# Learning Multi-Index Models with Neural Networks via Mean-Field Langevin Dynamics

**Alireza Mousavi-Hosseini** MOUSAVI@CS.TORONTO.EDU
*University of Toronto and Vector Institute*

**Denny Wu** DENNYWU@NYU.EDU
*New York University and Flatiron Institute*

**Murat A. Erdogdu** ERDOGDU@CS.TORONTO.EDU
*University of Toronto and Vector Institute*

## Abstract

We study the problem of learning multi-index models in high-dimensions using a two-layer neural network trained with the mean-field Langevin algorithm. Under mild distributional assumptions on the data, we characterize the *effective dimension* $d_{\text{eff}}$ that controls both sample and computational complexity by utilizing the adaptivity of neural networks to latent low-dimensional structures. When the data exhibit such a structure, $d_{\text{eff}}$ can be significantly smaller than the ambient dimension. We prove that the sample complexity grows linearly with $d_{\text{eff}}$, bypassing the limitations of the information exponent or the leap complexity that appeared in recent analyses of gradient-based feature learning. On the other hand, the computational complexity may inevitably grow exponentially with $d_{\text{eff}}$ in the worst-case scenario.

## 1. Introduction

A key characteristic of neural networks is their adaptability to the underlying statistical model. Several works have shown that shallow neural networks trained by (variants of) gradient descent can efficiently learn functions of low-dimensional projections (i.e., multi-index models) with a sample complexity that depends on the properties of the nonlinear link function known as the *information exponent* [9] or the *leap complexity* [2]. Specifically, to learn a target function with information exponent or leap complexity $k \in \mathbb{N}_+$ on isotropic Gaussian data, a sample size of $n \gtrsim d^{\Theta(k)}$ is typically required in these analyses [1, 11, 12, 21, 35]. This sample complexity is also predicted by the framework of correlational statistical query (CSQ) lower bounds [2, 23].

On the other hand, if the (polynomial) optimization budget is not taken into consideration, [7] showed that neural networks can learn multi-index models with a sample complexity that does not depend on the information or leap exponent. However, thus far it has been relatively unclear whether standard first-order optimization algorithms inherit this optimality.

A promising approach to obtain statistically optimal sample complexity is to consider training neural networks in the *mean-field regime* [17, 34, 37, 41, 43], where overparameterization lifts the gradient descent dynamics into the space of measures with global convergence guarantees. While most existing results in this regime focus on optimization instead of generalization/learnability guarantees, recent works have shown that in certain settings, neural networks in the mean-field regime can achieve a sample complexity that does not depend on the leap complexity [18, 45, 47, 51]. However, these guarantees rely on stringent assumptions on data (isotropic Gaussian, hypercube, etc.) as well as single-index models with specific link functions [10, 33], or $k$-sparse partiy classifi-

cation [45, 47, 51]. Moreover, the computational complexity of the training algorithm in [45, 47] is exponential in the ambient (input) dimension, without adapting to the potential structure in the inputs. An exception in this direction is the recent work of [39], which considers the $k$-parity problem on anisotropic data.

**Our contributions.** Motivated by the above discussion, we address two key questions. First,

*Can we train two-layer neural networks using the MFLA to learn arbitrary multi-index models with an optimal sample complexity?*

We answer this in the affirmative by showing that empirical risk minimization on a standard variant of a two-layer neural network can be achieved by the MFLA. This result handles arbitrary multi-index models on subGaussian data with general covariance, hence enabling us to achieve the optimal sample complexity with standard gradient-based training. However, such a universal guarantee will inevitably suffer from an exponential computational complexity; thus, the second fundamental question we aim to answer is

*Are there conditions under which the computational complexity of the MFLA can be improved by adapting to data structure?*

We provide a positive answer by showing that the computational complexity of MFLA is governed by the *effective dimension* of the learning problem, instead of its ambient dimension; this implies an improved efficiency of MFLA when the data is anisotropic as in prior works [28, 36].

**Related works.** The training dynamics of neural networks in the mean-field regime is described by a nonlinear partial differential equation in the space of parameter distributions [17, 34, 41]. Unlike the NTK description [19, 30] that freezes the parameters around the random initialization, the mean-field regime allows for the parameters to travel and learn useful features, leading to improved statistical efficiency. While convergence analyses for mean-field neural networks are typically *qualitative* in nature, in that they do not specify the rate of convergence or finite-width discrepancy, the mean-field Langevin algorithm that we study is a noticeable exception, for which the convergence rate [16, 29, 38] as well as uniform-in-time propagation of chaos [13, 44] have been established.

The benefit of feature learning has also been studied in a "narrow-width" setting for learning low-dimensional target functions. Examples of low-dimensional targets include single-index models [5, 9, 11, 21, 35] and multi-index models [1, 2, 12, 20, 23, 24]. While the information-theoretic threshold for learning such functions is $n \gtrsim d$ [8, 22], the complexity of gradient-based learning is governed by properties of the link function. For instance, in the single-index setting, prior works established a sufficient sample size of $n \gtrsim d^{\Theta(k)}$ where $k$ is the *information exponent* [9, 11, 21] or the *generative exponent* [4, 22, 25, 32]. This presents a gap between the information-theoretically achievable sample complexity and the performance of neural networks optimized by gradient descent, which we aim to close by studying the statistical efficiency of mean-field neural networks.

## 2. Preliminaries: Statistical Model and Training Algorithm

**Statistical model.** In this paper, we consider the regression setting where the input $\boldsymbol{x} \in \mathbb{R}^d$ is generated from some distribution and the response $y \in \mathbb{R}$ is given by the multi-index model

$$y = g\Big(\frac{\langle \boldsymbol{u}_1, \boldsymbol{x}\rangle}{\sqrt{k}}, \dots, \frac{\langle \boldsymbol{u}_k, \boldsymbol{x}\rangle}{\sqrt{k}}\Big) + \xi. \tag{2.1}$$

Here, $g : \mathbb{R}^k \to \mathbb{R}$ is the unknown link function, $\xi$ is a zero-mean $\varsigma$-subGaussian noise independent from $\boldsymbol{x}$; for simplicity, we assume that $\varsigma^2 \lesssim 1$. Without loss of generality, we assume $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$

are orthonormal and form the matrix $\boldsymbol{U} = (\boldsymbol{u}_1/\sqrt{k}, \dots, \boldsymbol{u}_k/\sqrt{k})^\top \in \mathbb{R}^{k\times d}$; thus, we can use the shorthand notation $y = g(\boldsymbol{U}\boldsymbol{x}) + \xi$. Throughout the paper, we consider the setting $k \ll d$, and treat $k$ as an absolute constant independent from the ambient input dimension $d$.

