambda_{\min}\|\mathrm{vec}(\Delta)\|_2^2 - \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{T}}^2\Big] \\
&\overset{(d)}{=} \frac{\exp(-\boldsymbol{r}^T\boldsymbol{d})}{2+2\boldsymbol{r}^T\boldsymbol{d}} \times \frac{\lambda_{\min}}{2}\|\Delta\|_{\mathrm{T}}^2,
\end{aligned}
$$

where (a) follows because $\|\Delta\|_{1,1,1} \leq \sqrt{k_1k_2k_3}\|\Delta\|_{\mathrm{T}}$, (b) follows from (24), (c) follows from the Courant-Fischer theorem (because $\mathbb{E}[\mathrm{vec}(\Phi(\mathbf{x}))\mathrm{vec}(\Phi(\mathbf{x}))^T]$ is a symmetric matrix) and Assumption 4.1, and (d) follows because $\|\mathrm{vec}(\Delta)\|_2 = \|\Delta\|_{\mathrm{T}}$. □

F Bounds on the tensor maximum norm of the gradient of the loss function

In this Section, we will show that, with enough samples, the tensor maximum norm of the gradient of the loss function evaluated at the true natural parameter is bounded with high probability. This result will allow us to prove Theorem 4.3 in Appendix G.

We will first state the main result of this Section (Proposition F.1). Next, we will provide a supporting Lemma wherein we show that the expected value of a random variable of interest is zero. Finally, we will prove Proposition F.1.

Proposition F.1. *Let Assumptions 2.1, 2.2 and 2.3 be satisfied. For any $\delta_4 \in (0,1)$, any $\epsilon_4 > 0$, the components of the gradient of the loss function $\mathcal{L}_n(\Theta)$[7] evaluated at Θ^* are bounded from above as*

$$\|\nabla \mathcal{L}_n(\Theta^*)\|_{\max} \leq \epsilon_4,$$

with probability at least $1-\delta_4$ as long as

$$n > \frac{2\phi_{\max}^2 \exp(2\boldsymbol{r}^T\boldsymbol{d})}{\epsilon_4^2} \log\Big(\frac{2k_1k_2k_3}{\delta_4}\Big).$$

F.1 Supporting Lemma for Proposition F.1

Lemma F.1. *For any $u \in [k_1]$, $v \in [k_2]$ and $w \in [k_3]$, define the random variable*

$$\mathsf{x}_{uvw} = -\Phi_{uvw}(\mathbf{x}) \exp\big(-\big\langle\big\langle \Theta^*, \Phi(\mathbf{x})\big\rangle\big\rangle\big). \tag{27}$$

We have

$$\mathbb{E}[\mathsf{x}_{uvw}] = 0,$$

where the expectation is with respect to $f_{\mathbf{x}}(\mathbf{x};\Theta^*)$.

Proof of Lemma F.1. Fix any $u \in [k_1]$, $v \in [k_2]$ and $w \in [k_3]$. Using (27), we have

$$\mathbb{E}[\mathsf{x}_{uvw}] = -\int_{\mathbf{x}\in\mathcal{X}} f_{\mathbf{x}}(\mathbf{x};\Theta^*)\Phi_{uvw}(\mathbf{x}) \exp\big(-\big\langle\big\langle \Theta^*, \Phi(\mathbf{x})\big\rangle\big\rangle\big) d\mathbf{x} \overset{(a)}{=} \frac{-\int_{\mathbf{x}\in\mathcal{X}} \Phi_{uvw}(\mathbf{x}) d\mathbf{x}}{\int_{\mathbf{y}\in\mathcal{X}} \exp\big(\big\langle\big\langle \Theta^*, \Phi(\mathbf{y})\big\rangle\big\rangle\big) d\mathbf{y}}$$

$$\overset{(b)}{=} 0,$$

where (a) follows from the definition of $f_{\mathbf{x}}(\mathbf{x};\Theta^*)$, and because $\mathbb{E}_{\mathcal{U}_{\mathcal{X}}}[\Phi(\mathbf{x})]$ is a constant, and (b) follows because $\int_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) d\mathbf{x} = 0$ from Definition 2.1 □

F.2 Proof of Proposition F.1

Proof of Proposition F.1. Fix $u \in [k_1]$, $v \in [k_2]$ and $w \in [k_3]$. We will start by simplifying the gradient of the $\mathcal{L}_n(\Theta)$ evaluated at Θ^*. The component of the gradient of $\mathcal{L}_n(\Theta)$ corresponding to Θ_{uvw} evaluated at Θ^* is given by

$$\frac{\partial \mathcal{L}_n(\Theta^*)}{\partial \Theta_{uvw}} = -\frac{1}{n}\sum_{t=1}^{n} \Phi_{uvw}(\mathbf{x}^{(t)}) \exp\Big(-\Big\langle\Big\langle \Theta^*, \Phi(\mathbf{x}^{(t)})\Big\rangle\Big\rangle\Big).$$

