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 $\ell_2$ error of $\alpha$ in the parameter estimation with sample complexity $O(poly(k/\alpha))$ and computational complexity $O(poly(k/\alpha))$. 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.

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 $x = (x_1, \ldots, x_p)$ 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

$$f_x(x; \theta) \propto \exp(\theta^T \phi(x) + \beta(x)),$$

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

$$f_x(x; \Theta) \propto \exp(\langle\langle \Theta, \Phi(x) \rangle\rangle) = \exp\left(\sum_{i \in [k_1], j \in [k_2], l \in [k_3]} \Theta_{ijl} \times \Phi_{ijl}(x)\right),$$

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 $\langle\langle \Theta, \Phi(x) \rangle\rangle$ denotes the tensor inner product, i.e., the sum of product of entries of $\Theta$ and $\Phi(x)$. An exponential family is minimal if there does not exist a nonzero tensor $U \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ such that $\langle\langle U, \Phi(x) \rangle\rangle$ is equal to a constant for all $x \in \mathcal{X}$.

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 families, first introduced by Hogg and Craig [20], are 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 $\Theta$ specifies a particular distribution in the exponential family. If the natural statistic $\Phi$ and the support of $x$ (i.e., $X$) are known, then learning a distribution in the exponential family is equivalent to learning the corresponding natural parameter $\Theta$. Despite having a long history, there has been limited progress on learning natural parameter $\Theta$ 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 $\Theta^*$ from i.i.d. samples of $x$ obtained from $f_x(\cdot; \Theta^*)$. We focus on the setting where $\Theta^*$ and $\Phi$ 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 $x^{(1)}, \ldots, x^{(n)}$ of $x$, we propose the following novel loss function to learn a distribution belonging to the exponential family in $f_x(\cdot; \Theta)$:

$$
\mathcal{L}_n(\Theta) = \frac{1}{n} \sum_{t=1}^{n} \exp \left( - \langle \Theta, \Phi(x^{(t)}) \rangle \right),
$$

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

$$
\hat{\Theta}_n \in \arg \min_{\Theta \in \Lambda} \mathcal{L}_n(\Theta),
$$

is consistent and (under mild further restrictions) asymptotically normal (see Theorem 4.2). We obtain an $\epsilon$-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 $O(\text{poly}(k_1 k_2/\epsilon))$ 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 $\alpha$ in the tensor $\ell_2$ norm with respect to the true natural parameter $\Theta^*$ with $O(\text{poly}(k_1 k_2/\alpha))$ samples 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 $f_x(\cdot; \Theta^* - \Theta)$. We show that the estimator that minimizes the population version of the loss function in (3).

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

is equivalent to the estimator that minimizes the Kullback-Leibler (KL) divergence between $U_X$ (the uniform distribution on $X$) and $f_x(\cdot; \Theta^* - \Theta)$ (see Theorem 4.1). Therefore, at the population level, our method can be viewed as the MLE of the parametric family $f_x(\cdot; \Theta^* - \Theta)$. 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 We let $k_3 = O(1)$. See Section 2.
1.2 Related Works

In this section, we look at the related works on learning exponential family. Broadly speaking, there are two main approaches to overcome the computational hardness of the MLE: (a) approximating the MLE and (b) selecting a surrogate objective. Given the richness of both 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 [37]; (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 [3] 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 [56], 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 [53]). 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:

\[
f_{x_i|x_{-i}}(x_i|x_{-i} = x_{-i}) \propto \exp \left( \sum_{j \in [p], j \neq i} \theta_{ij} \phi(x_j) \right) \phi(x_i),
\]

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. Yang et al. [62] showed that only the following joint distribution is consistent with the node-conditional distributions in (5):

\[
f_x(x) \propto \exp \left( \sum_{i \in [p]} \theta_i \phi(x_i) + \sum_{j \neq i} \theta_{ij} \phi(x_i) \phi(x_j) \right).
\]

To learn the node-conditional distribution in (5) for linear \( \phi(\cdot) \) (i.e., \( \phi(x) = x \)), Yang et al. [62] proposed an \( \ell_1 \) regularized node-conditional log-likelihood. However, their finite sample analysis

\[\text{Under node-wise-sparsity, } \sum_{j \in [p], j \neq i} \theta_{ij} \text{ is bounded by a constant for every } i \in [p].\]
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 on 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 [55]) for vector-space MRFs and non-linear $\phi(\cdot)$: with $\mathcal{U}_X$, being the uniform distribution on $X$, and $\tilde{\phi}(x_i) = \phi(x_i) - \int x_i \phi(x_i')d\mathcal{U}_X(x_i')dx_i'$

