Finite Sample Bounds for Learning with Score Matching

Devin Smedira SMEDIRA@MIT.EDU
Operations Research Center, Massachusetts Institute of Technology;
Theoretical Division, Los Alamos National Laboratory

Abhijith Jayakumar ABHIJITHJ@LANL.GOV
Theoretical Division, Los Alamos National Laboratory

Sidhant Misra SIDHANT@LANL.GOV
Theoretical Division, Los Alamos National Laboratory

Marc Vuffray VUFFRAY@LANL.GOV
Theoretical Division, Los Alamos National Laboratory

Andrey Y. Lokhov LOKHOV@LANL.GOV
Theoretical Division, Los Alamos National Laboratory

Editors: Steve Hanneke and Tor Lattimore

Abstract

Learning of continuous exponential family distributions with unbounded support remains an important area of research for both theory and applications in high-dimensional statistics. In recent years, score matching has become a widely used method for learning exponential families with continuous variables due to its computational ease when compared against maximum likelihood estimation. However, theoretical understanding of the statistical properties of score matching is still lacking. In this work, we provide a non-asymptotic sample complexity analysis for learning the structure of exponential families of polynomials with score matching. The derived sample bounds show a polynomial dependence on the model dimension. These bounds are the first of its kind, as all prior work has shown only asymptotic bounds on the sample complexity.

Keywords: Score Matching, Exponential Families, Model Selection, Structure Learning

1. Introduction

Learning of many-variable distributions is a fundamental primitive in machine learning. Often, it is natural to parametrize a target distribution via a class of energy functions E_θ as

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}$$

up to a normalizing constant Z_θ, reducing the problem of learning the distribution to learning the parameter θ of the energy function. Though statistically efficient, the standard Maximum-Likelihood estimation method applied to this problem suffers from a computational bottleneck by which the learner is forced to implicitly or explicitly compute the partition functions Z_θ of models in the hypothesis class (Montanari, 2015).

In recent years, score matching (Hyvärinen, 2005) has emerged as a popular method to tackle such learning tasks, primarily owing to it's computational tractability. Score matching works directly with the *score* of the distribution ($\nabla_x \log(p_\theta(x))$), minimizing the expected difference of this

quantity with the score of a parametrized hypothesis. While initially seeming impossible to compute without prior knowledge of the true model, this loss can be rewritten via integration by parts as something easily computable from samples. This technique has several benefits, first among them being that the reliance on derivatives of the Gibbs distribution avoids the difficulty of computing a partition function.

In this work, we study distributions parameterized by polynomial energy functions with learnable coefficients, called *exponential families*. Exponential families are a common family of parametric distributions which have been studied since the works of Darmois (1935), Koopman (1936), and Pitman (1936). Importantly for our designs, the score-matching loss on exponential families can be reduced to a fairly benign quadratic optimization problem over the polynomial coefficients. Our primary contribution is to provide a finite sample bound to learn the *structure* of exponential families, meaning the coefficients for the maximal monomials of the energy function, with score-matching. Learning such structural results expose full conditional independence and Markov property of the model. Additionally, we will show a finite sample bound for the total parameter recovery of models parameterized by multi-linear polynomials. In both cases, these bounds will be polynomial in the dimension of the model.

While many prior works have looked at the statistical properties of score matching, a few are highly relevant to this study of exponential families. Sriperumbudur et al. (2017) derives a model agnostic asymptotic sample convergence rate for the score matching estimator on exponential families. A more refined analysis of the model dependent properties of score matching is provided by Koehler et al. (2022) for the case of distribution families with good isoperimetric properties. These techniques are applied to exponential families in Pabbaraju et al. (2023) to derive an asymptotic sample bound which fully captures the dependence on the target family's dimension. In contrast, the aim of our present work is to prove first of their kind non-asymptotic sample complexity bounds for parameter recovery with score matching under natural model assumptions.

Unlike previous works on score-matching which revolve around the study of the restricted Poincaré constant to relate score-matching to the maximum-likelihood estimator, our results are derived from the explicit study of the curvature of the score-matching loss. We devise a method to relate local properties of the distribution to loss for the entire distribution. We expect this method to be easily generalized for other models with continuous energy models.

1.1. Further related work

Approximations to Maximum Likelihood A classical alternative to exact maximum likelihood for energy-based models is to replace the intractable log-partition gradient by a Markov chain estimate, as done in *contrastive divergence* (CD-k), introduced in the context of training restricted Boltzmann machines (Hinton, 2002). Another broad family of approximations replaces the exact likelihood by variational bounds or mean-field approximations that trade statistical efficiency for tractability (Wainwright and Jordan, 2008). In contrast, score matching sidesteps the partition function entirely by moving from likelihood to a score-based objective.

Discrete alphabets and graphical model learning For discrete distributions (e.g., Ising/Potts models), there is a large literature on algorithms that learn the energy that avoid the partition function. Prominent examples include logistic regression (Ravikumar et al., 2010; Wu et al., 2019), interaction screening (Vuffray et al., 2016, 2020), and Sparsitron (Klivans and Meka, 2017). In addition, ratio matching (Hyvärinen, 2007) can be viewed as a direct discrete analogue of score

matching that yields a tractable objective while retaining consistency under suitable conditions. However, the theory of ratio matching is not as well developed as the aforementioned examples. Conversely, several works attempted to generalize the estimators for discrete models to exponential families with continuous variables. (Shah et al., 2021a,b) applied the interaction screening technique to the setting of continuous exponential families with pairwise interactions, however restricted to distributions with bounded support, and showed an exponential dependence of sample complexity on the domain bound. (Ren et al., 2021) introduced a modification of the interaction screening estimator that works for exponential families with unbounded support and numerically compared it to the pseudolikelihood estimator, but did not provide the statistical complexity analysis of the estimator.

Score matching and generative modeling Score matching also plays a central role in modern diffusion/score-based generative modeling, where one learns (variants of) the score of noise-perturbed data distributions and then samples by simulating a reverse-time SDE (Song and Ermon, 2019; Song et al., 2021). There is a growing theory literature that studies when approximate scores yield guarantees on the distribution produced by the discretized reverse process (Azangulov et al., 2024; Oko et al., 2023; Tang and Yang, 2024; Yakovlev and Puchkin, 2025; Yakovlev et al., 2025). Lee et al. (2023), proves polynomial-time convergence guarantees for denoising diffusion / score-based generative modeling under very general assumptions on the data distribution, assuming L^2-accurate score estimates. Related convergence analyses include Wasserstein guarantees for broader classes of score-based models (Gao et al., 2023) and more recent results covering weaker smoothness regimes (Bruno and Sabanis, 2025).

2. Problem Formulation and Main Results

In this section, we will formulate our model selection problem for exponential families as well as introduce the score-matching loss which underlies our technique.

2.1. Exponential Families

A generic exponential family on $\mathbb{R}^n$ has the form $p_\theta(x) \propto h(x)\exp(\langle\theta, T(x)\rangle)$, where $h(x)$ is some base measure, θ is a vector of parameters, and $T(x)$ is a vector of basis functions. In this work, we will limit ourselves to exponential families with the basis functions in $T(x)$ consisting only of monomials in $x_1, \dots x_n$ with degree at most d.

We will use $\mathcal{K}$ to denote the *factor set* which indexes the set of the basis functions (as well as the parameters) for a given exponential family. For each $k \in \mathcal{K}$ we will let k_i denote the degree of x_i in the corresponding basis function $f_k(x) = \prod_{i=1}^n x_i^{k_i}$. We will use $\partial k = \{i | k_i > 0\}$ to denote the support of k, $w = \max_{k\in\mathcal{K}} |\partial k|$ to denote the *interaction order* of the model family and let $\mathcal{K}_i = \{k \in \mathcal{K} | i \in \partial k\}$. Throughout this work, we will use θ^* to denote the parameters of the true model we sample from. Further, we will assume there is some integer ℓ_1-bound B on the parameters associated with each variable, meaning $\sum_{k\in\mathcal{K}_i} |\theta^*_k| \le B$. Finally, our analysis will require bounds on the tail decay of the true model, so we will limit ourselves to distributions p_{θ^*} satisfying the following condition.

Condition 1. There exists some constant $k > 0$ and integer C_t satisfying $\max((\ln(2)/k)^{1/(d-1)}, 1) \le C_t \le e^n$ such that for any $s \ge C_t$, $\Pr(\|x\|_\infty > s) \le \exp(-ks^{d-1})$ when $x \sim p_{\theta^*}$.

This assumption is satisfied by several natural exponential families. For example, we have the following pertaining to the exponential families studied in Pabbaraju et al. (2023), proven in Appendix B.

Fact 1 *If $T(x)$ contains only monomials of degree at most $d-1$ and $h(x) = \exp\left(-\sum_{i=1}^{n} x_i^d\right)$, then p_{θ^*} satisfies Assumption 1 with $k = 1$ and $C_t = nB + 1$.*

2.2. The Structure Graph of Exponential Families

In order to define the structure and factor graph of a model, we will now introduce some additional notation pertaining to graphical models. The structure of a model is of particular interest in learning, as it encodes the conditional independence between variables in the model.