Each term in the above summation is distributed as the random variable x_{uvw} (see (27)). The random variable x_{uvw} has zero mean (see Lemma F.1) and satisfies $|\mathsf{x}_{uvw}| \leq \phi_{\max}\exp(\boldsymbol{r}^T\boldsymbol{d})$ (from Assumption 2.3 and (15)). Using the Hoeffding's inequality, we have

$$\mathbb{P}\left(\left|\frac{\partial \mathcal{L}_n(\Theta^*)}{\partial \Theta_{uvw}}\right| > \epsilon_4\right) < 2\exp\left(-\frac{n\epsilon_4^2}{2\phi_{\max}^2\exp(2\boldsymbol{r}^T\boldsymbol{d})}\right). \tag{28}$$

The proof follows by using (28) and the union bound over all $u \in [k_1]$, $v \in [k_2]$ and $w \in [k_3]$. □

[7] Ideally, one would consider the gradient of $\mathcal{L}_n(\mathrm{vec}(\Theta))$. However, for the ease of the exposition we abuse the terminology.

G Proof of Theorem 4.3

In this Section, we will prove Theorem 4.3. We restate the Theorem below and then provide the proof.

Theorem 4.3. *Let $\hat{\Theta}_{\epsilon,n}$ be an ϵ-optimal solution of $\hat{\Theta}_n$ obtained from Algorithm 1 for ϵ of the order $O(\alpha^2\lambda_{\min})$. Let Assumptions 2.1, 2.2, 2.3, and 4.1 be satisfied. Recall Property 4.1. Then, for any $\delta \in (0,1)$, we have $\|\hat{\Theta}_{\epsilon,n} - \Theta^*\|_{\mathrm{T}} \leq \alpha$ with probability at least $1-\delta$ as long as*

$$n \geq O\Big(\frac{k_1^2 k_2^2}{\alpha^4 \lambda_{\min}^2} \log\Big(\frac{k_1 k_2}{\delta}\Big)\Big). \tag{13}$$

The computational cost scales as $O\Big(\frac{k_1 k_2}{\alpha^2} \max\big(k_1 k_2 n, c(\Lambda)\big)\Big)$ where $c(\Lambda)$ is the cost of projection onto Λ. Further, ignoring the dependence on δ, $\lambda_{\min}$, and $c(\Lambda)$, n in (13) (as well as the associated computational cost) scales as $O\big(\mathrm{poly}\big(\frac{k_1 k_2}{\alpha}\big)\big)$.

Proof of Theorem 4.3. Let the number of samples satisfy

$$\begin{aligned} n \geq \max\Bigg\{ & \frac{8\phi_{\max}^4 k_1^2 k_2^2 k_3^2}{\lambda_{\min}^2} \log\Big(\frac{4k_1^2 k_2^2 k_3^2}{\delta}\Big), \\ & \frac{2^9 \phi_{\max}^2 k_1^2 k_2^2 (\boldsymbol{r}^T \boldsymbol{g})^2 (1+\boldsymbol{r}^T \boldsymbol{d})^2 \exp(4\boldsymbol{r}^T \boldsymbol{d})}{\alpha^4 \lambda_{\min}^2} \log\Big(\frac{4k_1 k_2 k_3}{\delta}\Big) \Bigg\} \\ \overset{(a)}{\approx} & O\Big(\frac{k_1^2 k_2^2}{\alpha^4 \lambda_{\min}^2} \log\big(\frac{k_1 k_2}{\delta}\big)\Big) \approx O\Big(\mathrm{poly}\Big(\frac{k_1 k_2}{\alpha}\Big)\Big). \end{aligned}$$

where (a) follows because $k_3, \phi_{\max}, \boldsymbol{r}, \boldsymbol{g}, \boldsymbol{d} = O(1)$.

Let $\Delta = \hat{\Theta}_{\epsilon,n} - \Theta^*$. Define the residual of the first-order Taylor expansion as

$$\delta\mathcal{L}_n(\Delta, \Theta^*) = \mathcal{L}_n(\Theta^* + \Delta) - \mathcal{L}_n(\Theta^*) - \langle\langle \nabla\mathcal{L}_n(\Theta^*), \Delta\rangle\rangle. \tag{29}$$