For our predictor, we use a two-layer neural network coupled with $\ell_2$ regularization to learn the statistical model (2.1). Denoting the $m$ neurons with a matrix $\boldsymbol{W} \coloneqq (\boldsymbol{w}_1, \dots, \boldsymbol{w}_m)^\top$, the student model and the $\ell_2$-regularizer are given as

$$\hat{y}_m(\boldsymbol{x}; \boldsymbol{W}) \coloneqq \frac{1}{m}\sum_{j=1}^{m} \Psi(\boldsymbol{x}; \boldsymbol{w}_j) \quad \text{and} \quad R(\boldsymbol{W}) \coloneqq \frac{1}{m}\|\boldsymbol{W}\|_{\mathrm{F}}^2 = \frac{1}{m}\sum_{j=1}^{m}\|\boldsymbol{w}_j\|^2, \tag{2.2}$$

where $\Psi : \mathbb{R}^d \times \mathcal{W} \to \mathbb{R}$ is the activation function, and $\boldsymbol{w}_j \in \mathcal{W}$ with $\mathcal{W}$ denoting the weight space. In this formulation, the second layer weights are all fixed to be $+1$. In the classical regression setting where we observe $n$ i.i.d. samples $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^n$ from the data distribution, the regularized population and empirical risks are defined respectively as

$$J_\lambda(\boldsymbol{W}) \coloneqq \mathbb{E}[\ell(\hat{y}_m(\boldsymbol{x}; \boldsymbol{W}), y)] + \frac{\lambda}{2}R(\boldsymbol{W}) \;\text{ and }\; \hat{J}_\lambda(\boldsymbol{W}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_m(\boldsymbol{x}^{(i)}; \boldsymbol{W}), y^{(i)}) + \frac{\lambda}{2}R(\boldsymbol{W}),$$

where $\ell(\hat{y}, y) = \rho(\hat{y} - y)$ with $\rho : \mathbb{R} \to \mathbb{R}_+$ being a convex loss.

**Training Algorithm.** We minimize the regularized empirical risk $\hat{J}_\lambda(\boldsymbol{W})$ via the *mean-field Langevin algorithm* (MFLA) with stepsize $\eta$, which updates the weights at iteration $l$ by

$$\boldsymbol{w}_j^{l+1} = \boldsymbol{w}_j^l - m\eta\nabla_{\boldsymbol{w}_j}\hat{J}_\lambda(\boldsymbol{W}) + \sqrt{\frac{2\eta}{\beta}}\boldsymbol{\xi}_j^l, \quad 1 \le j \le m, \tag{2.3}$$

where $\boldsymbol{\xi}_j^l$ are independent standard Gaussian random vectors. In Appendix A, we will introduce the necessary background on the mean-field Langevin dynamics (MFLD) and optimization on the space of measures, where we observe that (2.3) is a simple time-discretization of the MFLD.

## 3. Learning Multi-index Models

In this section, we will provide the learning guarantees. For technical reasons, we use an approximation of ReLU denoted by $z \mapsto \phi_{\kappa,\iota}(z)$ for some $\kappa, \iota > 1$, which is given by $\phi_{\kappa,\iota}(z) = \kappa^{-1}\ln(1 + \exp(\kappa z))$ for $z \in (-\infty, \iota/2]$ and extended on $(\iota, \infty)$ such that $\phi_{\kappa,\iota}$ is $C^2$ smooth, $|\phi_{\kappa,\iota}| \le \iota$, $|\phi'_{\kappa,\iota}| \le 1$, and $|\phi''_{\kappa,\iota}| \le \kappa$. Note that $\phi_{\kappa,\iota}$ recovers ReLU as $\kappa, \iota \to \infty$. To be able to learn functions with positive and negative parts, we choose $\mathcal{W} = \mathbb{R}^{2d+2}$, and use the notation $\boldsymbol{w} = (\boldsymbol{\omega}_1^\top, \boldsymbol{\omega}_2^\top)^\top$ with $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2 \in \mathbb{R}^{d+1}$, and ultimately use

$$\Psi(\boldsymbol{x}; \boldsymbol{w}) \coloneqq \phi_\kappa(\langle\tilde{\boldsymbol{x}}, \boldsymbol{\omega}_1\rangle) - \phi_\kappa(\langle\tilde{\boldsymbol{x}}, \boldsymbol{\omega}_2\rangle), \tag{3.1}$$

where $\tilde{\boldsymbol{x}} \coloneqq (\boldsymbol{x}, \tilde{r}_x)^\top \in \mathbb{R}^{d+1}$ for a constant $\tilde{r}_x$ corresponding to bias, to be specified later. The above can also be viewed as a 2-layer neural network with second-layer weights frozen at $\pm 1$. We make the following assumption on the input distribution.

**Assumption 1** *The input $\boldsymbol{x}$ has zero mean and covariance $\boldsymbol{\Sigma}$. Further, $\|\boldsymbol{x}\|$ and $\|\boldsymbol{U}\boldsymbol{x}\|$ are sub-Gaussian with respective norms $\sigma_n\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^{1/2}$ and $\sigma_u\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}^\top\|_{\mathrm{F}}$ for some absolute constants $\sigma_n, \sigma_u$.*

Even though the above assumption covers a wide range of input distributions, it is mainly motivated by the Gaussian case which satisfies the above assumption. Without loss of generality, we will consider a scaling where $\|\mathbf{\Sigma}\| \lesssim 1$. A key quantity in our analysis is the *effective dimension* which governs the algorithmic guarantees.

**Definition 1 (Effective dimension)** *Define* $d_{\text{eff}} \coloneqq c_x^2/r_x^2$ *where* $c_x \coloneqq \text{tr}(\mathbf{\Sigma})^{1/2}$, $r_x \coloneqq \|\mathbf{\Sigma}^{1/2}\boldsymbol{U}^\top\|_{\text{F}}$.

The effective dimension $d_{\text{eff}}$ can be significantly smaller than the ambient dimension $d$, leading to particularly favorable results in the following when $d_{\text{eff}} = \text{polylog}(d)$. This concept has numerous applications from learning theory to statistical estimation; see e.g. [6, 27, 48, 50]. We make the following assumption on the link function in (2.1).

**Assumption 2** *$g$ is locally Lipschitz, i.e. $|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)| \le L\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$ for $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^k$ satisfying $\|\boldsymbol{z}_1\| \vee \|\boldsymbol{z}_2\| \le \tilde{r}_x \coloneqq r_x(1 + \sigma_u\sqrt{2\ln(n)^q})$ for some $q > 0$ and $L = \mathcal{O}(1/r_x)$.*

We emphasize that the above Lipshitz condition is only local; allowing e.g. polynomially growing link functions. We scale the Lipschitz constant with $1/r_x$ to make sure $|y| \asymp 1$ with high probability. The main result of this section is stated in the following theorem.

**Theorem 2** *For an appropriate choice of hyperparameters $\eta, \kappa, \iota, \lambda$, and $\beta$, with a sufficient number of samples, number of neurons, and number of iterations that can be bounded by*

$$n \le \tilde{\mathcal{O}}(d_{\text{eff}}), \quad m \le \tilde{\mathcal{O}}(d^2 e^{\tilde{\mathcal{O}}(d_{\text{eff}})}), \quad l \le \tilde{\mathcal{O}}(d^3 e^{\tilde{\mathcal{O}}(d_{\text{eff}})}), \tag{3.2}$$

*with probability at least $1 - \mathcal{O}(n^{-q})$ for some $q > 0$, MFLA can achieve the excess risk bound*

$$\mathbb{E}_{\boldsymbol{W}^l}\mathbb{E}_{y,\boldsymbol{x}}[\rho(y - \hat{y}_m(\boldsymbol{x}; \boldsymbol{W}^l))] - \mathbb{E}_\xi[\rho(\xi)] \le o_n(1). \tag{3.3}$$

We refer to Theorem 28 in Appendix B for a more precise statement with the choice of hyperparameters. The theorem above demonstrates a certain adaptivity to the effective low-dimensional structure, both in terms of *statistical* and *computational* complexity, which occurs without explicitly encoding any information about the covariance structure in the algorithm. In contrast, "fixed-grid" methods (see [15] and references therein), that fix the first-layer of a two-layer network's representation similar to random features regression [40], and then train the second-layer by solving a convex problem, do not show this type of adaptivity to low dimensions. In particular their computational complexity always scales exponentially with the ambient dimension $d$, unless information about the covariance structure is explicitly used when specifying the fixed representation.