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

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 $\psi(\cdot)$

$$f_x(x) \propto \exp \left( \sum_{i \in [p]} \theta_i \phi(x_i) + \sum_{j \neq i} \theta_{ij} \psi(x_i, x_j) \right).$$

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 [38]. 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 $S(x, Q)$ is a numerical score assigned to a realization $x$ of a random variable $x$ and it measures the quality of a predictive distribution $Q$ (with probability density $q(\cdot)$). If $P$ is the true distribution of $x$, the divergence $D(P, Q)$ associated with a scoring rule is defined as $\mathbb{E}_P[S(x, Q) - S(x, P)]$. 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 $\Delta$ is the Laplacian operator, $\nabla$ is the gradient and $\| \cdot \|_2$ is the $\ell_2$ 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 $[t] := \{1, \ldots, t\}$. For a deterministic sequence $v_1, \ldots, v_t$, we let $v := (v_1, \ldots, v_t)$. For a random sequence $v_1, \ldots, v_t$, we let $v := (v_1, \ldots, v_t)$. For a matrix $M \in \mathbb{R}^{u \times v}$, we denote the element in $i^{th}$ row and $j^{th}$ column by $M_{ij}$, the singular values of the matrix by $\sigma_i(M)$ for $i \in [\min\{u, v\}]$, the matrix maximum norm by $\|M\|_{\max} := \max_{i \in [u], j \in [v]} |M_{ij}|$, the entry-wise $L_{1,1}$ norm by $\|M\|_{1,1} := \sum_{i \in [u], j \in [v]} |M_{ij}|$, the nuclear norm by $\|M\|_* := \sum_{i \in [\min\{u, v\}]} \sigma_i(M)$. We denote the Frobenius or Trace inner product of matrices $M, N \in \mathbb{R}^{u \times v}$ by $\langle M, N \rangle := \sum_{i \in [u], j \in [v]} M_{ij}N_{ij}$. For a matrix $M \in \mathbb{R}^{u \times v}$, we denote a generic norm on $\mathbb{R}^{u \times v}$ by $\|M\|_p$ and denote the associated dual norm by $\|M\|_p^*$. For a tensor $U \in \mathbb{R}^{u \times v \times w}$, we denote its $(i, j, t)$ entry by $U_{ijt}$, its $l^{th}$ slice (obtained by fixing the last index) by $U_{i:j}$ or $U^{(l)}$, the tensor maximum norm (with a slight abuse of notation) by $\|U\|_{\max} := \max_{i \in [u], j \in [v], t \in [w]} |U_{ijt}|$, and the tensor inner product of tensors $U, V \in \mathbb{R}^{u \times v \times w}$ by $\langle U, V \rangle := \sum_{i \in [u], j \in [v], t \in [w]} U_{ijt}V_{ijt}$. We denote the vectorization of the tensor $U \in \mathbb{R}^{u \times v \times w}$ by $\text{vec}(U) \in \mathbb{R}^{uwv}$ (the ordering of the elements is not important as long as it is consistent). Let $0 \in \mathbb{R}^{k_1 \times k_2 \times k_3}$ denote the tensor with every entry zero. We denote a $p$-dimensional ball of radius $b$ centered at 0 by $B(0, b)$.

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 $f_{\theta}(\cdot; \Theta = \Theta)$, 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 $x = (x_1, \ldots, x_p)$ be a $p-$dimensional vector of continuous random variables. For any $i \in [p]$, let the support of $x_i$ be $X_i \subset \mathbb{R}$. Define $X := \prod_{i=1}^p X_i$. Let $x = (x_1, \ldots, x_p) \in X$ be a realization of $x$. In this work, we assume that the random vector $x$ belongs to an exponential family with bounded support (i.e., length of $X_i$ 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 $\Phi(x) : X \to \mathbb{R}^{k_1 \times k_2 \times k_3}$ as follows.