Let $\nabla\mathcal{L}_n^{(i)}(\Theta^*)$ denote the i^{th} slice of $\nabla\mathcal{L}_n(\Theta^*)$. From the definition of an ϵ-optimal solution of $\hat{\Theta}_n$, we have

$$\begin{aligned} \epsilon &\geq \mathcal{L}_n(\hat{\Theta}_{\epsilon,n}) - \min_{\Theta\in\Lambda} \mathcal{L}_n(\Theta) \\ &\geq \mathcal{L}_n(\hat{\Theta}_{\epsilon,n}) - \mathcal{L}_n(\Theta^*) \\ &\overset{(a)}{=} \langle\langle \nabla\mathcal{L}_n(\Theta^*), \hat{\Theta}_{\epsilon,n} - \Theta^*\rangle\rangle + \delta\mathcal{L}_n(\Delta, \Theta^*) \\ &\overset{(b)}{=} \sum_{i=1}^{k_3} \langle \nabla\mathcal{L}_n^{(i)}(\Theta^*), \hat{\Theta}_{\epsilon,n}^{(i)} - \Theta^{*(i)}\rangle + \delta\mathcal{L}_n(\Delta, \Theta^*) \\ &\overset{(c)}{\geq} -\sum_{i=1}^{k_3} \mathcal{R}_i^*(\nabla\mathcal{L}_n^{(i)}(\Theta^*)) \times \mathcal{R}(\hat{\Theta}_{\epsilon,n}^{(i)} - \Theta^{*(i)}) + \delta\mathcal{L}_n(\Delta, \Theta^*) \\ &\overset{(d)}{\geq} -2\sum_{i=1}^{k_3} \mathcal{R}_i^*(\nabla\mathcal{L}_n^{(i)}(\Theta^*)) \times r_i + \delta\mathcal{L}_n(\Delta, \Theta^*) \\ &\overset{(e)}{\geq} -2k_1 k_2 \sum_{i=1}^{k_3} g_i \times \times \|\nabla\mathcal{L}_n^{(i)}(\Theta^*)\|_{\max} \times r_i + \delta\mathcal{L}_n(\Delta, \Theta^*) \\ &\overset{(f)}{\geq} -2k_1 k_2 \|\nabla\mathcal{L}_n(\Theta^*)\|_{\max} \sum_{i=1}^{k_3} g_i \times r_i + \delta\mathcal{L}_n(\Delta, \Theta^*), \end{aligned}$$

where (a) follows from (29), (b) follows from the definitions of a slice of a tensor, tensor inner product, and Frobenius inner product, (c) follows from the definition of a dual norm, (d) follows because $\mathcal{R}(\hat{\Theta}^{(i)}_{\epsilon,n} - \Theta^{*(i)}) \leq \mathcal{R}(\hat{\Theta}^{(i)}_{\epsilon,n}) + \mathcal{R}(\Theta^{*(i)}) \leq 2r_i$ from Assumption 2.1, (e) follows from Property 4.1 in Section 4, and (f) follows because $\|\nabla \mathcal{L}^{(i)}_n(\Theta^*)\|_{\max} \leq \|\nabla \mathcal{L}_n(\Theta^*)\|_{\max}\ \forall i \in [k_3]$.

Using Proposition E.1 with $\delta_3 = \frac{\delta}{2}$, and Proposition F.1 with $\delta_4 = \frac{\delta}{2}$, we have the following with probability at least $1 - \delta$.

$$\epsilon \geq -2k_1k_2\epsilon_4 \times \boldsymbol{r}^T\boldsymbol{g} + \frac{\lambda_{\min}\exp(-\boldsymbol{r}^T\boldsymbol{d})}{4(1+\boldsymbol{r}^T\boldsymbol{d})}\|\Delta\|^2_{\mathrm{T}}.$$

This can be rearranged

$$\|\Delta\|^2_{\mathrm{T}} \leq \frac{\epsilon + 2k_1k_2\epsilon_4 \times \boldsymbol{r}^T\boldsymbol{g}}{\lambda_{\min}} \times 4(1+\boldsymbol{r}^T\boldsymbol{d})\exp(\boldsymbol{r}^T\boldsymbol{d}). \tag{30}$$

Now, let

$$\epsilon = \frac{\alpha^2\lambda_{\min}}{8(1+\boldsymbol{r}^T\boldsymbol{d})\exp(\boldsymbol{r}^T\boldsymbol{d})} \qquad \text{and} \qquad \epsilon_4 = \frac{\alpha^2\lambda_{\min}}{16k_1k_2 \times \boldsymbol{r}^T\boldsymbol{g} \times (1+\boldsymbol{r}^T\boldsymbol{d}) \times \exp(\boldsymbol{r}^T\boldsymbol{d})} \tag{31}$$