**Comparison with prior bounds.** Here, we compare the guarantee of Theorem 2 with two prior works that are particularly relevent. First, [7] requires $d^{\frac{k+3}{2}}$ sample complexity for learning general multi-index models with $k$ indices, which is worse than the complexity $d_{\text{eff}}$ of Theorem 2 even in the worst case $d_{\text{eff}} = d$. The improvement in our bound is due to a refined control over $\|\boldsymbol{U}\boldsymbol{x}\|$. Further, [7] does not provide a quantitative analysis of the optimization complexity, and it is not clear if their algorithm is adaptive to the covariance structure. Moreover, [39] studies learning $k$-sparse parities, a subclass of multi-index models we considered, for which it is considerably simpler to construct optimal neural networks with a bounded activation. While the effective dimension (and the resulting sample complexity) of [39] is not explicitly scale-invariant, we derive a scale-invariant translation of their bound in Appendix C, and show that it is always lower bounded by our effective dimension, especially when $\mathbf{\Sigma}$ is nearly rank-deficient.

## 4. Interpreting the Effective Dimension

To better demonstrate the impact of effective dimension, we consider two covariance models.

**Spiked covariance.** We consider the spiked covariance model of [36]. Namely, given a spike direction $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, suppose the covariance and the target directions satisfy

$$\boldsymbol{\Sigma} = \frac{\mathbf{I}_d + \alpha \boldsymbol{\theta}\boldsymbol{\theta}^\top}{1+\alpha}, \quad \alpha \asymp d^{\gamma_2}, \quad \|\boldsymbol{U}\boldsymbol{\theta}\| \asymp d^{-\gamma_1}, \quad \gamma_2 \in [0,1], \quad \gamma_1 \in [0, 1/2]. \tag{4.1}$$

Note that in high-dimensional settings, $\gamma_1 = 1/2$ corresponds to a regime where $\boldsymbol{\theta}$ is sampled uniformly over $\mathbb{S}^{d-1}$, whereas $\gamma_1 = 0$ corresponds to the case where $\boldsymbol{\theta}$ has a strong correlation with $\boldsymbol{U}$. We only consider $\gamma_2 \leq 1$ since $\gamma_2 > 1$ corresponds to a setting where the input is effectively one-dimensional. In this setting, effective dimension depends on $\gamma_1$ and $\gamma_2$.

**Corollary 3** *Under the spiked covariance model (4.1), we have* $d_{\text{eff}} \asymp d^{1-\{(\gamma_2-2\gamma_1)\vee 0\}}$.

To get improvements over the isotropic effective dimension $d$, either the spike magnitude $\alpha$ or the spike-target alignment $\|\boldsymbol{U}\boldsymbol{\theta}\|$ needs to be sufficiently large so that $\gamma_2 > 2\gamma_1$. As $\gamma_2 \to 1$ and $\gamma_1 \to 0$, the effective dimension will be smaller than $\mathrm{polylog}(d)$, leading to a computational complexity that is quasipolynomial in $d$.

**Covariance with decaying eigenvalues.** Next, we consider a more general power-law decay for the eigenspectrum. Suppose $\boldsymbol{\Sigma} = \sum_{i=1}^d \lambda_i \boldsymbol{\theta}_i \boldsymbol{\theta}_i^\top$ is the spectral decomposition of $\boldsymbol{\Sigma}$, and

$$\frac{\lambda_i}{\lambda_1} \asymp i^{-\alpha}, \quad \frac{\|\boldsymbol{U}\boldsymbol{\theta}_i\|^2}{\|\boldsymbol{U}\boldsymbol{\theta}_1\|^2} \asymp i^{-\gamma}, \quad \text{for} \quad 1 \leq i \leq d, \tag{4.2}$$

for some absolute constants $\alpha, \gamma > 0$. Notice that $\sum_{i=1}^d \|\boldsymbol{U}\boldsymbol{\theta}_i\|^2 = \|\boldsymbol{U}\|_{\text{F}}^2 = 1$. The following corollary characterizes $d_{\text{eff}}$ in terms of the parameters $\alpha$ and $\gamma$.

**Corollary 4** *Under the power-law eigenspectrum for the covariance matrix (4.2), we have*

$$d_{\text{eff}} \asymp \begin{cases} d^{1\wedge(2-\alpha-\gamma)} & \alpha < 1, \gamma < 1 \\ d^{1-\alpha} & \alpha < 1, \gamma \geq 1 \\ d^{(1-\gamma)\vee 0} & \alpha \geq 1 \end{cases},$$

*where $\asymp$ above hides* $\mathrm{polylog}(d)$ *dependencies.*

Thus, the computational complexity becomes quasipolynomial in $d$ when $\alpha, \gamma \geq 1$. This happens when $\boldsymbol{\Sigma}$ is approximately low-rank with most of its eigenspectrum concentrated around the first few eigenvalues, and the corresponding eigenvectors are well-aligned with the row space of $\boldsymbol{U}$.

## 5. Conclusion

In this paper, we investigated the mean-field Langevin algorithm for learning multi-index models. We proved that the statistical and computational complexity of this problem can be characterized by an effective dimension which captures the low-dimensional structure in the input covariance, along with its correlation with the target directions. In particular, the sample complexity scales almost linearly with the effective dimension, while without additional assumptions, the computational complexity may scale exponentially with this quantity. We leave open the question of under which assumptions the computaitonal complexity will be polynomial even in the effective dimension as an interesting direction for future research.

## References


[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, 2022.

[2] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[4] Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

[5] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. *arXiv preprint arXiv:2205.01445*, 2022.

[6] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2023.

[7] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[8] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[9] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.*, 22:106–1, 2021.

[10] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

[11] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[12] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[13] Fan Chen, Zhenjie Ren, and Songbo Wang. Uniform-in-time propagation of chaos for mean field langevin dynamics. *arXiv preprint arXiv:2212.03050*, 2022.

[14] Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33:13363–13373, 2020.

[15] Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *Open Journal of Mathematical Optimization*, 3:1–19, 2022.

[16] Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

[17] Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems*, 2018.

[18] Lénaïc Chizat and Francis Bach. Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss. In *Conference on Learning Theory*, 2020.

[19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, 2019.

[20] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *arXiv preprint arXiv:2308.08977*, 2023.

[21] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. The computational complexity of learning gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

[23] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural Networks can Learn Representations with Gradient Descent. In *Conference on Learning Theory*, 2022.

[24] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[25] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

[26] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36 (2):183–212, 1983.

[27] B. Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of Lazy Training of Two-layers Neural Networks. In *Advances in Neural Information Processing Systems*, 2019.

[28] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When Do Neural Networks Outperform Kernel Methods? In *Advances in Neural Information Processing Systems*, 2020.

[29] Kaitong Hu, Zhenjie Ren, David Siska, and Lukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769*, 2019.