Given an exponential family with associated factor set $\mathcal{K}$ as above, the associated factor graph is a bipartite graph $G = ([n], \mathcal{K}, E)$ with an edge set

$$E = \{(i, k) \in [n] \times \mathcal{K} | i \in [k]\}. \tag{1}$$

We see from the above that an edge (i, k) occurs when the variable x_i appears in the associated basis function f_k. Thus, we can see that the neighborhood of any factor $k \in \mathcal{K}$ in the factor graph is precisely its support ∂k. Notice that the definition in 1 depends only on the set of basis functions $\mathcal{K}$ associated with our family of functions, and is independent of the true underlying model. The factor graph $G^* = ([n], \mathcal{K}^*, E^*)$ associated with a specific model p_{θ^*} in the family is the induced subgraph of G obtained by taking $\mathcal{K}^* = \{k \in \mathcal{K} | \theta_k^* \neq 0\}$.

For any given factor graph G, the *maximal factors* $\mathcal{M}_{\text{fac}}$ consists of every factor whose neighborhood is not strictly contained within the neighborhood of another factor,

$$\mathcal{M}_{\text{fac}}(G) = \{k \in \mathcal{K} | \ \nexists k' \in \mathcal{K}, \partial k \subset \partial k'\}.$$

It is possible that multiple distinct maximal factors will have the same neighborhood. For example if the basis function $x_1 x_2$ corresponds to a maximal factor, then the basis function $x_1^2 x_2^2$ will as well, as they have the same neighborhood in the factor graph. It will be convenient to define *maximal cliques* which correspond to the neighborhoods of all the maximal factors as

$$\mathcal{M}_{\text{cli}}(G) = \{c \in \mathcal{P}([n]) | \exists k \in \mathcal{M}_{\text{fac}}(G), c = \partial k\},$$

where $\mathcal{P}$ denotes the powerset. We will call the span of a clique c to be the set of all factors whose neighborhood is exactly c denoted by $[c]_{\text{sp}} = \{k \in \mathcal{M}_{\text{cli}} | c = \partial k\}$. Finally, we can define the structure $\mathcal{S}$ of a model to be the set of maximal cliques in the models factor graph G^*, meaning

$$\mathcal{S} = \mathcal{M}_{\text{cli}}(G^*).$$

2.3. Score Matching

Formally, the score matching loss for some parameter vector θ and sample x is

$$\mathcal{L}(\theta, x) = \frac{1}{2} \|\nabla_x \log p_{\theta^*}(x) - \nabla_x \log p_\theta(x)\|^2.$$

This defines the following loss function when averaged over the true distribution, $\mathcal{L}(\theta) = \mathbb{E}_{x\sim p_{\theta^*}}[\mathcal{L}_i(\theta, x)]$.

This loss is computable only with knowledge of the true parameters. It can be shown (Hyvärinen (2005)), under some mild assumptions on the decay of the true distribution, that $\mathcal{L}(\theta)$ is expressible in the following way up to a constant that does not depend on the optimization variables,

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim p_{\theta^*}} \mathrm{Tr}\nabla_x^2 \log p_\theta(x) + \frac{1}{2}\|\nabla_x \log p_\theta(x)\|^2 + C. \tag{1}$$

In this work, we will be interested primarily in analyzing the local score matching loss around each vertex $i \in [n]$, which we define as,

$$\mathcal{L}_i(\theta, x) := \frac{\partial}{\partial^2 x_i} \log p_\theta(x) + \frac{1}{2}\left(\frac{\partial}{\partial x_i} \log p_\theta(x)\right)^2. \tag{2}$$

We let $\mathcal{L}_i(\theta) = \mathbb{E}_{x\sim p_{\theta^*}}[\mathcal{L}_i(\theta, x)]$. Further, for any collection of samples $x^{(1)}, \dots x^{(M)}$, we let $\mathcal{L}_i(\theta, x^{(M)}) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_i(\theta, x^{(m)})$. In the case of exponential families, this loss function can be computed by a quadratic optimization problem.

2.4. Main Results

We are now ready to formally state our result on structure learning, which we prove in Section 4.

Theorem 1 (Family Structure Learning) *Fix some exponential family p_{θ^*} with base measure $h(x) = 1$. For a fixed index $i \in [n]$, let $\hat{\mathcal{K}}$ be the maximal factors with i as a neighbor in the family factor graph G, meaning $\hat{\mathcal{K}} = \{k \in \mathcal{M}_{\mathrm{fac}}(G) | i \in \partial k\}$. Further, for M independent samples $x_1, \dots x_M \sim p_{\theta^*}$, let $\hat{\theta}$ be the minimizer of $\mathcal{L}_i(\theta, x^{(M)})$ subject to $\sum_{k\in\mathcal{K}_j} |\theta_k| \le B$ for every index j. There exists some $M^* = (dBC_t^d)^{O(d^2 w)}$ such that for every $\rho \ge 1$ and $\epsilon \le 1$, we have that $(\theta^*_k - \hat{\theta}_k)^2 \le \epsilon$ for every $k \in \hat{\mathcal{K}}$ with probability greater than $1 - \frac{1}{\rho n C_t}$ when $M \ge \rho \frac{n^{d+1} M^*}{\epsilon^2}$.*

Notice that this theorem learns the parameters for *family* structure, not the structure of the specific model. In particular, we only learn $k \in \mathcal{M}_{\mathrm{fac}}(G)$ the maximal factors for the *family* factor graph G, not the model graph G^*. However, provided a sufficient number of samples, one can run an iterative algorithm to identify and remove cliques which are present in the family but not the particular model, allowing for the recovery of the *model* structure. This idea is made precise in Theorem 4. Though we present the theorem here assuming the score-matching minimizer can be computed exactly, this may be intractable in practice. A version of this result is presented in Appendix D which provides recovery guarantees so long as $\hat{\theta}$ has a loss sufficiently close to the minimum's.

We state Theorem 1 assuming the base measure $h(x) = 1$. However, the result can be applied to any distribution satisfying Condition 1 so long as its base measure can be encoded as a polynomial exponential family. In particular, this theorem applies directly to the family from Fact 1, as the base measure consists of bounded degree monomials which do not supersede any maximal factors

When each basis function f_k is linear in each variable, it is possible to show a stronger result recovering the entire parameter vector θ^* using score matching. In particular, we define a model p_{θ^*} to be a *multi-linear exponential model* if $k_i \in \{0, 1\}$ for every $k \in \mathcal{K}$ and every $i \in [n]$. Under this assumption, we have the following theorem proven in Section 6.

Theorem 2 (Total Recovery of Multi-Linear Models) *Suppose p_{θ^*} is a multi-linear exponential model with base measure $h(x) = \exp\left(-\sum_{i=1}^{n} x_i^d\right)$. For M independent samples $x_1, \dots x_M \sim p_{\theta^*}$, let $\hat{\theta}$ be the minimizer of $\mathcal{L}_i(\theta, x^{(M)})$ subject to $\sum_{k \in \mathcal{K}_j} |\theta_k| \leq B$ for every index j. There exists some $M^* = (BC_t^d)^{O(d)}$ such that for every $\rho \geq 1$ and $\epsilon \leq 1$, $(\theta_k^* - \hat{\theta}_k)^2 \leq \epsilon$ for every $k \in \mathcal{K}_i$ with probability greater than $1 - \frac{1}{\rho n C_t}$ when $M \geq \rho \frac{n^{2d+1} M^*}{\epsilon^2}$.*

3. Curvature Analysis

This section will focus on analyzing the curvature of the loss function $\mathcal{L}_i$ and deriving a few key properties which will be useful in our proofs. Bounding the curvature of the loss function allows us to show that a difference in loss implies a minimal distance between parameter vectors in euclidean space. This is the key ingredient in our proof of the finite sample bound.

Suppose $\hat{\theta}$ is some candidate parameter vector. We define $\Delta = \hat{\theta} - \theta^*$, $E_i(x, \theta) = \frac{\partial}{\partial x_i} \log p_{\theta(x)} = \frac{\partial}{\partial x_i} \sum_{k \in \mathcal{K}} \theta_k f_k(x)$, and $\phi(y) = y^2/2$. Notice that E_i is linear in each θ_k term. We have that

$$\begin{aligned}
\mathcal{L}_i(\theta^*) - \mathcal{L}_i(\hat{\theta}) &= \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \theta^*)) + \frac{\partial}{\partial^2 x_i} \log p_{\theta^*} \right] - \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \hat{\theta})) + \frac{\partial}{\partial^2 x_i} \log p_{\hat{\theta}} \right] \\
&= \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \theta^*)) \right] - \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \hat{\theta})) \right] - \sum_{k \in \mathcal{K}} \Delta_k \mathbb{E}_{x \sim p_{\theta^*}} \left[\frac{\partial}{\partial^2 x_i} f_k(x) \right].
\end{aligned}$$

Further, we have that

$$\begin{aligned}
\langle \Delta, \nabla \mathcal{L}_i(\theta^*) \rangle &= \mathbb{E}_{x \sim p_{\theta^*}} \phi'(E_i(x, \theta^*)) \left(\sum_{k \in \mathcal{K}} \frac{\partial}{\partial \theta_{\mathbf{k}} \partial x_i} \log p_{\theta^*}(x) \Delta_k \right) - \sum_{k \in \mathcal{K}} \Delta_k \mathbb{E}_{x \sim p_{\theta^*}} \left[\frac{\partial}{\partial^2 x_i} f_k(x) \right] \\
&= \mathbb{E}_{x \sim p_{\theta^*}} \phi'(E_i(x, \theta^*)) E_i(x, \Delta) - \sum_{k \in \mathcal{K}} \Delta_k \mathbb{E}_{x \sim p_{\theta^*}} \left[\frac{\partial}{\partial^2 x_i} f_k(x) \right].
\end{aligned}$$