Plugging in ϵ and ϵ_4 from (31) in (30), we obtain that

$$\|\Delta\|_{\mathrm{T}} \leq \alpha.$$

The computational cost of the operation $\Theta_{(t)} - \eta\nabla\mathcal{L}_n(\Theta_{(t)}) - \Theta$ in Algorithm 1 is of the order k_1k_2n (because $k_3 = O(1)$). Therefore, the computational cost of the step $\Theta_{(t+1)} \leftarrow \arg\min_{\Theta\in\Lambda}\|\Theta_{(t)} - \eta\nabla\mathcal{L}_n(\Theta_{(t)}) - \Theta\|$ of Algorithm 1 is of the order $\max\{k_1k_2n, c(\Lambda)\}$. From Lemma 3.1, with $\epsilon = O(\alpha^2\lambda_{\min})$, Algorithm 1 returns an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ as long as $\tau = O\Big(\mathrm{poly}\Big(\frac{k_1k_2}{\alpha^2\lambda_{\min}}\Big)\Big)$. Therefore, the total computational cost scales as $O\Big(\frac{k_1k_2}{\alpha^2\lambda_{\min}}\max\big(k_1k_2n, c(\Lambda)\big)\Big)$. Whenever the cost of projection onto Λ is $O\big(\mathrm{poly}(k_1k_2)\big)$, we have the total computational cost scaling as $O\Big(\mathrm{poly}\Big(\frac{k_1k_2}{\alpha}\Big)\Big)$. □

H Computational cost for the example constraints on the natural parameters

In this Section, we provide Corollary H.1, Corollary H.2, and Corollary H.3. These Corollaries provide the computational cost to produce an ϵ-optimal solution of $\hat{\Theta}_n$ for sparse decomposition of Θ, low-rank decomposition of Θ, and sparse-plus-low-rank decomposition of Θ. respectively. Recall the convex relaxations of these constraints from Section 2.1.

H.1 Sparse Decomposition

Corollary H.1. *(Sparse decomposition) Suppose Θ^* has a sparse decomposition i.e., $\Theta^* = (\Theta^{*(1)})$ and $\|\Theta^{*(1)}\|_{1,1} \leq r_1$. Let Assumptions 2.1, 2.2, 2.3, and 4.1 be satisfied. Let*

$$n \geq O\bigg(\frac{k_1^2k_2^2}{\alpha^4\lambda^2_{\min}}\log\Big(\frac{k_1k_2}{\delta}\Big)\bigg).$$

Let $\eta = 1/k_1k_2k_3\phi^2_{\max}\exp(r_1d_1)$ and $\Theta^{(0)} = \mathbf{0}$. Then, Algorithm 1 is guaranteed to produce an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ such that $\|\hat{\Theta}_{\epsilon,n} - \Theta^*\|_{\mathrm{T}} \leq \alpha$, with probability at least $1-\delta$ and with number of computations of the order

$$O\bigg(\frac{k_1^4k_2^4}{\alpha^6\lambda^3_{\min}}\log\Big(\frac{k_1k_2}{\delta}\Big)\bigg).$$

Proof of Corollary H.1 . The computational cost of projecting on the $L_{1,1}$ ball is $O(k_1k_2)$ (see [15] and note $k_3 = O(1)$). The computational cost of the operation $\Theta_{(t)} - \eta\nabla\mathcal{L}_n(\Theta_{(t)}) - \Theta$ is $O(k_1k_2n)$ (because $k_3 = O(1)$). Therefore, the computational cost of the step $\Theta_{(t+1)} \leftarrow \arg\min_{\Theta\in\Lambda} \|\Theta_{(t)} - \eta\nabla\mathcal{L}_n(\Theta_{(t)}) - \Theta\|$ of Algorithm 1 is $O(k_1k_2n)$.

From Lemma 3.1, Algorithm 1 returns an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ as long as

$$\tau \geq \frac{2k_1k_2\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})}{\epsilon}\|\hat{\Theta}_n\|_{\mathrm{T}}^2.$$

Also, $\|\hat{\Theta}_n\|_{\mathrm{T}}^2 = \|\hat{\Theta}_n^{(1)}\|_{\mathrm{F}}^2 \leq \|\hat{\Theta}_n^{(1)}\|_{1,1}^2 \leq r_1^2$. Combining everything, the computational cost scales as $O\Big(\frac{k_1^2k_2^2n}{\epsilon}\Big)$. Using Theorem 4.3, and plugging in $n = O\Big(\frac{k_1^2k_2^2}{\alpha^4\lambda_{\min}^2}\log\Big(\frac{k_1k_2}{\delta}\Big)\Big)$ and $\epsilon = O(\alpha^2\lambda_{\min})$ completes the proof. □

H.2 Low-rank decomposition

Corollary H.2. *(Low-rank decomposition) Suppose Θ^* has a low-rank decomposition i.e., $\Theta^* = (\Theta^{*(1)})$ and $\|\Theta^*\|_\star \leq r_1$. Let Assumptions 2.1, 2.2, 2.3, and 4.1 be satisfied. Let*

$$n \geq O\Big(\frac{k_1^2k_2^2}{\alpha^4\lambda_{\min}^2}\log\Big(\frac{k_1k_2}{\delta}\Big)\Big).$$