[30] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2018.

[31] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[32] Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

[33] Arvind Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. *Advances in Neural Information Processing Systems*, 36, 2023.

[34] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

[35] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36, 2023.

[37] Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*, 2017.

[38] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, 2022.

[39] Atsushi Nitanda, Kazusato Oko, Taiji Suzuki, and Denny Wu. Improved statistical and computational complexity of the mean-field langevin dynamics under structured data. In *The Twelfth International Conference on Learning Representations*, 2024.

[40] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

[41] Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as Interacting Particle Systems: Asymptotic convexity of the Loss Landscape and Universal Scaling of the Approximation Error. *arXiv preprint arXiv:1805.00915*, 2018.

[42] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[43] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[44] Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: Time and space discretization, stochastic gradient, and variance reduction. *arXiv preprint arXiv:2306.07221*, 2023.

[45] Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: classifying sparse parities and beyond. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[46] Alain-Sol Sznitman. Topics in propagation of chaos. *Lecture notes in mathematics*, pages 165–251, 1991.

[47] Matus Telgarsky. Feature selection and low test error in shallow low-rotation relu networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[48] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[49] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[50] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[51] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

## Appendix A. Background on Optimization in Measure Space and MFLD

**Notation.** We denote the Euclidean inner product with $\langle \cdot, \cdot \rangle$, the Euclidean norm for vectors and the operator norm for matrices with $\|\cdot\|$, and the Frobenius norm with $\|\cdot\|_{\mathrm{F}}$. We use $\mathcal{P}(\mathcal{W})$, $\mathcal{P}_2(\mathcal{W})$, and $\mathcal{P}_2^{\mathrm{ac}}(\mathcal{W})$ to denote the set of (Borel) probability measures, the set of probability measures with finite second moment, and the set of absolutely continuous probability measures with finite second moment on the weight space $\mathcal{W}$, respectively. Finally, $\delta_{\boldsymbol{w}_0}$ denotes the Dirac measure at $\boldsymbol{w}_0$.

**Optimization in measure space.** To minimize the regularized empirical risk $\hat{J}_\lambda$ defined in Section 2, we will consider a discretization of the following set of SDEs, which essentially define an interacting particle system over $m$ neurons:

$$\mathrm{d}\boldsymbol{w}_j^t = -m\nabla_{\boldsymbol{w}_j} J(\boldsymbol{w}_1^t, \dots, \boldsymbol{w}_m^t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}\boldsymbol{B}_t^j \quad \text{for} \quad 1 \le j \le m, \tag{A.1}$$

where $(\boldsymbol{B}_t^j)_{j=1}^m$ is a set of independent Brownian motions on the weight space $\mathcal{W}$.

Notice that the neural network and the regularizer in (2.2) are both invariant under permutations of the weights $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_m)$; thus, an equivalent integral representation is given by

$$\hat{y}(\boldsymbol{x}; \mu_{\boldsymbol{W}}) \coloneqq \int \Psi(\boldsymbol{x}; \cdot)\mathrm{d}\mu_{\boldsymbol{W}} \quad \text{and} \quad \mathcal{R}(\mu_{\boldsymbol{W}}) \coloneqq \int \|\cdot\|^2 \mathrm{d}\mu_{\boldsymbol{W}} \quad \text{with} \quad \mu_{\boldsymbol{W}} = \frac{1}{m}\sum_{j=1}^m \delta_{\boldsymbol{w}_j}. \tag{A.2}$$

Indeed, $\hat{y}(\boldsymbol{x}; \mu_{\boldsymbol{W}}) = \hat{y}_m(\boldsymbol{x}; \boldsymbol{W})$ and $\mathcal{R}(\mu_{\boldsymbol{W}}) = R(\boldsymbol{W})$, and this formulation allows extension to infinite-width networks by letting $\mu \in \mathcal{P}_2(\mathcal{W})$. Thus, we rewrite the population and the empirical risks in the space of measures as

$$\mathcal{J}_\lambda(\mu_{\boldsymbol{W}}) \coloneqq J_\lambda(\boldsymbol{W}) \quad \text{and} \quad \hat{\mathcal{J}}_\lambda(\mu_{\boldsymbol{W}}) \coloneqq \hat{J}_\lambda(\boldsymbol{W}),$$

and allow their domain to be all $\mu \in \mathcal{P}_2(\mathcal{W})$. We can equivalently state the interacting SDE system (A.1) as (see e.g. [16, Proposition 2.4])

$$\mathrm{d}\boldsymbol{w}_j^t = -\nabla_{\boldsymbol{w}} \hat{\mathcal{J}}_\lambda'[\mu_{\boldsymbol{W}^t}](\boldsymbol{w}_j^t) + \sqrt{\frac{2}{\beta}}\mathrm{d}\boldsymbol{B}_t^j \quad \text{for} \quad 1 \le j \le m, \tag{A.3}$$

where $\hat{\mathcal{J}}_\lambda'[\mu] \in L^2(\mathcal{W})$ denotes the first variation [42, Definition 7.12] of $\hat{\mathcal{J}}_\lambda(\mu)$.

As $m \to \infty$, the stochastic empirical measure $\mu_{\boldsymbol{W}^t}$ weakly converges to a deterministic measure $\mu_t$ for all fixed $t$, a phenomenon known as the *propagation of chaos* [46]. Furthermore, $\mu_t$ can be characterized as the law of the solution of the following SDE and non-linear Fokker-Planck equation

$$\mathrm{d}\boldsymbol{w}^t = -\nabla_{\boldsymbol{w}} \hat{\mathcal{J}}_\lambda'[\mu_t](\boldsymbol{w}^t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}\boldsymbol{B}_t \quad \text{and} \quad \partial_t \mu_t = \boldsymbol{\nabla} \cdot (\mu_t \nabla \hat{\mathcal{J}}_\lambda'[\mu_t]) + \beta^{-1}\boldsymbol{\Delta}\mu_t, \tag{A.4}$$

where $\boldsymbol{\nabla}\cdot$ and $\boldsymbol{\Delta}$ are the divergence and Laplacian operators, respectively. Due to the existence of mean-field interactions, (A.4) is known as the *mean-field Langevin dynamics* (MFLD).

For a pair of probability measures $\mu \ll \nu$ both in $\mathcal{P}(\mathcal{W})$, we define the relative entropy $\mathcal{H}(\mu \,|\, \nu)$ and the relative Fisher information $\mathcal{I}(\mu \,|\, \nu)$ respectively as

$$\mathcal{H}(\mu \,|\, \nu) \coloneqq \int_{\mathcal{W}} \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\mu \quad \text{and} \quad \mathcal{I}(\mu \,|\, \nu) \coloneqq \int_{\mathcal{W}} \left\| \nabla \ln \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|^2 \mathrm{d}\mu. \tag{A.5}$$