We define the curvature of a function ψ as it's deviation from it's linear approximation at a point, $\delta\psi(\Delta, x) = \psi(x + \Delta) - \psi(x) - \langle \Delta, \nabla \psi(x) \rangle$. Combining this definition with the previous two equations, we see that

$$\begin{aligned}
\delta \mathcal{L}_i(\Delta, \theta^*) &= \mathcal{L}_i(\hat{\theta}) - \mathcal{L}_i(\theta^*) - \langle \Delta, \nabla \mathcal{L}_i(\theta^*) \rangle \\
&= \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \hat{\theta})) \right] - \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \theta^*)) \right] - \mathbb{E}_{x \sim p_{\theta^*}} \phi'(E_i(x, \theta^*)) E_i(x, \Delta) \\
&= \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \theta^* + \Delta)) \right] - \mathbb{E}_{x \sim p_{\theta^*}} \left[\phi(E_i(x, \theta^*)) \right] - \mathbb{E}_{x \sim p_{\theta^*}} \phi'(E_i(x, \theta^*)) E_i(x, \Delta) \\
&= \mathbb{E}_{x \sim p_{\theta^*}} \delta\phi(E_i(x, \theta^*), E_i(x, \Delta)) \\
&= \frac{1}{2} \mathbb{E}_{x \sim p_{\theta^*}} E_i[x, \Delta]^2, \qquad (3)
\end{aligned}$$

where the final equality follows from $\delta\phi(x, \Delta) = (x + \Delta)^2/2 - x^2/2 - \Delta x = \Delta^2/2$. Let $p_t(x)$ denote the distribution $p_{\theta^*}(x)$ conditioned on $\|x\|_\infty \leq C_t$. By applying Condition 1 to equation 3, we see that

$$\frac{1}{2} \mathbb{E}_{x \sim p_{\theta^*}} E_i[x, \Delta]^2 \geq \frac{1}{4} \mathbb{E}_{x \sim p_t} E_i[x, \Delta]^2, \qquad (4)$$

which will provide a more convenient reference distribution to use. Showing a bound on the above in terms of Δ will be sufficient to prove our main results. To see why, suppose we have some collection of samples $x^{(M)}$ and that $\hat{\theta}$ is the minimizer of $\mathcal{L}_i(\theta, x^{(M)})$. Then, we have that

$$\begin{aligned} 0 \geq \mathcal{L}_i(\hat{\theta}, x^{(M)}) - \mathcal{L}_i(\theta^*, x^{(M)}) &= \delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) + \langle \Delta, \nabla_\theta \mathcal{L}_i(\theta^*, x^{(M)}) \rangle \\ &\geq \delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) - 2B\|\nabla_\theta \mathcal{L}_i(\theta^*, x^{(M)})\|_\infty. \end{aligned} \tag{5}$$

In Appendix C, we prove a high probability bound on $\|\nabla_\theta \mathcal{L}_i(\theta^*, x^{(M)})\|_\infty$ under Assumption 1 and that $\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)})$ concentrates around its expectation using standard techniques. This culminates in Lemma 9, which allows us to derive our main theorems directly from a bound on $\mathbb{E}_{x\sim p_t} E_i[x, \Delta]^2$. Thus, our primary technical challenge shifts to bounding the right hand side of Equation (4).

4. Recovery of Family Structure

This section will be dedicated to proving Theorem 1, a finite sample bound to recover a model's parameters corresponding to the *model family's* maximal factors. Specifically, this section will work towards validating Equation (12), which immediately implies Theorem 1 when combined with Lemma 9, a concentration result for the model curvature. In what follows, we will assume that we have a specific model p_{θ^*} parameterized by the vector θ^* that lies within an exponential model family as defined above.

4.1. Grid Points

We will now fix some specific maximal clique $c \ni i$. In order to obtain structure recovery, we will need to lower bound the expression in Equation (4) in terms of Δ_k for each $k \in [c]_{\text{sp}}$. To this end, we will first observe that

$$\mathbb{E}_{x\sim p_t} E_i[x, \Delta]^2 = \mathbb{E}_{x_{\setminus c}\sim p_t}[\mathbb{E}_{x_c|x_{\setminus c}\sim p_t}[E_i[x, \Delta]^2]]. \tag{6}$$

This section will be dedicated to finding some simple and separable distribution $q_{x_{\setminus c}}(x_c)$ which we can take the inner expectation with respect to instead.

Fix some value for $x_{\setminus c}$ satisfying $\|x_{\setminus c}\|_\infty \leq C_t$. Let $\rho(x_c|x_{\setminus c})$ be the probability density function for the distribution $x_c|x_{\setminus c} \sim p_t$ defined as

$$\rho(x_c|x_{\setminus c}) = \frac{1}{Z(x_{\setminus c})} \exp\left(\sum_{k\in\mathcal{K}} \theta^*_k f_k(x) - \sum_{i\in c} x_i^d\right) \mathbf{1}(\|x_c\|_\infty \leq C_t).$$

We let $\hat{x}_c = \text{argmax}_{x_c} \rho(x_c|x_{\setminus c})$ and take $\gamma = d^2 C_t^{2d} B$. We will further take l_i to be some set of integers such that $\frac{l_i}{\gamma} \leq (\hat{x}_c)_i \leq \frac{l_i+1}{\gamma}$ for each $i \in c$. Take any alternative x_c satisfying $\frac{l_i}{\gamma} \leq (x_c)_i \leq \frac{l_i+1}{\gamma}$ and let $\epsilon = \hat{x}_c - x_c$. We see that

$$|f_k(\hat{x}) - f_k(x)| = \left|\prod_{i\notin c} x_i^{k_i} \left(\prod_{i\in c} \hat{x}_i^{k_i} - \prod_{i\in c} x_i^{k_i}\right)\right|$$

$$
\begin{aligned}
&= \prod_{i \notin c} |x_i|^{k_i} \left(\prod_{i \in c} (|\hat{x}_i| + |\epsilon_i|)^{k_i} - \prod_{i \in c} |\hat{x}_i|^{k_i} \right) \\
&\leq \prod_{i \notin c} C_t^{k_i} \left(\prod_{i \in c} (C_t + \frac{1}{\gamma})^{k_i} - \prod_{i \in c} C_t^{k_i} \right) \\
&\leq C_t^d ((C_t + \frac{1}{\gamma})^d - C_t^d) \\
&\leq \frac{dC_t^{2d}}{\gamma} + \frac{2^d C_t^{2d}}{\gamma^2} \\
&\leq \frac{1}{dB}.
\end{aligned}
$$

Thus, we conclude that

$$
\begin{aligned}
\frac{\rho(x_c|x_{\setminus c})}{\rho(\hat{x}_c|x_{\setminus c})} &= \exp\left(\sum_{k \in \mathcal{K}} \theta_k^* (f_k(x) - f_k(\hat{x})) \right) \\
&\geq \exp\left(- \sum_{k \in \mathcal{K}} \theta_k^* |f_k(x) - f_k(\hat{x})| \right) \\
&\geq \exp\left(\frac{-1}{dB} (\sum_{k \in \mathcal{K}, \partial k \ni i} \theta_k^*) \right) \\
&\geq e^{-1}.
\end{aligned}
$$

Next, we see that

$$
1 = \int_{\|x_c'\|_\infty \leq C_t} \rho(x_c'|x_{\setminus c}) dx \leq \int_{\|x_c'\|_\infty \leq C_t} \rho(\hat{x}_c|x_{\setminus c}) dx = C_t^d \rho(\hat{x}_c|x_{\setminus c}),
$$

which in turn implies that $\rho(x_c|x_{\setminus c}) \geq \frac{1}{eC_t^d}$. Therefore, we define our *centering distribution* $q_{x_{\setminus c}}(x_c)$ to be the uniform probability distribution over all x_c satisfying $\frac{l_i}{\gamma} \leq (x_c)_i \leq \frac{l_i+1}{\gamma}$, so that we have

$$
\mathbb{E}_{x_{\setminus c} \sim p_t} [\mathbb{E}_{x_c|x_{\setminus c} \sim p_t} [E_i[x, \Delta]^2]] \geq \frac{1}{eC_t^d} \mathbb{E}_{x_{\setminus c} \sim p_t} [\mathbb{E}_{x_c \sim q_{x_{\setminus c}}} [E_i[x, \Delta]^2]], \tag{7}
$$

fulfilling our initial goal. For the rest of the section, we will refer to $q_{x_{\setminus c}}$ simply as q for simplicity.

4.2. Bounds for Centered Basis Function

In the course of the following proof, we will want to work with basis functions which are centered with respect to our centering distributions q. For each factor k where $\partial k = c$, we will define a corresponding *mostly centered basis function* $h_{i,k}$ as

$$
h_{i,k}(x) = k_i x_i^{k_i - 1} \sum_{r \in \mathcal{P}(c \setminus i)} (-1)^{|r|} \mathbb{E}_{x_r \sim q^{|r|}} [f_k(x)] = k_i x_i^{k_i - 1} \prod_{i \in c \setminus i} (x_i^{k_i} - \mathbb{E}_{x_i \sim q} [x_i^{k_i}]).
$$

We do not center around the variable x_i above as it is not necessary for the remainder of the proof, and doing so would force $h_{i,k} = 0$ whenever $k_i = 1$. We introduce the following Lemma.