Let $\eta = 1/k_1k_2k_3\phi_{\max}^2\exp(r_1d_1)$ and $\Theta^{(0)} = \mathbf{0}$. Then, Algorithm 1 is guaranteed to produce an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ such that $\|\hat{\Theta}_{\epsilon,n} - \Theta^*\|_{\mathrm{T}} \leq \alpha$, with probability at least $1-\delta$ and with number of computations of the order

$$O\Big(\frac{k_1^4k_2^4}{\alpha^6\lambda_{\min}^3}\log\Big(\frac{k_1k_2}{\delta}\Big)\Big).$$

Proof of Corollary H.2 . The computational cost of projecting on the nuclear ball is $O(k_1k_2\min\{k_1,k_2\})$ (see [23] and note $k_3 = O(1)$). The computational cost of the operation $\Theta_{(t)} - \eta\nabla\mathcal{L}_n(\Theta_{(t)}) - \Theta$ is $O(k_1k_2n)$ because ($k_3 = O(1)$). Therefore, the computational cost of the step $\Theta_{(t+1)} \leftarrow \arg\min_{\Theta\in\Lambda} \|\Theta_{(t)} - \eta\nabla\mathcal{L}_n(\Theta_{(t)}) - \Theta\|$ of Algorithm 1 is $O(k_1k_2\max\{\min\{k_1,k_2\},n\})$.

From Lemma 3.1, Algorithm 1 returns an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ scales as

$$\tau \geq \frac{2k_1k_2\phi_{\max}^2 \exp(\boldsymbol{r}^T\boldsymbol{d})}{\epsilon}\|\hat{\Theta}_n\|_{\mathrm{F}}^2.$$

Also, $\|\hat{\Theta}_n\|_{\mathrm{F}}^2 \leq \|\hat{\Theta}_n\|_\star^2 \leq r_1^2$. Combining everything, the computational cost is of the order $O\Big(\frac{k_1^2k_2^2\max\{\min\{k_1,k_2\},n\}}{\epsilon}\Big)$. Using Theorem 4.3, and plugging in $n = O\Big(\frac{k_1^2k_2^2}{\alpha^4\lambda_{\min}^2}\log\Big(\frac{k_1k_2}{\delta}\Big)\Big)$ and $\epsilon = O(\alpha^2\lambda_{\min})$ completes the proof. □

H.3 Sparse-plus-low-rank decomposition

Corollary H.3. *(Sparse-plus-low-rank decomposition) Suppose Θ^* has a sparse-plus-low-rank decomposition i.e., $\Theta^* = (\Theta^{*(1)}, \Theta^{*(2)})$ such that $\|\Theta^{*(1)}\|_{1,1} \leq r_1$ and $\|\Theta^{*(2)}\|_\star \leq r_2$. Let Assumptions 2.1, 2.2, 2.3, and 4.1 be satisfied. Let*

$$n \geq O\Big(\frac{k_1^2k_2^2}{\alpha^4\lambda_{\min}^2}\log\Big(\frac{k_1k_2}{\delta}\Big)\Big).$$

Let $\eta = 1/k_1k_2k_3\phi_{\max}^2\exp(r_1d_1 + r_2d_2)$ and $\Theta^{(0)} = \mathbf{0}$. Then, Algorithm 1 is guaranteed to produce an ϵ-optimal solution $\hat{\Theta}_{\epsilon,n}$ such that $\|\hat{\Theta}_{\epsilon,n} - \Theta^*\|_{\mathrm{T}} \leq \alpha$, with probability at least $1-\delta$ and with number of computations of the order

$$O\Big(\frac{k_1^4k_2^4}{\alpha^6\lambda_{\min}^3}\log\Big(\frac{k_1k_2}{\delta}\Big)\Big).$$

Proof of Corollary H.3 . The proof follows directly from the proofs of Corollary H.1 and Corollary H.2. □

I Examples

In this Section, we provide a more elaborate discussion on the examples of natural parameters and statistics from Section 2.1.

I.1 Sparse-plus-low-rank decomposition

The natural statistic Φ of an exponential family is such that for any $i_1 \neq i_2 \in [k_1], j_1 \neq j_2 \in [k_2], l_1 \neq l_2 \in [k_3]$, $\Phi_{i_1 j_1 l_1} \neq \Phi_{i_2 j_2 l_2}$. Further, an exponential family is minimal if there does not exist a non-zero tensor $\mathbf{U} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ such that $\sum_{i\in[k_1], j\in[k_2], l\in[k_3]} \mathbf{U}_{ijl}\Phi_{ijl}(\mathbf{x})$ is equal to a constant for all $\mathbf{x} \in \mathcal{X}$. However, for the sparse-plus-low-rank decomposition, it is desirable to let $\Phi^{(1)} = \Phi^{(2)}$ (see [8, 37]). In this scenario, there exists a non-zero tensor $\mathbf{U} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ such that $\sum_{i\in[k_1], j\in[k_2], l\in[k_3]} \mathbf{U}_{ijl}\Phi_{ijl}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$ for e.g., this is true if $\mathbf{U}^{(1)} = -\mathbf{U}^{(2)}$. In this situation, we say an exponential family is minimal if there does not exist a non-zero tensor $\mathbf{U} \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ such that $\sum_{l\in[k_3]} \mathbf{U}^{(l)} \neq 0$ as well as $\sum_{i\in[k_1], j\in[k_2], l\in[k_3]} \mathbf{U}_{ijl}\Phi_{ijl}(\mathbf{x})$ is equal to a constant for all $\mathbf{x} \in \mathcal{X}$. Therefore, it is often convenient to represent the tensor $\mathbf{U}$ in terms of a matrix and define minimality of an exponential family in terms of this new matrix.