It is well-known at this point that $\mu_t$ in (A.4) can be interpreted as the Wasserstein gradient flow of the entropic regularized functional $\mathcal{F}_\beta(\mu) \coloneqq \hat{\mathcal{J}}_\lambda(\mu) + \frac{1}{\beta}\mathcal{H}(\mu \,|\, \tau)$, where $\tau$ is the uniform measure on compact $\mathcal{W}$ or the Lebesgue measure on a Euclidean space [3, 31, 49]. For this gradient flow to converge exponentially fast towards the minimizer $\mu^*_\beta \coloneqq \arg\min_\mu \mathcal{F}_\beta(\mu)$, we require a *gradient domination* condition on $\mu^*_\beta$ in the space of probability measures, given as

$$\mathcal{H}(\mu \,|\, \mu^*_\beta) \le \frac{C_{\mathrm{LSI}}}{2}\mathcal{I}(\mu \,|\, \mu^*_\beta), \quad \forall \mu \in \mathcal{P}(\mathcal{W}), \tag{A.6}$$

which is referred to as the log-Sobolev inequality (LSI). If the measure $\mathrm{d}\nu_{\mu_t} \propto \exp(-\beta \mathcal{F}'[\mu_t])\mathrm{d}\tau$ satisfies LSI with constant $C_{\mathrm{LSI}}$ for all $t \ge 0$, $\mu_t$ enjoys the following exponential convergence

$$\mathcal{F}_\beta(\mu_t) - \mathcal{F}_\beta(\mu^*_\beta) \le e^{\frac{-2t}{\beta C_{\mathrm{LSI}}}}(\mathcal{F}_\beta(\mu_0) - \mathcal{F}_\beta(\mu^*_\beta)); \tag{A.7}$$

see e.g. [16, Theorem 3.2] and [38, Theorem 1].

## Appendix B. Proofs of Section 3

Before presenting the layout of proofs, we introduce a useful reformulation of the objective $\mathcal{F}_{\beta,\lambda}(\mu)$. Recall that

$$\mathcal{F}_{\beta,\lambda}(\mu) = \hat{\mathcal{J}}_0(\mu) + \frac{\lambda}{2}\mathcal{R}(\mu) + \frac{1}{\beta}\mathcal{H}(\mu).$$

Let $\gamma \propto \exp\left(\frac{-\lambda\beta}{2}\|\boldsymbol{w}\|^2\right)$ be the centered Gaussian measure on $\mathbb{R}^{2d+2}$ with variance $1/(\lambda\beta)$. Then, we can rewrite the above as

$$\mathcal{F}_{\beta,\lambda}(\mu) = \hat{\mathcal{J}}_0(\mu) + \frac{1}{\beta}\mathcal{H}(\mu \,|\, \gamma) - \frac{d}{2}\ln\left(\frac{\lambda\beta}{2}\right).$$

As a result, we can define

$$\tilde{\mathcal{F}}_{\beta,\lambda}(\mu) \coloneqq \hat{\mathcal{J}}_0(\mu) + \frac{1}{\beta}\mathcal{H}(\mu \,|\, \gamma), \tag{B.1}$$

which is non-negative and equivalent to $\mathcal{F}_\beta$ up to an additive constant. Notice that

$$\mu^*_\beta \coloneqq \arg\min_\mu \mathcal{F}_{\beta,\lambda}(\mu) = \arg\min_\mu \tilde{\mathcal{F}}_{\beta,\lambda}(\mu).$$

This reformulation, which was also used in [45], allows us to combine the effect of weight decay and entropic regularization into a single non-negative term $\mathcal{H}(\mu \,|\, \gamma)$. Furthermore, the simple density expression for the Gaussian measure $\gamma$ allows us to achieve useful estimates for $\mathcal{H}(\mu \,|\, \gamma)$. In particular, as we will show below, it is possible to control $\mathcal{H}(\mu^*_\beta \,|\, \gamma)$ with effective dimension rather than ambient dimension, which leads to dependence on $d_{\mathrm{eff}}$ rather than $d$ in our bounds.

We break down the proof of Theorem 2 into three steps:

1. In Section B.2 we show that there exists a measure $\mu^* \in \mathcal{P}_2(\mathbb{R}^{2d+2})$ where $\hat{y}(\cdot; \mu^*)$ can approximate $g$ on the training set with bounds on $\mathcal{R}(\mu^*)$. This construction provides upper bounds on $\hat{\mathcal{J}}_0(\mu^*_\beta)$ and $\mathcal{H}(\mu^*_\beta \,|\, \gamma)$.

2. In Section B.3, we perform a generalization analysis via Rademacher complexity tools given the bound on $\mathcal{H}(\mu^*_\beta \,|\, \gamma)$, leading to a bound on $\mathcal{J}_0(\mu^*_\beta)$.

3. Finally, in Section B.4, we estimate the LSI constant and constants related to smoothness/discretization along the trajectory, which imply that $\mathcal{F}^m_{\beta,\lambda}(\mu^m_l)$ converges to $\mathcal{F}_\beta(\mu^*_\beta)$, where $\mathcal{F}^m_{\beta,\lambda}$ is an adjusted objective over $\mathcal{P}(\mathbb{R}^{(2d+2)m})$ defined in (B.6). This bound implies the convergence of $\mathbb{E}_{\boldsymbol{W}\sim\mu^m_l}[J_0(\boldsymbol{W})]$ to $J_0(\mu^*_\beta)$, which was bounded in the previous step.

Before laying out these steps, in Section B.1, we will introduce the required concentration results. In the following, we will use the unregularized population $\mathcal{J}_0(\mu) \coloneqq \mathbb{E}[\ell(\hat{y}(\boldsymbol{x};\mu), y)]$ and empirical $\hat{\mathcal{J}}_0(\mu) = \mathbb{E}_{S_n}[\ell(\hat{y}(\boldsymbol{x};\mu), y)]$ risks, and also consider the finite-width versions $J_0(\boldsymbol{W}) \coloneqq \mathcal{J}_0(\mu_{\boldsymbol{W}})$ and $\hat{J}_0(\boldsymbol{W}) \coloneqq \hat{\mathcal{J}}_0(\mu_{\boldsymbol{W}})$.

### B.1. Concentration Bounds

We begin by specifying the definition of subGaussian and subexponential random variables in our setting.

**Definition 5** [50] *A random variable $x$ is $\sigma$-subGaussian if $\mathbb{E}\big[e^{\lambda(x-\mathbb{E}[x])}\big] \le e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$, and is $(\nu,\alpha)$-subexponential if $\mathbb{E}\big[e^{\lambda(x-\mathbb{E}[x])}\big] \le e^{\lambda^2\nu^2/2}$ for all $|\lambda| \le 1/\alpha$. If $x$ is $\sigma$-subGaussian, then*

$$\mathbb{P}(x - \mathbb{E}[x] \ge t) \le \exp\big(\frac{-t^2}{2\sigma^2}\big). \tag{B.2}$$

*If $x$ is $(\nu,\alpha)$-subexponential, then*

$$\mathbb{P}(x - \mathbb{E}[x] \ge t) \le \exp\big(-\frac{1}{2}\min\big(\frac{t^2}{\nu^2}, \frac{t}{\alpha}\big)\big) \tag{B.3}$$

Moreover, for centered random variables, let $|\cdot|_{\psi_2}$ and $|\cdot|_{\psi_1}$ denote the subGaussian and subexponential norm respectively [48, Definitions 2.5.6 and 2.7.5]. Then $x$ is $\sigma$-subGaussian if and only if $\sigma \asymp |x - \mathbb{E}[x]|_{\psi_2}$, and is $(\nu,\nu)$-subexponential if and only if $\nu \asymp |x - \mathbb{E}[x]|_{\psi_1}$.

Next, we bound several quantities that appear in various parts of our proofs.