Lemma 3 *For any reference index i, any maximal clique $c \ni i$, the Nonsingular Parametrization of Cliques constant B_{NPC}, defined as*

$$B_{\mathrm{NPC}} = \min_{\|x\|_2=1} \mathbb{E}_q \left[\left(\sum_{k \in [c]_{\mathrm{sp}}} x_k h_k(x) \right)^2 \right],$$

will satisfy $B_{\mathrm{NPC}} \geq \exp(-O(d^2|c| \log \gamma))$.

Proof We first define the matrix $M_{k,k'} = \mathbb{E}_q[h_k h_{k'}]$ for each $k, k' \in [c]_{\mathrm{sp}}$, and observe that

$$\left(\sum_{k \in [c]_{\mathrm{sp}}} x_k h_k(x) \right)^2 = x_k^T M x_k,$$

which implies $B_{\mathrm{NPC}} = \lambda_{\min}(M)$. For each index $j \neq i$, we now define a new $d \times d$ matrix A^j where

$$\begin{aligned} A^j_{m,n} &= \mathbb{E}_{x \sim q}[(x^m - \mathbb{E}_q[x^m])(x^n - \mathbb{E}_q[x^n])] \\ &= \frac{(l_j+1)^{m+n+1} - l_j^{m+n+1}}{\gamma^{m+n+1}(m+n+1)} - \frac{((l_j+1)^{m+1} - l_j^{m+1})((l_j+1)^{n+1} - l_j^{n+1})}{\gamma^{m+1}\gamma^{n+1}(m+1)(n+1)}. \end{aligned}$$

This matrix A^j will have a few key properties. First, A^j is positive-definite, as $y^T A^j y = \mathbb{E}_q[(\sum_j y_j(x^j - \mathbb{E}_q[x^j])^2] > 0$. Second, $\lambda_{\max}(A_j) \leq d$ since each entry in A^j is bounded above by 1. Finally, $\gamma^{2d+1}(2d+1)!(d+1)!A^j$ has only integer entries, since γ is an integer by the assumption that C_t and B are integers. And, since integer matrices have integer determinants, this implies it has a determinant of at least 1 when combined with the positive-definiteness. Thus, since we know that $\gamma > d$, we have $\det(A^j) \geq \exp(-O(d^2 \log \gamma))$, ultimately implying

$$\lambda_{\min}(A^j) \geq \frac{\det(A^j)}{\lambda_{\max}(A^j)^{d-1}} \geq \frac{\det(A^j)}{d^{d-1}} = \exp(-O(d^2 \log \gamma)).$$

For index i, we will define the matrix A^i such that $A^i_{m,n} = \mathbb{E}_q[x^{m+n}]$, and we follow an analogous argument to show $\lambda_{\min}(A^i) \geq \exp(-O(d^2 \log \gamma))$. Fixing any particular k and k', we see that

$$\begin{aligned} M_{k,k'} =& \mathbb{E}_{x \sim q}[h_k(x) h_{k'}(x)] \\ =& \mathbb{E}_q \left[x_i^{k_i + k_i' - 2} \prod_{j \in c \setminus i} (x_j^{k_j} - \mathbb{E}_q[x_j^{k_j}])(x_j^{k_j'} - \mathbb{E}_q[x_j^{k_j'}]) \right] \\ =& \mathbb{E}_q[x_i^{k_i + k_i' - 2}] \prod_{j \in c \setminus i} \mathbb{E}_q \left[(x_j^{k_j} - \mathbb{E}_q[x_j^{k_j}])(x_j^{k_j'} - \mathbb{E}_q[x_j^{k_j'}]) \right] \\ =& \prod_{j \in c} A^j_{k_j, k_j'}. \end{aligned}$$

Thus, M will be a proper sub-matrix of $\bigotimes_{j \in c} A^j$, implying that

$$\lambda_{\min}(M) \geq \prod_{j \in c} \lambda_{\min}(A^j) = \exp(-O(d^2|c| \log \gamma)).$$

■

4.3. Curvature Bounds for Maximal Cliques

This section will be dedicated to completing our bound on the expected curvature. For convenience, we will let $f_{i,k}(x) = \frac{\partial}{\partial x_i} f_k(x)$. We know $f_{i,k}(x) = h_{i,k}(x) + R_{i,k}(x)$ for $\partial k = c$, where

$$R_{i,k}(x) = \sum_{r \in \mathcal{P}(c \setminus i) \setminus \emptyset} (-1)^{|r|} \mathbb{E}_{x_r \sim q}[f_{i,k}(x)].$$

Picking up from Equation (7), we see that

$$\mathbb{E}_{x_c \sim q}\left[\left(\sum_{k \in \mathcal{K}_i} \Delta_k f_{i,k}(x)\right)^2\right] = \mathbb{E}_{x_c \sim q}\left[\left(\sum_{k \in [c]_{\text{sp}}} \Delta_k h_{i,k}(x)\right)^2\right] \tag{8}$$

$$+ 2 \sum_{k \in [c]_{\text{sp}}} \sum_{l' \in \mathcal{K}_i \setminus [c]_{\text{sp}}} \Delta_k \Delta_{k'} \mathbb{E}_{x_c \sim q}[h_{i,k}(x) f_{i,k'}(x)] \tag{9}$$

$$+ \sum_{k,k' \in [c]_{\text{sp}}} \Delta_k \Delta_{k'} \mathbb{E}_{x_c \sim q}[h_{i,k}(x) R_{i,k'}(x)] \tag{10}$$

$$+ \mathbb{E}_{x_c \sim q}\left[\left(\sum_{k \in \mathcal{K}_i \setminus [c]_{\text{sp}}} \Delta_k f_{i,k}(x) + \sum_{k \in [c]_{\text{sp}}} \Delta_k R_{i,k}(x)\right)^2\right]. \tag{11}$$

We now look at the contribution from each line of the above. In Equation (9), for any given k and k', there must be some index $j \neq i$ where $j \in c$ and $j \notin \partial k'$. Since $f_{i,k'}$ will not depend on x_j, we see

$$\mathbb{E}_{x_c \sim q}[h_{i,k}(x) f_{i,k'}(x)] = \mathbb{E}_{x_{c \setminus j} \sim q}[f_{i,k'}(x) \mathbb{E}_{x_j \sim q}[h_{i,k}(x)]] = 0,$$

meaning the contribution from line (9) is 0. Expanding the definition of $R_{i,k'}$ in line (10), we see

$$\mathbb{E}_{x_c \sim q}[h_{i,k}(x) R_{i,k'}(x)] = - \sum_{r \in \mathcal{P}(c \setminus i) \setminus \emptyset} (-1)^{|r|} \mathbb{E}_{x_c \sim q}[h_{i,k}(x) \mathbb{E}_{x_r \sim q}[f_{i,k'}(x)]] = 0,$$

since r always contains an index $j \neq i$ which $\mathbb{E}_{x_r \sim q}[f_{i,k'}(x)]$ will not depend on, implying the contribution from line (10) is also 0. Since Equation (11) is the expectation of a squared value, its contribution must be strictly non-negative. Thus, we conclude that

$$\mathbb{E}_{x_c \sim q}\left[\left(\sum_{k \in \mathcal{K}_i} \Delta_k f_{i,k}(x)\right)^2\right] \geq \mathbb{E}_{x_c \sim q}\left[\left(\sum_{k \in [c]_{\text{sp}}} \Delta_k h_{i,k}(x)\right)^2\right] \geq B_{\text{NPC}} \sum_{k \in [c]_{\text{sp}}} \Delta_k^2,$$

where B_{NPC} is defined in Lemma 3. Combining the above with Equations (6) and (7), we conclude that

$$\mathbb{E}_{x \sim p_t} E_i(x, \Delta)^2 \geq \frac{B_{\text{NPC}}}{e C_t^d} \sum_{k \in [c]_{\text{sp}}} \Delta_k^2. \tag{12}$$

The above equation, when combined with Lemma 9, immediately implies Theorem 1.

5. Recovery of Model Structure

In this section, we will present an approach to strengthen Theorem 1 to allow for the identification of the structure graph for a specific model and recover the parameter weights for the corresponding maximal factors. Suppose we have some $2\sqrt{\epsilon}$-lower bound with $\epsilon \leq 1$ on the strength of parameters associated with any maximal factor of the model, meaning that $\theta^*_k \geq 2\sqrt{\epsilon}$ for every $k \in \mathcal{M}_{\text{fac}}(G^*)$. Then, for sufficiently many samples M, Algorithm 1 will be able to recover both the structure $\mathcal{S}$ of the underlying model and the associated parameter θ^*_k for the maximal factors k of the model. In particular, we have the following theorem.