I.2 Assumptions 2.1 and 2.2

While we expect the constants $\boldsymbol{r}$ in Assumption 2.1 and $\boldsymbol{d}$ in Assumption 2.2 to be $O(1)$ for most applications, the sample complexity and the computational complexity in Theorem 4.3 would still be $O\Big(\text{poly}\Big(\frac{k_1 k_2}{\alpha}\Big)\Big)$ as long as $\boldsymbol{r}$ and $\boldsymbol{d}$ are $O\Big(\log(k_1 k_2)\Big)$.

I.3 Polynomial natural statistic

Suppose the natural statistics are polynomials of $\mathbf{x}$ with maximum degree l, i.e., $\prod_{i\in[p]} x_i^{l_i}$ such that $l_i \geq 0\ \forall i \in [p]$ and $\sum_{i\in[p]} l_i \leq l$.

- Let $\mathcal{X} = [0, b]$ for $b \in \mathbb{R}$. We will first show that $\phi_{\max} = 2b^l$. We have

$$\begin{aligned}
\|\Phi(\mathbf{x})\|_{\max} &= \max_{u\in[k_1], v\in[k_2], w\in[k_3]} |\Phi_{uvw}(\mathbf{x})| \\
&\overset{(a)}{=} \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\Phi_{uvw}(\mathbf{x}) - \mathbb{E}_{\mathcal{U}_\mathcal{X}}[\Phi_{uvw}(\mathbf{x})]\Big| \\
&\overset{(b)}{\leq} \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\Phi_{uvw}(\mathbf{x})\Big| + \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\mathbb{E}_{\mathcal{U}_\mathcal{X}}[\Phi_{uvw}(\mathbf{x})]\Big| \\
&\leq 2 \max_{\mathbf{x}\in\mathcal{X}} \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\Phi_{uvw}(\mathbf{x})\Big| \leq 2b^l.
\end{aligned}$$

  where (a) follows from Definition 2.1 and (b) follows from the triangle inequality.

- Suppose Θ^* has a sparse decomposition i.e., $\Theta^* = (\Theta^{*(1)})$ and $\|\Theta^{*(1)}\|_{1,1} \leq r_1$. The dual norm of the matrix $L_{1,1}$ norm is the matrix maximum norm. Then, if $\mathcal{X} = [0, b]$ for $b \in \mathbb{R}$,

$$\mathcal{R}_1^*(\Phi^{(1)}(\mathbf{x})) = \|\Phi^{(1)}(\mathbf{x})\|_{\max} = \|\Phi(\mathbf{x})\|_{\max} \leq \phi_{\max} = 2b^l.$$

- Suppose Θ^* has a low-rank decomposition i.e., $\Theta^* = (\Theta^{*(1)})$ and $\|\Theta^*\|_\star \leq r_1$. The dual norm of the matrix nuclear norm is the matrix spectral norm. Then,

$$\mathcal{R}_1^*(\Phi^{(1)}(\mathbf{x})) = \|\Phi^{(1)}(\mathbf{x})\|.$$

  Let $l = 2$, and $\mathcal{X} = \mathcal{B}(0, b)$. Observe that by writing $\Phi^{(1)}(\mathbf{x}) = \tilde{x}\tilde{x}^T$ where $\tilde{x} = (1, x_1, \cdots, x_p)$, we have

$$\|\Phi^{(1)}(\mathbf{x})\| \leq 2\Big(1 + \sum_{i\in[p]} \mathbf{x}_i^2\Big) \leq 2(1 + b^2).$$

- Suppose Θ^* has a sparse-plus-low-rank decomposition i.e., $\Theta^* = (\Theta^{*(1)}, \Theta^{*(2)})$ such that $\|\Theta^{*(1)}\|_{1,1} \leq r_1$ and $\|\Theta^{*(2)}\|_\star \leq r_2$. The dual norm of the matrix $L_{1,1}$ norm is the matrix maximum norm and the dual norm of the matrix nuclear norm is the matrix spectral norm. Let $l = 2$, and $\mathcal{X} = \mathcal{B}(0, b)$. Then,

$$\boldsymbol{\mathcal{R}}^*(\Phi(\mathbf{x})) \leq (\|\Phi^{(1)}(\mathbf{x})\|_{\max}, \|\Phi^{(2)}(\mathbf{x})\|) \leq (2b^2, 2 + 2b^2).$$

I.4 Trigonometric natural statistic

Suppose the natural statistics are sines and cosines of $\mathbf{x}$ with l different frequencies, i.e., $\sin(\sum_{i\in[p]} l_i x_i) \cup \cos(\sum_{i\in[p]} l_i x_i)$ such that $l_i \in [l] \cup \{0\}$.