**Lemma 6** *Under Assumption 1, for any $q > 0$ and all $1 \le i \le n$, with probability at least $1 - n^{-q}$,*

$$\Big\|\boldsymbol{U}\boldsymbol{x}^{(i)}\Big\| \le r_x\big(1 + \sigma_u\sqrt{2(q+1)\ln n}\big) = \tilde{r}_x. \tag{B.4}$$

**Proof** By subGaussianity of $\|\boldsymbol{U}\boldsymbol{x}\|$ from Assumption 1 and the subGaussian tail bound, with probability at least $1 - n^{-q-1}$

$$\begin{aligned}\Big\|\boldsymbol{U}\boldsymbol{x}^{(i)}\Big\| &\le \mathbb{E}[\|\boldsymbol{U}\boldsymbol{x}\|] + \sigma_u r_x\sqrt{2(q+1)\ln n} \\ &= r_x + \sigma_u r_x\sqrt{2(q+1)\ln n}.\end{aligned}$$

The statement of lemma follows from a union bound over $1 \le i \le n$. ∎

**Lemma 7** *Under Assumption 1, we have $\mathbb{E}_{S_n}\Big[\|\boldsymbol{x}\|^2\Big] \lesssim c_x^2$ with probability at least $1 - \exp(-\Omega(n))$.*

**Proof** By the triangle inequality,

$$|||\boldsymbol{x}|||_{\psi_2} \leq |||\boldsymbol{x}|| - \mathbb{E}[||\boldsymbol{x}||]|_{\psi_2} + |\mathbb{E}[||\boldsymbol{x}||]|_{\psi_2} \lesssim \sigma_n \left\|\boldsymbol{\Sigma}^{1/2}\right\|_{\mathrm{F}} + \mathrm{tr}(\boldsymbol{\Sigma})^{1/2} \lesssim \mathrm{tr}(\boldsymbol{\Sigma})^{1/2}.$$

Recall $c_x^2 \coloneqq \mathrm{tr}(\boldsymbol{\Sigma})$. Furthermore, by [48, Lemma 2.7.6] we have

$$\left|\|\boldsymbol{x}\|^2\right|_{\psi_1} = |||\boldsymbol{x}|||_{\psi_2}^2 \lesssim c_x^2.$$

We arrive at a similar result for the centered random variable $\|\boldsymbol{x}\|^2 - \mathbb{E}[\|\boldsymbol{x}\|]^2 = \|\boldsymbol{x}\|^2 - c_x^2$. We conclude the proof by the subexponential tail inequality,

$$\mathbb{P}\Big(\mathbb{E}_{S_n}\Big[\|\boldsymbol{x}\|^2\Big] - c_x^2 \geq t c_x^2\Big) \leq \exp(-\min(t, t^2)\Omega(n)).$$

■

**Lemma 8** *Under Assumption 1, we have* $\mathbb{E}_{S_n}\big[y^2\big] \lesssim 1$ *with probability at least* $1 - 2\exp(-\Omega(n))$.

**Proof** We have

$$|y|^2 \leq 3g(0)^2 + 3\mathcal{O}(1/r_x^2)\|\boldsymbol{U}\boldsymbol{x}\|^2 + 3\xi^2.$$

By a similar argument to Lemma 7 we have

$$\left|\|\boldsymbol{U}\boldsymbol{x}\|^2\right|_{\psi_1} = |||\boldsymbol{U}\boldsymbol{x}|||_{\psi_2}^2 \leq 2|||\boldsymbol{U}\boldsymbol{x}|| - \mathbb{E}[||\boldsymbol{U}\boldsymbol{x}||]|_{\psi_2}^2 + 2\,\mathbb{E}[||\boldsymbol{U}\boldsymbol{x}||]^2 \lesssim (1+\sigma_u^2) r_x^2,$$

since $\mathbb{E}\Big[\|\boldsymbol{U}\boldsymbol{x}\|^2\Big] = r_x^2$. As a result, by the subexponential tail bound,

$$\mathbb{E}_{S_n}\Big[\|\boldsymbol{U}\boldsymbol{x}\|^2\Big] - \mathbb{E}\Big[\|\boldsymbol{U}\boldsymbol{x}\|^2\Big] \lesssim (1+\sigma_u^2) r_x^2 \lesssim r_x^2,$$

with probability at least $1 - \exp(-\Omega(n))$. Similarly, $\left|\xi^2\right|_{\psi_1} \leq |\xi|_{\psi_2}^2 \lesssim \varsigma^2$, therefore,

$$\mathbb{E}_{S_n}\big[\xi^2\big] - \mathbb{E}\big[\xi^2\big] \lesssim \varsigma^2 \lesssim 1,$$

with probability at least $1 - \exp(-\Omega(n))$. The statement of the lemma follows by a union bound. ■

**Lemma 9** *Under Assumption 1, for any* $q > 0$ *and* $n \gtrsim \frac{c_x^2}{\|\boldsymbol{\Sigma}\|}(1 + \sigma_n^2(q+1)\ln(n))\ln(dn^q)$, *with probability at least* $1 - \mathcal{O}(n^{-q})$ *we have* $\left\|\mathbb{E}_{S_n}\big[\boldsymbol{x}\boldsymbol{x}^\top\big]\right\| \lesssim \|\boldsymbol{\Sigma}\|$. *Further, if* $q \geq 1$, *then* $\mathbb{E}\Big[\left\|\mathbb{E}_{S_n}\big[\boldsymbol{x}\boldsymbol{x}^\top\big]\right\|^{1/2}\Big] \lesssim \|\boldsymbol{\Sigma}\|^{1/2}$.

**Proof** First, note that by subGaussianity of $\|\boldsymbol{x}\|$, for every fixed $i$, we have with probability at least $1 - n^{-q-1}$,

$$\left\|\boldsymbol{x}^{(i)}\right\| - \mathbb{E}[\|\boldsymbol{x}\|] \leq \sigma_n \left\|\boldsymbol{\Sigma}^{1/2}\right\|_{\mathrm{F}} \sqrt{2(q+1)\ln n}.$$

Since $\mathbb{E}[\|\boldsymbol{x}\|] \le c_x$, via a union bound, with probability at least $1 - n^{-q}$,

$$\left\|\boldsymbol{x}^{(i)}\right\| \le c_x + \sigma_n c_x \sqrt{2(q+1)\ln n} =: \tilde{c}_x.$$

Define the clipped version of $\boldsymbol{x}$ via $\boldsymbol{x}_c = \boldsymbol{x}(1 \wedge \frac{\tilde{c}_x}{\|\boldsymbol{x}\|})$. Then, on the above event,

$$\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right] = \mathbb{E}_{S_n}\left[\boldsymbol{x}_c\boldsymbol{x}_c^\top\right].$$

Moreover,

$$\left\|\mathbb{E}\left[\boldsymbol{x}_c\boldsymbol{x}_c^\top\right]\right\| = \sup_{\|\boldsymbol{v}\|\le 1} \mathbb{E}\left[\langle \boldsymbol{x}_c, \boldsymbol{v}\rangle^2\right] \le \sup_{\|\boldsymbol{v}\|\le 1} \mathbb{E}\left[\langle \boldsymbol{x}, \boldsymbol{v}\rangle^2\right] = \left\|\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|.$$

Finally, by the covariance estimation bound of [50, Corollary 6.20] for centered subGaussian random vectors and the condition on $n$ given in the statement of the lemma,

$$\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}_c\boldsymbol{x}_c^\top\right]\right\| - \left\|\mathbb{E}\left[\boldsymbol{x}_c\boldsymbol{x}_c^\top\right]\right\| \lesssim \left\|\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|$$

with probability at least $1 - \mathcal{O}(n^{-q})$. Consequently, we have $\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\| \lesssim \|\boldsymbol{\Sigma}\|$ with probability at least $1 - \mathcal{O}(n^{-q})$.