Theorem 4 (Iterative Model Structure Learning) *The output $(\hat{\mathcal{S}}, \hat{\theta}_i)$ of Algorithm 1 satisfies $\hat{\mathcal{S}} = \mathcal{S}$ and $\left((\hat{\theta}_i)_k - \theta^*_k\right)^2 \leq \epsilon$ with probability greater than $1 - \frac{1}{\rho C_t}$ as long as $M \geq \rho \frac{wn^{2d+1}M^*}{\epsilon^2}$, where w is the interaction order of the model family and $M^* = (dBC_t^d)^{O(d^2w)}$ is defined as in Theorem 1.*

Proof By Theorem 1, we know that each θ^s_i defined on line 8 of Algorithm 1 will satisfy

$$\left((\hat{\theta}^s_i)_k - \theta^*_k\right)^2 \leq \epsilon \tag{13}$$

for each $k \in \hat{\mathcal{K}}_i$ with probability at least $1 - \frac{1}{\rho n w C_t}$, where $\hat{\mathcal{K}}_i$ is defined on line 7. Thus, by the union bound, the probability that Equation (13) is satisfied by every $\hat{\theta}^s_i$ is at least $1 - \frac{1}{\rho C_t}$, since there are w different values of s and n unique values of i.

Now, suppose Equation (13) is satisfied for every $\hat{\theta}^s_i$. Take some clique $c \in \mathcal{M}_{\text{cli}}(G^s)$ which is not in $\mathcal{M}_{\text{cli}}(G^*)$. For each $k \in [c]_{\text{sp}}$, $\theta^*_k = 0$, meaning that $\hat{(\theta^s_i)}^2_k \leq \epsilon$. Thus, every element of $[c]_{\text{sp}}$ will be absent from $\mathcal{K}^{s+1}$, meaning that $c \notin \mathcal{M}_{\text{cli}}(G^{s+1})$. Therefore, every clique $c \in \mathcal{M}_{\text{cli}}(G^s)$ with $|c| > m - s$ must also be in $\mathcal{M}_{\text{cli}}(G^*)$, since it would have otherwise been removed in the previous iteration. We conclude that $\mathcal{S} = \mathcal{M}_{\text{cli}}(G^w) = \hat{S}$ and $\mathcal{M}_{\text{fac}}(G^w) \supseteq \mathcal{M}_{\text{fac}}(G^*)$, proving the theorem. ■

6. Total Recovery of Multi-linear Models

This section will be dedicated to proving Theorem 2, a finite sample bound for the total recovery of multi-linear model. As in Section 4, we will prove a lower bound on $\mathbb{E}_{x\sim p_t} E_i(x, \Delta)^2$ in Equation (14) which immediately implies Theorem 2 when combined with Lemma 9. We will now assume that p_{θ^*} is a multi-linear model, meaning that for each $k \in \mathcal{K}$, $k_j \in \{0, 1\}$ for every $j \in [n]$, with base measure $h(x) = \exp\left(-\sum_{i=1}^{n} x_i^d\right)$. We will fix both some index i and a second index $j \neq i$ and continue from Equation (4). By the Law of Conditional Variances, we have that

$$\begin{aligned}\mathbb{E}_{x\sim p_t} E_i(x, \Delta)^2 \geq \text{Var}_{\text{x}\sim \text{p}_\text{t}}(\text{E}_\text{i}(\text{x}, \Delta)) &\geq \mathbb{E}_{x_{\setminus j}\sim p_t} \text{Var}_{\text{x}_\text{j}|\text{x}_{\setminus\text{j}}\sim \text{p}_\text{t}} \text{E}_\text{i}(\text{x}, \Delta) \\ &= \mathbb{E}_{x_{\setminus j}\sim p_t} \left(\sum_{k\ni j,i} \Delta_k x_{k\setminus i,j}\right)^2 \text{Var}_{\text{x}_\text{j}|\text{x}_{\setminus\text{j}}\sim \text{p}_\text{t}} \text{x}_\text{j}.\end{aligned}$$

Algorithm 1 An Iterative Algorithm for Model Structure Recovery

Let $x_1, \dots x_M \sim p_{\theta^*}$ be independent samples
Let G be the bipartite factor graph for the model family
Let $\mathcal{K}^0 \leftarrow \mathcal{K}$ be the current set of potential factors
for $s = 0, \dots w$ **do**
  Let $G^s \leftarrow G[\mathcal{K}^s]$ be the subgraph of G induced by $\mathcal{K}^s$
  for $i \in [n]$ **do**
    Let $\hat{\mathcal{K}}_i \leftarrow \{k \in \mathcal{M}_{\text{fac}}(G^s) | i \in \partial k\}$
    Let $\hat{\theta}_i^s \leftarrow \text{argmin}_\theta(\mathcal{L}_i(\theta, x^{(M)}))$ subject to $\sum_{k \in \mathcal{K}_j} |\theta_k| \leq B$ for all j with factor set $\mathcal{K}^s$
  end for
  Let $\mathcal{N}^s \leftarrow \emptyset$
  for $c \in \mathcal{M}_{\text{cli}}(G^s)$ **do**
    if $\nexists k \in [c]_{\text{sp}}$ where $\hat{(\theta_i^s)}_k > \epsilon$ for some i **then**
      $\mathcal{N}^s \leftarrow \mathcal{N}^s \cup [c]_{\text{sp}}$
    end if
  end for
  Let $\mathcal{K}^{s+1} \leftarrow \mathcal{K}^s \setminus \mathcal{N}^s$ be the new set of potential factors
end for
return $\hat{\mathcal{S}} = \mathcal{M}_{\text{cli}}(G^w)$ and $\hat{\theta}_i = \hat{\theta}_i^w$

We now observe that the variable $x_j | x_{\setminus j} \propto \exp(\eta x_j - x_j^d)$ on $[-C_t, C_t]$ for some η dependent on $x_{\setminus j}$ and θ^*. In particular, based on the structure of the function, we know $|\eta| \leq BC_t^{d-1}$. Thus, by applying Lemma 6, a bound on the variance of $\exp(\eta x_j - x_j^d)$ in terms of η, we have

$$\mathbb{E}_{x \sim p_t} E_i(x, \Delta)^2 \geq \left(\frac{1}{2eBC_t^{d-1}}\right)^2 \mathbb{E}_{x_{\setminus j} \sim p_t} \left(\sum_{k \ni j,i} \Delta_k x_{k \setminus i,j}\right)^2.$$

We can generalize this argument. Take any index set $\mathbf{j} \not\ni i$, and any index $l \notin \mathbf{j} \cup \{i\}$. We will have

$$\begin{aligned}
\mathbb{E}_{x_{\setminus \mathbf{j}} \sim p_t} \left(\sum_{k \ni \mathbf{j},i} \Delta_k x_{k \setminus \{i,\mathbf{j}\}}\right)^2 &\geq \mathbb{E}_{x_{\setminus \mathbf{j},l} \sim p_t} \text{Var}_{\mathrm{x_l} | \mathrm{x}_{\setminus \mathbf{j},\mathrm{l}} \sim \mathrm{p_t}} \left(\sum_{\mathrm{k} \ni \mathbf{j},\mathrm{i}} \Delta_{\mathrm{k}} \mathrm{x}_{\mathrm{k} \setminus \{\mathrm{i},\mathbf{j}\}}\right) \\
&= \mathbb{E}_{x_{\mathbf{j}} \sim p_t} \left[\mathbb{E}_{x_{\setminus \mathbf{j},l} | x_{\mathbf{j}} \sim p_t} \left(\sum_{k \ni \mathbf{j},i,l} \Delta_k x_{k \setminus i,l,\mathbf{j}}\right) \text{Var}_{\mathrm{x_l} | \mathrm{x}_{\setminus \mathrm{l}} \sim \mathrm{p_t}}(\mathrm{x_l})\right] \\
&\geq \left(\frac{1}{2eBC_t^{d-1}}\right)^2 \mathbb{E}_{x_{\setminus \mathbf{j},l} \sim p_t} \left(\sum_{k \ni \mathbf{j},i,l} \Delta_k x_{k \setminus i,l,\mathbf{j}}\right).
\end{aligned}$$

Applying this logic repeatedly, we can see that for any k where $i \in \partial k$,

$$\mathbb{E}_{x \sim p_t} E_i(x, \Delta)^2 \geq \left(\frac{1}{2eBC_t^{d-1}}\right)^{2(d-1)} \Delta_k^2. \tag{14}$$

The proof of Theorem 2 follows from plugging Equation (14) into our curvature Lemma 9.

7. Conclusion

We have provided the first finite sample complexity bound for learning the structure of exponential families parametrized by bounded-degree polynomials using score matching. The sample bounds scale polynomially in the the model dimension and thus establish strong statistical merits of score matching in addition to its known computational properties. Information-theoretic bounds for recovery of exponential families with continuous variables are currently unknown. An important open question left for future exploration includes deriving such a lower bound, and comparing the rates to the scalings derived in this work. Our work establishes the properties of the optimal solution to the convex score matching objective under finite samples. It would be useful to complement our sample complexity analysis with the convergence analysis of gradient descent methods and establish the optimal rates. Finally, it would be interesting to extend our analysis to the families of distributions with general parametrization of the energy function beyond polynomials.

Acknowledgment

This work has been supported by the U.S. Department of Energy/Office of Science Advanced Scientific Computing Research Program.

References

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Stefano Bruno and Sotirios Sabanis. Wasserstein convergence of score-based generative models under semiconvexity and discontinuous gradients. arXiv preprint arXiv:2505.03432, 2025.

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

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. arXiv preprint arXiv:2311.11003, 2023.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. In *Neural Computation*, 2002.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 946–985, 2023.