Natural parameter $\Theta$. We focus on natural parameters with bounded norms. However, instead of having such constraints on the natural parameter $\Theta$ as it is, we decompose $\Theta$ 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 efficiently capture the sparse-plus-low-rank decomposition of $\Theta$ in addition to only sparse or only low-rank decompositions of $\Theta$ (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 $\Theta$.) For every $i \in [k_3]$, we let $R_i(\Theta^{(i)}) \leq r_i$ where $\Theta^{(i)} \in \mathbb{R}^{k_1 \times k_2}$ is the $i^{th}$ slice of $\Theta$, $R_i : \mathbb{R}^{k_1 \times k_2} \to \mathbb{R}_+$ is a norm and $r_i$ is a known constant. This decomposition is represented compactly by $R(\Theta) \leq r$ where $R(\Theta) = (R_1(\Theta^{(1)}), \ldots, R_{k_3}(\Theta^{(k_3)}))$ and $r = (r_1, \ldots, r_{k_3})$.

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

---

1 Even though we focus on continuous variables, our framework applies equally to discrete variables.
Natural Statistic $\Phi$. 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., $\mathcal{U}_X$) is zero. $\mathcal{U}_X$ 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:

$$\Phi(\cdot) := \Phi(\cdot) - \mathbb{E}_{d \in \mathcal{X}}[\Phi(x)].$$

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 $\Theta$ and the centered natural statistic $\Phi(\cdot)$ (see Appendix [B.1]).

**Assumption 2.2.** (Bounded dual norms of $\Phi$). 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 $x \in \mathcal{X}$, $\mathcal{R}_i^*(\Phi(x)) \leq d_i$. This is represented compactly by $\mathcal{R}^*(\Phi(x)) \leq d$ where $\mathcal{R}^*(\Phi(x)) = (\mathcal{R}_1^*(\Phi^{(1)}(x)), \ldots, \mathcal{R}_{k_3}^*(\Phi^{(k_3)}(x)))$ and $d = (d_1, \ldots, d_{k_3})$.

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 $\Phi$). For any $x \in \mathcal{X}$, $\|\Phi(x)\|_{\max} \leq \phi_{\max}$.

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

$$f_x(x; \Theta) \propto \exp \left( \left\langle \left\langle \Theta, \Phi(x) \right\rangle \right\rangle \right). \quad (9)$$

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

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

**Goal.** (Natural Parameter Recovery). Given $n$ independent samples of $x$ i.e., $x^{(1)}, \ldots, x^{(n)}$ obtained from $f_x(x; \Theta^*)$, compute an estimate $\hat{\Theta}$ of $\Theta^*$ in polynomial time such that $\|\Theta^* - \hat{\Theta}\|_T$ 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 [I] and [J] for more discussion on these examples.

**Examples of natural parameter.** We provide examples in Table 1 to illustrate the decomposability of $\Theta$ as in Assumption 2.1. We will revisit these examples briefly in Section 4 and in-depth in Appendix [I]. 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>
<thead>
<tr>
<th>Decomposition</th>
<th>$k_3$</th>
<th>Convex Relaxation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse decomposition ($\Theta^* = (\Theta^{(1)})$)</td>
<td>1</td>
<td>$|\Theta^{(1)}|_{1,1} \leq r_1$</td>
</tr>
<tr>
<td>Low-rank decomposition ($\Theta^* = (\Theta^{(1)})$)</td>
<td>1</td>
<td>$|\Theta^{(1)}|_{*} \leq r_1$</td>
</tr>
<tr>
<td>Sparse-plus-low-rank decomposition ($\Theta^* = (\Theta^{(1)}, \Theta^{(2)})$)</td>
<td>2</td>
<td>$|\Theta^{(1)}|<em>{1,1} \leq r_1$ and $|\Theta^{(2)}|</em>{*} \leq r_2$</td>
</tr>
</tbody>
</table>

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 statistics. 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 $x$ 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 $X = [0, b]$ for $b \in \mathbb{R}$, then $\phi_{\text{max}} = 2b^l$. If $\Theta^*$ has a sparse decomposition and $X = [0, b]$ for $b \in \mathbb{R}$, then $\mathcal{R}^*(\Phi(x)) \leq 2b^l$. Further, if $\Theta^*$ has a low-rank decomposition, $l = 2$, and $X = B(0, b)$ for $b \in \mathbb{R}$, then $\mathcal{R}^*(\Phi(x)) \leq 2(1 + b^2)$. Finally, if $\Theta^*$ has a sparse-plus-low-rank decomposition, $l = 3$, and $X = B(0, b)$ for $b \in \mathbb{R}$, then $\mathcal{R}^*(\Phi(x)) \leq (2b^2, 2 + 2b^2)$.