For the second part of the lemma, let $E$ denote the event on which the above $\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\| \lesssim \|\boldsymbol{\Sigma}\|$ holds. Then,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|^{1/2}\right] &= \mathbb{E}\left[\mathbb{1}(E)\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|^{1/2}\right] + \mathbb{E}\left[\mathbb{1}(E^C)\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|^{1/2}\right] \\
&\lesssim \|\boldsymbol{\Sigma}\|^{1/2} + \mathbb{P}\big(E^C\big)^{1/2}\,\mathbb{E}\left[\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|\right]^{1/2} \\
&\lesssim \|\boldsymbol{\Sigma}\|^{1/2} + \mathcal{O}(n^{-q/2})c_x.
\end{aligned}$$

Suppose $q \ge 1$. Then for $n \gtrsim c_x^2/\|\boldsymbol{\Sigma}\|$, we have $\mathbb{E}\left[\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|^{1/2}\right] \lesssim \|\boldsymbol{\Sigma}\|^{1/2}$, which completes the proof. ■

We summarize the above results into a single event.

**Lemma 10** *Suppose $n \gtrsim \frac{c_x^2}{\|\boldsymbol{\Sigma}\|}(1 + \sigma_n^2(q+1)\ln(n))\ln(dn^q)$. There exists an event $\mathcal{E}$ such that $\mathbb{P}(\mathcal{E}) \ge 1 - \mathcal{O}(n^{-q})$, and on $\mathcal{E}$:*

1. $\left\|\boldsymbol{U}\boldsymbol{x}^{(i)}\right\| \le \tilde{r}_x$ *for all* $1 \le i \le n$.
2. $\mathbb{E}_{S_n}\left[\|\boldsymbol{x}\|^2\right] \lesssim c_x^2$.
3. $\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\| \lesssim \|\boldsymbol{\Sigma}\|$.
4. $\mathbb{E}\left[\left\|\mathbb{E}_{S_n}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right\|^{1/2}\right] \lesssim \|\boldsymbol{\Sigma}\|^{1/2}$.
5. $\mathbb{E}_{S_n}\left[y^2\right] \lesssim 1$.

We recall the variational lower bound for the KL divergence, which will be used at various stages of different proofs to relate certain expectations to the KL divergence.

**Lemma 11 (Donsker-Varadhan Variational Formula for KL Divergence [26])** *Let $\mu$ and $\nu$ be probability measures on $\mathcal{W}$. Then,*

$$\mathcal{H}(\mu \,|\, \nu) = \sup_{f:\mathcal{W}\to\mathbb{R}} \int f \mathrm{d}\mu - \ln\Big(\int e^f \mathrm{d}\nu\Big).$$

Finally, we state the following lemma which will be useful in estimating smoothness constants in the convergence analysis.

**Lemma 12** *Suppose $(z, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^d$ are drawn from a probability distribution $\mathcal{D}$. Then,*

$$\|\mathbb{E}_{\mathcal{D}}[z\boldsymbol{x}]\| \le \sqrt{\mathbb{E}_{\mathcal{D}}[z^2]\|\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}\boldsymbol{x}^\top]\|}.$$

**Proof** We have

$$\begin{aligned}
\|\mathbb{E}_{\mathcal{D}}[z\boldsymbol{x}]\| = \sup_{\|\boldsymbol{v}\|\le 1} \langle \boldsymbol{v}, \mathbb{E}_{\mathcal{D}}[z\boldsymbol{x}]\rangle &= \sup_{\|\boldsymbol{v}\|\le 1} \mathbb{E}_{\mathcal{D}}[z\langle \boldsymbol{v}, \boldsymbol{x}\rangle] \\
&\le \sup_{\|\boldsymbol{v}\|\le 1} \sqrt{\mathbb{E}_{\mathcal{D}}[z^2]\,\mathbb{E}_{\mathcal{D}}\Big[\langle \boldsymbol{v}, \boldsymbol{x}\rangle^2\Big]} \quad \text{(Cauchy-Schwartz)} \\
&\le \sqrt{\mathbb{E}_{\mathcal{D}}[z^2] \sup_{\|\boldsymbol{v}\|\le 1} \langle \boldsymbol{v}, \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{v}\rangle} \\
&= \sqrt{\mathbb{E}_{\mathcal{D}}[z^2]\|\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}\boldsymbol{x}^\top]\|}.
\end{aligned}$$

■

Notice that the distribution $\mathcal{D}$ can be both the empirical as well as the population distribution.

### B.2. Approximating the Target Function

We begin by stating the following approximation lemma which is the result of [7, Proposition 6] adapted to our setting.

**Proposition 13** *Suppose $g : \mathbb{R}^k \to \mathbb{R}$ is L-Lipschitz and $|g(0)| = \mathcal{O}(L\tilde{r}_x)$. On the event of Lemma 10, there exists a measure $\mu \in \mathcal{P}_2(\mathbb{R}^{2d+2})$ with $\mathcal{R}(\mu) \le \Delta^2/\tilde{r}_x^2$ such that*

$$\max_i \Big| g(\boldsymbol{U}\boldsymbol{x}^{(i)}) - \hat{y}(\boldsymbol{x}^{(i)}; \mu)\Big| \le C_k L\tilde{r}_x \Big(\frac{\Delta}{L\tilde{r}_x}\Big)^{\frac{-2}{k+1}} \ln\Big(\frac{\Delta}{L\tilde{r}_x}\Big) + \frac{\ln 4}{\kappa},$$

*for all $\Delta \ge C_k$, where $C_k$ is a constant depending only on $k$, provided that the hyperparameter $\iota$ satisfies $\iota \ge C_k L\tilde{r}_x \Big(\frac{\Delta}{L\tilde{r}_x}\Big)^{2k/(k+1)}$.*

**Proof** Throughout the proof, we will use $C_k$ to denote a constant that only depends on $k$, whose value may change across instantiations. Let $\boldsymbol{z} := \boldsymbol{U}\boldsymbol{x} \in \mathbb{R}^k$ and $\tilde{\boldsymbol{z}} := (\boldsymbol{z}^\top, \tilde{r}_x)^\top \in \mathbb{R}^{k+1}$. Recall that on the event of Lemma 6 we have $\|\boldsymbol{z}^{(i)}\| \le \tilde{r}_x$ and $|g(\boldsymbol{z}^{(i)})| \lesssim L\tilde{r}_x$ for all $1 \le i \le n$. Let $\tau$

denote the uniform probability measure on $\mathbb{S}^k$. By [7, Proposition 6], for all $\Delta \geq C_k$, there exists $p \in L^2(\tau)$ with $\|p\|_{L^2(\tau)} \leq \Delta$ such that

$$\max_i \left| g(\boldsymbol{z}^{(i)}) - \int_{\mathbb{S}^k} p(\boldsymbol{v}) \phi_\infty\Big(\frac{1}{\tilde{r}_x}\Big\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}^{(i)}\Big\rangle\Big) \mathrm{d}\tau(\boldsymbol{v}) \right| \leq C_k L \tilde{r}_x \big(\frac{\Delta}{L\tilde{r}_x}\big)^{\frac{-2}{k+1}} \ln\big(\frac{\Delta}{L\tilde{r}_x}\big).$$