Andrea Montanari. Computational Implications of Reducing Data to Sufficient Statistics, July 2015. URL http://arxiv.org/abs/1409.3821. arXiv:1409.3821.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36:61306–61326, 2023.

Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, volume 32, pages 567–579. Cambridge University Press, 1936.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Christopher X Ren, Sidhant Misra, Marc Vuffray, and Andrey Y Lokhov. Learning continuous exponential families beyond gaussian. *arXiv preprint arXiv:2102.09198*, 2021.

Abhin Shah, Devavrat Shah, and Gregory Wornell. A computationally efficient method for learning exponential family distributions. *Advances in neural information processing systems*, 34:15841–15854, 2021a.

Abhin Shah, Devavrat Shah, and Gregory Wornell. On learning continuous pairwise markov random fields. In *International conference on artificial intelligence and statistics*, pages 1153–1161. PMLR, 2021b.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International conference on artificial intelligence and statistics*, pages 1648–1656. PMLR, 2024.

Marc Vuffray, Andrey Y. Lokhov, Sidhant Misra, and Michael Chertkov. Interaction screening: Efficient and robust learning of ising models. arXiv preprint arXiv:1602.07014, 2016.

Marc Vuffray, Sidhant Misra, and Andrey Lokhov. Efficient learning of discrete graphical models. *Advances in Neural Information Processing Systems*, 33:13575–13585, 2020.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 12 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL https://doi.org/10.1561/2200000001.

Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. *Advances in Neural Information Processing Systems*, 32, 2019.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 5824–5891. PMLR, 2025.

Konstantin Yakovlev, Anna Markovich, and Nikita Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density hessian estimation. *arXiv preprint arXiv:2512.24378*, 2025.

Appendix A. Distribution Results

This section is dedicated to the proof of two useful technical lemmas.

Lemma 5 *Let α, β be two probability distributions supported on an interval S with pdfs f_α, f_β respectively. Suppose there exists some measurable set A where for any $x \in A$ and $y \in S \setminus A$ we have that $\frac{f_\alpha(x)}{f_\alpha(y)} \geq \frac{f_\beta(x)}{f_\beta(y)}$. Then,* $\Pr(\alpha \in A) \geq \Pr(\beta \in A)$.

Proof Rearranging, we have that $f_\alpha(x) f_\beta(y) \geq f_\beta(x) f_\alpha(y)$ for any $x \in A, y \in S \setminus A$. Thus we can compute

$$\begin{aligned}
\Pr(\alpha \in A) &= \int_{x \in A} f_\alpha(x) dx \\
&= \int_{x \in A} \int_{y \in S} f_\alpha(x) f_\beta(y) dy dx \\
&\geq \int_{x \in A} \left(\int_{y \in S \setminus A} f_\beta(x) f_\alpha(y) dy + \int_{y \in A} f_\alpha(x) f_\beta(y) dy \right) dx \\
&= \Pr(\beta \in A) \Pr(\alpha \in s \setminus A) + \Pr(\alpha \in A) \Pr(\beta \in A) \\
&= \Pr(\beta \in A).
\end{aligned}$$

■

Lemma 6 *Let $x_1 \propto \exp(\eta x - x^{d+1})$ be a random variable supported on $[-B, B]$ for some $d \geq 1$, $\eta \geq 4$ and some $B \geq 1$. Then,*

$$\mathrm{Var}(\mathrm{x}_1) \geq \frac{1}{4\mathrm{e}^2\eta^2}.$$

Proof The general approach for this proof will be as follows. We will first lower bound the variance as a function of the probability x_1 falls within a certain region. We will then find a chain of variables for which this probability is decreasing, until we reach a variable for which we can easily calculate this probability.

Assume WLOG that $\eta > 0$. We first observe that $\frac{d}{dx}\exp(\eta x - x^{d+1}) = 0 \implies x = \left(\frac{\eta}{d+1}\right)^d$, so the variable x_1 has a mode of $\left(\frac{\eta}{d+1}\right)^d$. Let $D = \min\left(B, \left(\frac{\eta}{d+1}\right)^d\right)$. We observe that $D \geq 3/\eta$, since $B = 1 > 3/4 \geq 3/\eta$ and $\left(\frac{\eta}{d+1}\right)^d \geq \left(\frac{4}{d+1}\right)^d \geq 0.9 > 3/4 \geq 3/\eta$ for any $d \geq 1$.

From the definition of variance, we know that $\mathrm{Var}(\mathrm{x}_1) \geq \frac{1}{\eta^2}\Pr(|\mathrm{x}_1 - \mathbb{E}[\mathrm{x}_1]| \geq \frac{1}{\eta})$. Since the pdf of x_1 is strictly increasing in the range $[-D, D]$, we can conclude that $\Pr(|x_1 - \mathbb{E}[x_1]| \geq \frac{1}{\eta}) \geq \Pr(|x_1| \leq D - 2/\eta)$. This leads us to our first inequality,

$$\mathrm{Var}(\mathrm{x}_1) \geq \frac{1}{\eta^2}\Pr(|\mathrm{x}_1| \leq \mathrm{D} - 2/\eta). \tag{15}$$

We now introduce a new variable $x_2 \propto \exp(\eta x - x^{d+1})$ with $x_2 \in [-\infty, B]$. Since x_2 will have a larger normalizing coefficient that x_1, we clearly have that

$$\Pr(|x_1| \leq D - 2/\eta) \geq \Pr(|x_2| \leq D - 2/\eta). \tag{16}$$

We now need to look at the shape of the distribution for x_2. Let $f_{x_2}(x) = \frac{\mathbf{1}(x<B)}{Z_2}\exp(\eta x - x^{d+1})$ denote a pdf function for x_2, where Z_2 is an appropriate normalizing coefficient. For any fixed c, we have that

$$\begin{aligned}
f_{x_2}(D+c) &= \frac{\mathbf{1}(D+c<B)}{Z_2}\exp(\eta(D+c) - (D+c)^{d+1}) \\
&\leq \frac{1}{Z_2}\exp\left(\eta D + \eta c - \sum_{j=0}^{d+1} D^j c^{d+1-j}\right) \\
&= \frac{1}{Z_2}\exp\left(\eta D - D^{d+1} + \eta c - (d+1)D^d c - \sum_{j=2}^{d+1} D^j c^{d+1-j}\right) \\
&= \frac{1}{Z_2}\exp\left(\eta D - D^{d+1} - \sum_{j=2}^{d+1} D^j c^{d+1-j}\right) \\
&\leq \frac{1}{Z_2}\exp\left(\eta D - D^{d+1} - \sum_{j=2}^{d+1} (-1)^j D^j c^{d+1-j}\right) \\
&= \frac{1}{Z_2}\exp\left(\eta D - D^{d+1} - \eta c + (d+1)D^d c - \sum_{j=2}^{d+1} (-1)^j D^j c^{d+1-j}\right)
\end{aligned}$$

$$= \frac{1}{Z_2} \exp\left(\eta(D-c) - (D-c)^{d+1}\right)$$
$$= f_{x_2}(D-c).$$

Thus, $\Pr(x_2 > D) \leq \Pr(x_2 < D)$. This inspires the creation of a third variable $x_3 \propto \exp(\eta x_3 - x_3^{d+1})$ with $x_3 \in [-\infty, D]$. Since $\Pr(x_2 > D) \leq \Pr(x_2 < D)$, we will have that $Z_3 \leq 2Z_2$, where Z_3 is the normalizing coefficient for x_3. Thus, we have that

$$\Pr(|x_2| \leq D - 2/\eta) \geq \frac{1}{2}\Pr(|x_3| \leq D - 2/\eta). \tag{17}$$

We are now ready to define our final variable $x_4 \propto \exp(\eta x_4)$ with $x_4 \in [-\infty, D]$. We first compute the normalizing constant $Z_4 = \int_{-\infty}^{D} \exp(\eta x)dx = \exp(\eta D)/\eta$. Next, we simply compute

$$\Pr(|x_4| \leq D - 2/\eta) = \frac{\eta}{\exp(\eta D)} \int_{-D+2/\eta}^{D-2/\eta} \exp(\eta x)dx = \frac{1}{\exp(\eta D)}(e^{D\eta-2} - e^{-D\eta+2}) \geq \frac{1}{2e^2}, \tag{18}$$

where the last inequality comes from the fact that $D \geq 3/\eta$. We let $f_{x_3} = \frac{1}{Z_3}\exp(\eta x_3 - x_3^{d+1})$ and $f_{x_4} = \frac{1}{Z_4}\exp(\eta x_4)$ be pdf functions for x_3 and x_4 respectively. Applying Lemma 5 with $A = [-D + 2/\eta, D - 2/\eta]$ to these pdfs, we can conclude that

$$\Pr(|x_3| \leq D - 2/\eta) \geq \Pr(|x_4| \leq D - 2/\eta). \tag{19}$$

Combining Equations (15) (16), (17), (18), and (19) together, we get our desired result. ■

Appendix B. Proof of Fact 1

Proof Let $v \in \mathbb{R}^n$ be some vector satisfying $\|v\|_\infty = 1$, and let x_v be the probability distribution for x conditioned on x lying in the span of v. If $\|x_v\|_\infty$ satisfies the tailbound for any choice of vector v, then $\|x\|_\infty$ will as well.