- Let $\mathcal{X} \subset \mathbb{R}^p$. We will first show that $\phi_{\max} = 2$. We have

$$\begin{aligned}
\|\Phi(\mathbf{x})\|_{\max} &= \max_{u\in[k_1], v\in[k_2], w\in[k_3]} |\Phi_{uvw}(\mathbf{x})| \\
&\overset{(a)}{=} \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\Phi_{uvw}(\mathbf{x}) - \mathbb{E}_{\mathcal{U}_\mathcal{X}}[\Phi_{uvw}(\mathbf{x})]\Big| \\
&\overset{(b)}{\leq} \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\Phi_{uvw}(\mathbf{x})\Big| + \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\mathbb{E}_{\mathcal{U}_\mathcal{X}}[\Phi_{uvw}(\mathbf{x})]\Big| \\
&\leq 2 \max_{\mathbf{x}\in\mathcal{X}} \max_{u\in[k_1], v\in[k_2], w\in[k_3]} \Big|\Phi_{uvw}(\mathbf{x})\Big| \leq 2.
\end{aligned}$$

  where (a) follows from Definition 2.1 and (b) follows from the triangle inequality.
- Suppose Θ^* has a sparse decomposition i.e., $\Theta^* = (\Theta^{*(1)})$ and $\|\Theta^{*(1)}\|_{1,1} \leq r_1$. The dual norm of the matrix $L_{1,1}$ norm is the matrix maximum norm. Then, for any $\mathcal{X} \subset \mathbb{R}^p$,

$$\mathcal{R}_1^*(\Phi^{(1)}(\mathbf{x})) = \|\Phi^{(1)}(\mathbf{x})\|_{\max} = \|\Phi(\mathbf{x})\|_{\max} \leq \phi_{\max} = 2.$$

I.5 Combinations of polynomial and trigonometric statistics

Suppose the natural statistics are combinations of polynomials of $\mathbf{x}$ with maximum degree l, i.e., $\prod_{i\in[p]} x_i^{l_i}$ such that $l_i \geq 0\ \forall i \in [p]$ and $\sum_{i\in[p]} l_i \leq l$ as well as sines and cosines of $\mathbf{x}$ with $\tilde{l}$ different frequencies, i.e., $\sin(\sum_{i\in[p]} l_i x_i) \cup \cos(\sum_{i\in[p]} l_i x_i)$ such that $l_i \in [\tilde{l}] \cup \{0\}$.

- Let $\mathcal{X} = [0, b]$ for $b \in \mathbb{R}$. From Appendix I.3 and Appendix I.4, it is easy to verify that $\phi_{\max} = \max\{2, 2b^l\}$.
- Suppose Θ^* has a sparse decomposition i.e., $\Theta^* = (\Theta^{*(1)})$ and $\|\Theta^{*(1)}\|_{1,1} \leq r_1$. The dual norm of the matrix $L_{1,1}$ norm is the matrix maximum norm. Then, if $\mathcal{X} = [0, b]$ for $b \in \mathbb{R}$, it is easy to verify that

$$\mathcal{R}_1^*(\Phi^{(1)}(\mathbf{x})) = \|\Phi^{(1)}(\mathbf{x})\|_{\max} = \|\Phi(\mathbf{x})\|_{\max} \leq \phi_{\max} = \max\{2, 2b^l\}.$$

J Property 4.1 for norms of interest

In this Section, we show that the g defined in Property 4.1 in Section 4 is 1 for the entry-wise $L_{p,q}$ norm ($p, q \geq 1$), the Schatten p-norm ($p \geq 1$), and the operator p-norm ($p \geq 1$).