2. **Trigonometric statistics:** Suppose the natural statistics are sines and cosines of $x$ with $l$ different frequencies, i.e., $\sin(\sum_{i \in [p]} l_i x_i)$ and $\cos(\sum_{i \in [p]} l_i x_i)$ such that $l_i \in [l] \cup \{0\}$. For any $X \subset \mathbb{R}^p$, $\phi_{\text{max}} = 2$. If $\Theta^*$ has a sparse decomposition, then $\mathcal{R}^*(\Phi(x)) \leq 2$ for any $X \subset \mathbb{R}^p$.

Our framework also allows combinations of polynomial and trigonometric statistics (see Appendix [1] for details).

### 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 $x$ that the probability density $f_x(x; \Theta)$ is proportional to (see (9)).

**Definition 3.1 (The loss function).** Given $n$ samples $x^{(1)}, \ldots, x^{(n)}$ of $x$, 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

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

(10)

The proposed estimator $\hat{\Theta}_n$ produces an estimate of $\Theta^*$ by minimizing the loss function $\mathcal{L}_n(\Theta)$ over all natural parameters $\Theta$ satisfying Assumption 2.3, i.e.,

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

(11)

For any $\epsilon > 0$, $\hat{\Theta}_{\epsilon,n}$ is an $\epsilon$-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 $\Lambda$) and has efficient implementations for finding an $\epsilon$-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:** $\eta, T, \Lambda$

**Output:** $\hat{\Theta}_{\epsilon,n}$

**Initialization:** $\Theta(0) = 0$

1. for $t = 0, \ldots, T$
2. $\Theta(t+1) \leftarrow \arg \min_{\Theta \in \Lambda} \|\Theta(t) - \eta \nabla \mathcal{L}_n(\Theta(t)) - \Theta\|_T$
3. $\hat{\Theta}_{\epsilon,n} \leftarrow \Theta(T)$

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

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

$$
\tau \geq \frac{2k_1 k_2 k_3 \phi_{\text{max}}^2 \exp(\tau^T d)}{\epsilon} \|\hat{\Theta}_n\|_T^2.
$$

(12)

Further, ignoring the dependence on $k_3, \phi_{\text{max}}, \tau$ and $d$, $\tau$ in (12) scales as $O\left(\text{poly}\left(\frac{\log k_2}{\epsilon}\right)\right)$.

---

We believe that for polynomial and/or trigonometric natural statistics, Assumptions 2.2 and 2.3 would hold whenever the domain of $X$ is appropriately bounded.
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 $\mathcal{L}(\Theta)$. (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 $f_k(\cdot; \Theta^* - \Theta)$. Then, we establish consistency and asymptotic normality of our estimator. Finally, we provide non-asymptotic finite sample guarantees to recover $\Theta^*$.

1. Connection with MLE of $f_k(\cdot; \Theta^* - \Theta)$. 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 $f_k(x; \Theta^* - \Theta)$. 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 $\Theta \in \Lambda$, $\mathcal{L}(\Theta) = \mathbb{E} \left[ \exp \left( - \langle \langle \Theta, \Phi(x) \rangle \rangle \right) \right]$. The following result shows that the population version of the estimator in (11) is equivalent to the maximum likelihood estimator of $f_k(x; \Theta^* - \Theta)$.

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

$$\arg \min_{\Theta \in \Lambda} \mathcal{L}(\Theta) = \arg \min_{\Theta \in \Lambda} D(U_{X}(\cdot) \parallel f_k(\cdot; \Theta^* - \Theta)).$$

Further, the true parameter $\Theta^*$ 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 $f_k(\cdot; \Theta^* - \Theta)$ in terms of $\mathcal{L}(\Theta)$ (b) Next, we complete the proof by simplifying the KL-divergence between $U_{X}(\cdot)$ and $f_k(\cdot; \Theta^* - \Theta)$.