Defining $x_v = \alpha v$, we see that $\|x_v\|_\infty = |\alpha|$. Furthermore, we see that α has a probability density function of

$$p_\alpha(\alpha) = \frac{1}{Z_\alpha}\exp(-\sum_{i\in[n]}(\alpha v_i)^d + \sum_{k\in\mathcal{K}}\theta_k^* f_k(\alpha v)),$$

for appropriate value of Z_α. We define a new variable β which is sampled according to a distribution with pdf

$$p_\beta(\beta) = \frac{1}{Z_\beta}\exp(-\beta^d + nB\beta^{d-1}),$$

for appropriate value of Z_β. For any $1 \leq s_2 < s_1$, we have that

$$\frac{p_\alpha(s_1)}{p_\alpha(s_2)} = \exp(\sum_{k\in\mathcal{K}}\theta_k^*(f_k(s_1 v) - f_k(s_2 v)) - \sum_{i\in[n]}((s_1 v_i)^d - (s_2 v_i)^d))$$
$$\geq \exp(\sum_{k\in\mathcal{K}}\theta_k^*(s_1^{d-1} - s_2^{d-1}) - s_1^d + s_2^d)$$
$$\geq \exp(nB(s_1^{d-1} - s_2^{d-1}) - s_1^d + s_2^d)$$

$$= \frac{p_\beta(s_1)}{p_\beta(s_2)}.$$

Thus, by applying Lemma 5, we can conclude that for any $s > 1$, $\Pr(\alpha > s) \leq \Pr(\beta > s)$. Further, by taking the same argument applied to $-\alpha$, we can conclude that for any $s > 1$, $\Pr(|\alpha| > s) \leq 2\Pr(\beta > s)$. All that remains is to prove a tailbound for β. First, we see that

$$Z_\beta = \int_{-\infty}^{\infty} \exp(-\beta^d + nB\beta^{d-1})d\beta \geq \int_1^{nB} \exp(-\beta^d + nB\beta^{d-1})d\beta \geq nB.$$

Thus, for any $s > nB + 1$, we can see that

$$\Pr(\beta > s) = \frac{1}{Z_\beta}\int_s^{\infty} \exp(-\beta^d + nB\beta^{d-1})d\beta \leq \int_s^{\infty} \exp(-\beta^{d-1})d\beta \leq \exp(-s^{d-1}).$$

The last inequality holds from substituting $u = \beta^{d-1}$ and seeing $\frac{du}{d\beta} = (d-1)\beta^{d-2} \geq 1$.

■

Appendix C. Curvature Concentration Bounds

This section is dedicated to proving concentration bounds on the curvature of the loss function and formalizing the argument in Section 3. We start with the concentration result for $\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)})$.

Lemma 7 *Let $x^{(1)}, \ldots x^{(M)}$ be independent samples drawn from p_{θ^*}. Then, for any $\epsilon > 0$,*

$$\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) \geq \mathbb{E}[\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)})] - \frac{\epsilon B^2}{2} = \mathcal{L}_i(\Delta, x^*) - \frac{\epsilon B^2}{2},$$

holds for every $\Delta = \hat{\theta} - \theta^*$ corresponding to valid parameter vector $\hat{\theta}$ with probability greater than $1 - \frac{3d^4 C_t^{4d-4} n^{2d}}{M\epsilon^2}$.

Proof Let $H_{k_1,k_2} = \mathbb{E}_{p_{\theta^*}}[\frac{\partial}{\partial x_i} f_{k_1}(x) \frac{\partial}{\partial x_i} f_{k_2}(x)]$ and let $\hat{H}_{k_1,k_2} = \hat{\mathbb{E}}[\frac{\partial}{\partial x_i} f_{k_1}(x) \frac{\partial}{\partial x_i} f_{k_2}(x)]$, where $\hat{\mathbb{E}}$ denote the empirical average over our M samples. We will show that the entries of these two matrices vary only slightly with high probability, and from that result show that the curvature is close to its expectation with high probability.

Fix some $k_1, k_2 \in \mathcal{K}_i$, and notice that $\left|\frac{\partial}{\partial x_i} f_{k_1}(x) \frac{\partial}{\partial x_i} f_{k_2}(x)\right| \leq d^2\|x\|_\infty^{2d-2}$ by construction. Thus, $\mathrm{Var}\left(\mathrm{d}^2\|\mathrm{x}\|_\infty^{2\mathrm{d}-2}\right) \leq \mathrm{d}^4\mathbb{E}[\|\mathrm{x}\|_\infty^{4\mathrm{d}-4}]$. From Condition 1, we know for any $s \geq C_t^{4d-4}$ that

$$\Pr(\|x\|_\infty^{4d-4} \geq s) = \Pr(\|x\|_\infty \geq s^{1/(4d-4)}) \leq \exp(-k\sqrt{s}).$$

Thus,

$$\begin{aligned}\mathbb{E}[\|x\|_\infty^{4d-4}] &\leq C_t^{4d-4} + \int_{C_t^{4d-4}}^{\infty} \Pr(\|x\|_\infty^{4d-4} > s)ds \\ &\leq C_t^{4d-4} + \int_{C_t^{4d-4}}^{\infty} \exp(-k\sqrt{s})ds\end{aligned}$$

$$\leq C_t^{4d-4} + \frac{2}{k^2}\exp(-kC_t^{2d-2})(1 + kC_t^{2d-2})$$
$$\leq 3C_t^{4d-4}. \tag{20}$$

We apply Chebyshev's Inequality to see that

$$\Pr(|H_{k_1,k_2} - \hat{H}_{k_1,k_2}| \geq \epsilon) \leq \frac{\mathrm{Var}(\hat{\mathrm{H}}_{\mathrm{k}_1,\mathrm{k}_2})}{\epsilon^2} \leq \frac{3d^4C_t^{4d-4}}{M\epsilon^2}.$$

Notice that $|\mathcal{K}_i| \leq n^d$. Taking a union bound over all possible $k_1, k_2 \in \mathcal{K}_i$, we get that

$$\Pr(\|H - \hat{H}\|_\infty \geq \epsilon) \leq \frac{3d^4n^{2d}C_t^{4d-4}}{M\epsilon^2}.$$

Following the logic in equation 3, and conditioned on $\|H - \hat{H}\|_\infty \geq \epsilon$, we see that for every Δ

$$\begin{aligned}
\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) &= \frac{\Delta_i^T\hat{H}\Delta_i}{2} \\
&= \frac{\Delta_i^T H\Delta_i + \Delta_i^T(\hat{H} - H)\Delta_i}{2} \\
&= \delta\mathcal{L}_i(\Delta, x^*) + \frac{\Delta_i^T(\hat{H} - H)\Delta_i}{2} \\
&\geq \mathcal{L}_i(\Delta, x^*) - \frac{\epsilon\|\Delta_i\|_1^2}{2} \\
&\geq \mathcal{L}_i(\Delta, x^*) - \frac{\epsilon B^2}{2},
\end{aligned}$$

where Δ_i denote the entries in Δ corresponding to some $k \in \mathcal{K}_i$. The last inequality follows from $\Delta = \hat{\theta} - \theta^*$, where both satisfy a B $l1$-bound. This conclude the proof.

■

Next, we prove a probabilistic upper bound for $\|\nabla\mathcal{L}_i(\theta, x^{(M)})\|_\infty$.

Lemma 8 *Let $x^{(1)}, \ldots x^{(M)}$ be independent samples drawn from p_{θ^*}. Then, for any $\epsilon > 0$,* $\Pr(\|\nabla\mathcal{L}_i(\theta^*, x^{(M)})\|_\infty \geq \epsilon) \leq \frac{12d^2B^2C_t^{4d-4}n^d}{M\epsilon^2}$.

Proof For any collection of samples $x_1, \ldots x_M$, we see that

$$\|\nabla_\theta\mathcal{L}_i(\theta^*, x^{(M)})\|_\infty = \max_{k\in\mathcal{K}} \frac{1}{M}\sum_{j=1}^{M}\left(\frac{\partial}{\partial x_i^2}f_k(x^{(j)}) + 2\frac{\partial}{\partial x_i}f_k(x^{(j)})\sum_{k'\in\mathcal{K}_i}\theta_{k'}\frac{\partial}{\partial x_i}f_{k'}(x^{(j)})\right).$$

We let α_k denote the expression inside of the max operator above for each k. We know that $\mathbb{E}_{p_{\theta^*}}[\alpha_k] = 0$, since θ^* is the expected minimizer of $\mathcal{L}_i$. Further, notice that
$\alpha_k \leq \frac{1}{M}\sum_{j=1}^{M} 2dB\|x^{(j)}\|_\infty^{2d-2}$, implying that

$$\mathrm{Var}(\alpha_\mathrm{k}) = \mathbb{E}[\alpha_\mathrm{k}^2] \leq 4\mathrm{d}^2\mathrm{B}^2\mathbb{E}[\|\mathrm{x}\|_\infty^{4\mathrm{d}-4}]/\mathrm{M} \leq 12\mathrm{d}^2\mathrm{B}^2\mathrm{C}_\mathrm{t}^{4\mathrm{d}-4}/\mathrm{M}$$

by Equation (20). We can now apply Chebyshev's Inequality to see that

$$\Pr(\frac{1}{M}\sum_{j=1}^{M}\alpha_i \geq \epsilon) \leq \frac{\mathrm{Var}(\frac{1}{\mathrm{M}}\sum_{\mathrm{j=1}}^{\mathrm{M}}\alpha_{\mathrm{i}})}{\epsilon^2} \leq \frac{12d^2B^2C_t^{4d-4}}{M\epsilon^2}.$$

Observing that $|\mathcal{K}_i| \leq n^d$ and applying a union bound to the above will prove our result. ■

Finally, we combine the previous two Lemmas with Equation (5) in the following Lemma to complete the argument outlined in Section 3.