J.1 The entry-wise $L_{p,q}$ norm

Let $\tilde{\mathcal{R}}(\cdot)$ denote the entry-wise $L_{p,q}$ norm for some $p, q \geq 1$. We will show that for any matrix $\mathbf{M} \in \mathbb{R}^{k_1 \times k_2}$

$$\tilde{\mathcal{R}}(\mathbf{M}) \leq \|\mathbf{M}\|_{\max} \times k_1^{\frac{1}{p}} k_2^{\frac{1}{q}}.$$

By the definition of the entry-wise $L_{p,q}$ norm, we have

$$\begin{aligned}\tilde{\mathcal{R}}(\mathbf{M}) = \Bigg(\sum_{j \in [k_2]} \Bigg(\sum_{i \in [k_1]} |M_{ij}|^p \Bigg)^{\frac{q}{p}} \Bigg)^{\frac{1}{q}} &\leq \Bigg(\sum_{j \in [k_2]} \Bigg(\sum_{i \in [k_1]} \|\mathbf{M}\|_{\max}^p \Bigg)^{\frac{q}{p}} \Bigg)^{\frac{1}{q}} \\ &= k_1^{\frac{1}{p}} k_2^{\frac{1}{q}} \|\mathbf{M}\|_{\max} \leq k_1 k_2 \|\mathbf{M}\|_{\max}.\end{aligned}$$

J.2 The Schatten p-norm

Let $\tilde{\mathcal{R}}(\cdot)$ denote the Schatten p-norm for some $p \geq 1$. We will show that for any matrix $\mathbf{M} \in \mathbb{R}^{k_1 \times k_2}$

$$\tilde{\mathcal{R}}(\mathbf{M}) \leq \|\mathbf{M}\|_{\max} \times \sqrt{\min\{k_1, k_2\} k_1 k_2}.$$

Let the rank of $\mathbf{M}$ be denoted by r and the singular values of $\mathbf{M}$ be denoted by $\sigma_i(\mathbf{M})$ for $i \in [r]$. By the definition of the Schatten p-norm, we have

$$\begin{aligned}\tilde{\mathcal{R}}(\mathbf{M}) = \Bigg(\sum_{i \in [r]} \sigma_i^p(\mathbf{M}) \Bigg)^{\frac{1}{p}} \overset{(a)}{\leq} \sum_{i \in [r]} \sigma_i(\mathbf{M}) &\overset{(b)}{\leq} \sqrt{r k_1 k_2} \|\mathbf{M}\|_{\max} \\ &\overset{(c)}{\leq} \sqrt{\min\{k_1, k_2\} k_1 k_2} \|\mathbf{M}\|_{\max} \leq k_1 k_2 \|\mathbf{M}\|_{\max}\end{aligned}$$

where (a) follows because of the monotonicity of the Schatten p-norms, (b) follows because $\|\mathbf{M}\|_\star \leq \sqrt{r k_1 k_2} \|\mathbf{M}\|_{\max}$, and (c) follows because $r \leq \min\{k_1, k_2\}$.

J.3 The operator p-norm

Let $\tilde{\mathcal{R}}(\cdot)$ denote the operator p-norm for some $p \geq 1$. We will show that for any matrix $\mathbf{M} \in \mathbb{R}^{k_1 \times k_2}$

$$\tilde{\mathcal{R}}(\mathbf{M}) \leq \|\mathbf{M}\|_{\max} \times k_1^{\frac{1}{p}} k_2^{1 - \frac{1}{p}}.$$

Let $q = \frac{p}{p-1}$. For $i \in k_1$, let $[\mathbf{M}]_i$ denote the i^{th} row of $\mathbf{M}$. By the definition of the operator p-norm, we have

$$\begin{aligned}\tilde{\mathcal{R}}(\mathbf{M}) = \max_{\mathbf{y}: \|\mathbf{y}\|_p = 1} \|\mathbf{M}\mathbf{y}\|_p &\overset{(a)}{\leq} k_1^{\frac{1}{p}} \max_{\mathbf{y}: \|\mathbf{y}\|_p = 1} \|\mathbf{M}\mathbf{y}\|_\infty \\ &\overset{(b)}{\leq} k_1^{\frac{1}{p}} \max_{\mathbf{y}: \|\mathbf{y}\|_p = 1} \max_{i \in [k_1]} \|[\mathbf{M}]_i\|_q \|\mathbf{y}\|_p \\ &\leq k_1^{\frac{1}{p}} \max_{i \in [k_1]} \|[\mathbf{M}]_i\|_q\end{aligned}$$

$$\overset{(c)}{\leq} k_1^{\frac{1}{p}} k_2^{\frac{1}{q}} \max_{i\in[k_1]} \|[\mathbf{M}]_i\|_\infty$$
$$= k_1^{\frac{1}{p}} k_2^{1-\frac{1}{p}} \|\mathbf{M}\|_{\max} \leq k_1 k_2 \|\mathbf{M}\|_{\max}$$

where (a) follows because $\|\mathbf{v}\|_p \leq m^{\frac{1}{p}} \|\mathbf{v}\|_\infty$ for any vector $\mathbf{v} \in \mathbb{R}^m$ and $p \geq 1$, (b) follows from the definition of the infinity norm of a vector and using the Hölder's inequality, and (c) follows because $\|\mathbf{v}\|_q \leq m^{\frac{1}{q}} \|\mathbf{v}\|_\infty$ for any vector $\mathbf{v} \in \mathbb{R}^m$ and $q \geq 1$.