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 $f_k(\cdot; \Theta^* - \Theta)$ and not $f_k(\cdot; \Theta)$. Moreover, there is no clear connection between $\hat{\Theta}_n$ and the finite sample maximum likelihood estimate of $f_k(\cdot; \Theta)$ or $f_k(\cdot; \Theta^* - \Theta)$. 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 $\text{vec}(\Phi(x) \exp (- \langle \langle \Theta^*, \Phi(x) \rangle \rangle))$. Let $B(\Theta^*)$ denote the cross-covariance matrix of $\text{vec}(\Phi(x))$ and $\text{vec}(\Phi(x) \exp (- \langle \langle \Theta^*, \Phi(x) \rangle \rangle))$. Let $N(\mu, \Sigma)$ represent the multi-variate Gaussian distribution with mean vector $\mu$ and covariance matrix $\Sigma$.

**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 $\Theta^* \in \text{interior}(\Lambda)$ and $B(\Theta^*)$ is invertible, we have $\sqrt{n} \times \text{vec}(\hat{\Theta}_n - \Theta^*) \xrightarrow{d} N(\text{vec}(0), B(\Theta^*)^{-1} A(\Theta^*) B(\Theta^*)^{-1})$.

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 $\Theta^*$.

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

**Assumption 4.1.** (Positive eigenvalue of the autocorrelation matrix of $\Phi$) Let $\lambda_{\min}$ denote the minimum eigenvalue of $\mathbb{E}_x [\text{vec}(\Phi(x) \cdot \text{vec}(\Phi(x))^T]$. 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 $M \in \mathbb{R}^{k_1 \times k_2}$, there exists $g$ such that $\tilde{\mathcal{R}}(M) \leq gk_1k_2 \|M\|_{\max}$.

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

Theorem 4.3 below shows that, with enough samples, the \( \epsilon \)-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 \( \epsilon \)-optimal solution of \( \hat{\Theta}_n \) obtained from Algorithm 1 for \( \epsilon \) of the order \( O(\alpha^2 \lambda_{\text{min}}) \). 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 \( \|\hat{\Theta}_{\epsilon,n} - \Theta^*\|_T \leq \alpha \) with probability at least \( 1 - \delta \) as long as

\[
n \geq O\left(\frac{k_1^2 k_2^2}{\alpha^4 \lambda_{\text{min}}^2} \log \left(\frac{k_1 k_2}{\delta}\right)\right).
\]  

(13)

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

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

The computational cost of projection onto \( \Lambda \) i.e., \( c(\Lambda) \) is typically polynomial in \( k_1, k_2 \). In Appendix D, we provide the computational cost for the example constraints on the natural parameter \( \Theta \) 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 \( f_x(\cdot; \Theta) \). The MLE of \( f_x(\cdot; \Theta) \) minimizes the following loss function

\[
\min -\frac{1}{n} \sum_{t=1}^{n} \langle \Theta, \Phi(x^{(t)}) \rangle + \log \int_{x \in X} \exp \left( \langle \Theta, \Phi(x) \rangle \right) dx.
\]  

(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 converges 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 propose 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 \( \alpha \)-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
Assumption 4.1. For node-wise-sparse pairwise exponential family MRFs (e.g., Ising models), which is computationally efficient [14].

version of our estimator has a nice interpretation in terms of maximum likelihood estimation, the and asymptotic efficiency for this class of exponential family could be an interesting future direction. as asymptotically efficient. Investigating the possibility of a single estimator that achieves computational be connections with the Bregman score and this is an important direction for immediate future work.

finite sample version of our estimator does not have a similar interpretation. We believe there could multiplicative regularizing distribution from [41] could also be used. Second, while the population P support. While, conceptually, most non-compact distributions could be truncated by introducing a

Limitations and Future Work. First, in our current framework, we assume boundedness of the parameters associated with each node. Since the focus of our work is beyond the exponential families considered in this work (which may not be sparse or Gaussian). Sample Complexity. We do not assume p (the dimension of 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(poly(p)). Further, the 1/α^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/α^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 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:

P(|x_i| ≥ δ log γ) ≤ cγ^{−d} where c > 0 is a constant and γ > 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.
Acknowledgments and Disclosure of Funding

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


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is theoretical in nature.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Each theoretical result is accompanied with the assumptions required.
   (b) Did you include complete proofs of all theoretical results? [Yes] All the proofs are in the supplementary. Each theoretical result is followed by an appropriate reference to the corresponding proof. Each theoretical result is also followed by either the proof outline or the key ideas used in the proof.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]