Lemma 9 *Take target factors $\hat{\mathcal{K}} \subseteq \mathcal{K}$ which satisfy $\mathbb{E}_{x\sim p_t}E_i(x,\Delta)^2 \geq C_p\Delta_k^2$ for every $k \in \hat{\mathcal{K}}$. Then, for any $\rho > 0$, the empirical minimizer $\hat{\theta}$ of $\mathcal{L}_i(\theta, x^{(M)})$ satisfies $(\theta_k^* - \hat{\theta}_k)^2 \leq \epsilon$ for every $k \in \hat{\mathcal{K}}$ with probability*

$$1 - \frac{1}{\rho n C_t},$$

as long as

$$M \geq \rho\frac{3120d^4B^4C_t^{4d}n^{2d+1}}{\epsilon^2C_p^2}.$$

Proof Recall that $\Delta = \hat{\theta} - \theta^*$ and let $m_\Delta = \max_{k\in\hat{\mathcal{K}}}\Delta_k^2$. Continuing from Equation (5), we observe that $\hat{\theta}$ being the empirical minimizer implies

$$\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) \leq 2B\|\nabla_\theta\mathcal{L}_i(\theta^*, x^{(M)})\|_\infty.$$

Combining Equations (3) and (4), we know that

$$\delta\mathcal{L}_i(\Delta, \theta^*) \geq \frac{1}{4}\mathbb{E}_{x\sim p_t}E_i[x,\Delta]^2 \geq \frac{1}{4}C_p m_\Delta.$$

By plugging $\epsilon' = \frac{C_p\epsilon}{4B^2}$ into Lemma 7, we see that if $m_\Delta > \epsilon$, then

$$\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) \geq \mathcal{L}_i(\Delta, x^*) - \frac{\epsilon' B^2}{2} \geq \frac{1}{4}C_p m_\Delta - \frac{1}{8}C_p\epsilon > \frac{1}{8}C_p\epsilon, \tag{21}$$

with probability at least $1 - \frac{48d^4B^4C_t^{4d-4}n^{2d}}{MC_p^2\epsilon^2}$. Further, by Lemma 8, we know that

$$\|\nabla_\theta\mathcal{L}_i(\theta^*, x^{(M)})\|_\infty \leq \frac{C_p\epsilon}{16B}, \tag{22}$$

with probability $1 - \frac{3072d^2B^4C_t^{4d-4}n^d}{MC_p^2\epsilon^2}$. Since Equations (21) and (22) cannot both hold without contradicting $\hat{\theta}$ being the empirical minimizer, we can conclude via the union bound that

$$\begin{aligned}\Pr(m_\Delta > \epsilon) &\leq \frac{48d^4B^4C_t^{4d-4}n^{2d}}{MC_p^2\epsilon^2} + \frac{3072d^2B^4C_t^{4d-4}n^d}{MC_p^2\epsilon^2} \\ &\leq \frac{3120d^4B^4C_t^{4d-4}n^{2d}}{MC_p^2\epsilon^2} \\ &\leq \frac{1}{\rho n C_t}\end{aligned}$$

■

Appendix D. Near-Optimality Recovery Guarantee

In practice, it may be difficult to compute the exact minimizer of the score-matching loss. However, our analysis of the loss's curvature allows us to prove recovery guarantees for *approximate* minimizers of the score-matching loss as well. Specifically, given samples $x^{(M)}$, we say a solution $\tilde{\theta}$ is a η-minimizer of $\mathcal{L}_i(\theta, x^{(M)})$ if for some $\eta > 0$,

$$\mathcal{L}_i(\tilde{\theta}, x^{(M)}) - \mathcal{L}_i(\hat{\theta}, x^{(M)}) \leq \eta,$$

where $\hat{\theta}$ is the true minimizer. This section will be dedicated to proving the following result, which extends our main result Theorem 1 for a η-minimizer $\tilde{\theta}$. A similar process can be followed for our other theorems as well.

Theorem 10 (Learning from Near-Optimality) *Fix some exponential family p_{θ^*} with base measure $h(x) = 1$. For a fixed index $i \in [n]$, let $\hat{\mathcal{K}}$ be the maximal factors with i as a neighbor in the family factor graph G, meaning $\hat{\mathcal{K}} = \{k \in \mathcal{M}_{\text{fac}}(G) | i \in \partial k\}$. Further, for M independent samples $x_1, \dots x_M \sim p_{\theta^*}$, let $\tilde{\theta}$ be a η-minimizer of $\mathcal{L}_i(\theta, x^{(M)})$ subject to $\sum_{k \in \mathcal{K}_j} |\theta_k| \leq B$ for every index j.*

There exists some M^* and η^* both of the order $(dBC_t^d)^{O(d^2 w)}$ such that for every $\rho \geq 1$ and $\epsilon \leq 1$, we have that $(\theta_k^* - \hat{\theta}_k)^2 \leq \epsilon$ for every $k \in \hat{\mathcal{K}}$ with probability greater than $\frac{1}{\rho n C_t}$ when $M \geq \rho \frac{n^{d+1} M^*}{\epsilon^2}$ and $\eta \leq \epsilon \eta^*$.

To proceed, we will first need to show the following adapted version of Lemma 9. With this, Theorem 10 follows directly from applying the Lemma to Equation (12), just as the main Theorem does.

Lemma 11 *Take target factors $\hat{\mathcal{K}} \subseteq \mathcal{K}$ which satisfy $\mathbb{E}_{x \sim p_t} E_i(x, \Delta)^2 \geq C_p \Delta_k^2$ for every $k \in \hat{\mathcal{K}}$ and fix some $\epsilon > 0$. Let $\tilde{\theta}$ be an η-minimizer of $\mathcal{L}_i(\theta, x^{(M)})$, for some $\eta < \frac{1}{8} C_p \epsilon$. Then, for any $\rho > 0$, $\tilde{\theta}$ satisfies $(\theta_k^* - \tilde{\theta}_k)^2 \leq \epsilon$ for every $k \in \hat{\mathcal{K}}$ with probability*

$$1 - \frac{1}{\rho n C_t},$$

as long as

$$M \geq \rho \frac{3264 d^4 B^4 C_t^{4d} n^{2d+1}}{\epsilon^2 C_p^2}.$$

Proof Set $\Delta = \tilde{\theta} - \theta^*$ and let $m_\Delta = \max_{k \in \hat{\mathcal{K}}} \Delta_k^2$. We first observe that $\delta \mathcal{L}_i(\Delta, \tilde{\theta}, x^{(M)}) - \delta \mathcal{L}_i(\Delta, \theta^*, x^{(M)}) \leq \eta$ by the definition of η-minimizer. Adapting Equation (5), we observe that

$$\delta \mathcal{L}_i(\Delta, \theta^*, x^{(M)}) \leq 2B \|\nabla_\theta \mathcal{L}_i(\theta^*, x^{(M)})\|_\infty + \eta < 2B \|\nabla_\theta \mathcal{L}_i(\theta^*, x^{(M)})\|_\infty + \frac{1}{8} C_p \epsilon.$$

Combining Equations (3) and (4), we know that

$$\delta \mathcal{L}_i(\Delta, \theta^*) \geq \frac{1}{4} \mathbb{E}_{x \sim p_t} E_i[x, \Delta]^2 \geq \frac{1}{4} C_p m_\Delta.$$

By plugging $\epsilon' = \frac{C_p \epsilon}{8B^2}$ into Lemma 7, we see that if $m_\Delta > \epsilon$, then

$$\delta\mathcal{L}_i(\Delta, \theta^*, x^{(M)}) \geq \mathcal{L}_i(\Delta, x^*) - \frac{\epsilon' B^2}{2} \geq \frac{1}{4} C_p m_\Delta - \frac{1}{16} C_p \epsilon > \frac{3}{16} C_p \epsilon, \tag{23}$$

with probability at least $1 - \frac{192 d^4 B^4 C_t^{4d-4} n^{2d}}{M C_p^2 \epsilon^2}$. Further, by Lemma 8, we know that

$$\|\nabla_\theta \mathcal{L}_i(\theta^*, x^{(M)})\|_\infty \leq \frac{C_p \epsilon}{16 B}, \tag{24}$$

with probability $1 - \frac{3072 d^2 B^4 C_t^{4d-4} n^d}{M C_p^2 \epsilon^2}$. Since Equations (23) and (24) cannot both hold without contradicting $\tilde{\theta}$ being an η-minimizer, we can conclude via the union bound that

$$\begin{aligned}
\Pr(m_\Delta > \epsilon) &\leq \frac{192 d^4 B^4 C_t^{4d-4} n^{2d}}{M C_p^2 \epsilon^2} + \frac{3072 d^2 B^4 C_t^{4d-4} n^d}{M C_p^2 \epsilon^2} \\
&\leq \frac{3264 d^4 B^4 C_t^{4d-4} n^{2d}}{M C_p^2 \epsilon^2} \\
&\leq \frac{1}{\rho n C_t}
\end{aligned}$$

■