In fact, we have a stronger guarantee on $p$. Specifically, $p(\boldsymbol{v})$ is given by

$$p(\boldsymbol{v}) = \sum_{j \geq 1} \lambda_j^{-1} r^j h_j(\boldsymbol{v}),$$

where $r \in (0,1), \lambda_j, h_j : \mathbb{S}^k \to \mathbb{R}$ are introduced by [7, Appendix D]. In particular,

$$h(\boldsymbol{v}) = g\Big(\frac{\tilde{r}_x \boldsymbol{v}_{1:k}}{\boldsymbol{v}_{k+1}}\Big)\boldsymbol{v}_{k+1},$$

with the spheircal harmonics decomposition $h(\boldsymbol{v}) = \sum_{j \geq 0} h_j(\boldsymbol{v})$. It is shown in [7, Appendix D.2] that $\lambda_j \leq C_k j^{(k+1)/2}$, and one can prove through spherical harmonics calculations (omitted here for brevity) that $|h_j(\boldsymbol{v})| \leq C_k \sup_{\boldsymbol{v} \in \mathbb{S}^k} h(\boldsymbol{v}) j^{(k-1)/2} \leq C_k L \tilde{r}_x j^{(k-1)/2}$. As a result,

$$|p(\boldsymbol{v})| \leq \sum_{j \geq 0} \lambda_j^{-1} r^j |h_j(\boldsymbol{v})| \leq \sum_{j \geq 1} \lambda_j^{-1} r^j |h_j(\boldsymbol{v})| \leq C_k L \tilde{r}_x \sum_{j \geq 1} j^k r^j \leq \frac{C_k L \tilde{r}_x}{(1-r)^k}.$$

Using $1 - r = \Big(C_k L \tilde{r}_x / \Delta\Big)^{2/(k+1)}$ as in [7, Appendix D.4] yields

$$|p(\boldsymbol{v})| \leq C_k L \tilde{r}_x \Big(\frac{\Delta}{L\tilde{r}_x}\Big)^{2k/(k+1)}.$$

Define $p_+(\boldsymbol{v}) \coloneqq p(\boldsymbol{v}) \vee 0$ and $p_-(\boldsymbol{v}) \coloneqq (-p(\boldsymbol{v})) \vee 0$. Then, by positive 1-homogeneity of ReLU,

$$\begin{aligned}
\int_{\mathbb{S}^k} p(\boldsymbol{v}) \phi_\infty\Big(\frac{1}{\tilde{r}_x}\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}\rangle\Big) \mathrm{d}\tau(\boldsymbol{v}) &= \int_{\mathbb{S}^k} p_+(\boldsymbol{v}) \phi_\infty\Big(\frac{1}{\tilde{r}_x}\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}\rangle\Big) \mathrm{d}\tau(\boldsymbol{v}) - \int_{\mathbb{S}^k} p_-(\boldsymbol{v}) \phi_\infty\Big(\frac{1}{\tilde{r}_x}\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}\rangle\Big) \mathrm{d}\tau(\boldsymbol{v}) \\
&= \int_{\mathbb{S}^k} \phi_\infty\Big(\frac{p_+(\boldsymbol{v})}{\tilde{r}_x}\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}\rangle\Big) \mathrm{d}\tau(\boldsymbol{v}) - \int_{\mathbb{S}^k} \phi_\infty\Big(\frac{p_-(\boldsymbol{v})}{\tilde{r}_x}\Big\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}^{(i)}\Big\rangle\Big) \mathrm{d}\tau(\boldsymbol{v}) \\
&= \int_{\mathbb{R}^{k+1}} \phi_\infty(\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}\rangle) \mathrm{d}\tilde{\mu}_1(\boldsymbol{v}) - \int_{\mathbb{R}^{k+1}} \phi_\infty(\langle \boldsymbol{v}, \tilde{\boldsymbol{z}}\rangle) \mathrm{d}\tilde{\mu}_2(\boldsymbol{v}) \\
&= \int_{\mathbb{R}^{d+1}} \phi_\infty(\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}\rangle) \mathrm{d}\mu_1(\boldsymbol{w}) - \int_{\mathbb{R}^{d+1}} \phi_\infty(\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}\rangle) \mathrm{d}\mu_2(\boldsymbol{w}),
\end{aligned}$$

where $\tilde{\mu}_1 \coloneqq \frac{(\cdot)p_+(\cdot)}{\tilde{r}_x}\#\tau$ and $\tilde{\mu}_2 \coloneqq \frac{(\cdot)p_-(\cdot)}{\tilde{r}_x}\#\tau$ are the corresponding pushforward measures, $\mu_1 = T_{\boldsymbol{U}}\#\tilde{\mu}_1$ and $\mu_2 = T_{\boldsymbol{U}}\#\tilde{\mu}_2$, where $T_{\boldsymbol{U}}(\boldsymbol{v}) = (\boldsymbol{U}^\top \boldsymbol{v}_k, v_{k+1})^\top \in \mathbb{R}^{d+1}$ for $\boldsymbol{v} = (\boldsymbol{v}_k^\top, v_{k+1})^\top \in \mathbb{R}^{k+1}$. In other words, $\boldsymbol{w} \sim \mu_1$ is generated by sampling $\boldsymbol{v} \sim \tilde{\mu}_1$ and letting $\boldsymbol{w} = (\boldsymbol{U}^\top \boldsymbol{v}_k, v_{k+1})^\top$, with a similar procedure for $\boldsymbol{w} \sim \mu_2$. Furthermore,

$$\begin{aligned}
\mathcal{R}(\mu) = \int_{\mathbb{R}^{d+1}} \|\boldsymbol{w}\|^2 \mathrm{d}\mu_1(\boldsymbol{w}) + \int_{\mathbb{R}^{d+1}} \|\boldsymbol{w}\|^2 \mathrm{d}\mu_2(\boldsymbol{w}) &= \int_{\mathbb{R}^{k+1}} \|\boldsymbol{v}\|^2 \mathrm{d}\tilde{\mu}_1(\boldsymbol{v}) + \int_{\mathbb{R}^{k+1}} \|\boldsymbol{v}\|^2 \mathrm{d}\tilde{\mu}_2(\boldsymbol{v}) \\
&= \int_{\mathbb{S}^k} \frac{p(\boldsymbol{v})^2}{\tilde{r}_x^2} \mathrm{d}\tau(\boldsymbol{v}) \leq \frac{\Delta^2}{\tilde{r}_x^2}.
\end{aligned}$$

The last step is to replace $\phi_\infty$ with $\phi_{\kappa,\iota}$. Note that for all $i$, and almost surely over $\boldsymbol{w} \sim \mu_1$, we have $\left|\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\right| \le p_+(\boldsymbol{v}) \le C_k L\tilde{r}_x \left(\frac{\Delta}{L\tilde{r}_x}\right)^{2k/(k+1)}$, with a similar bound holding for $\boldsymbol{w} \sim \mu_2$. As a result, by choosing $\iota \ge C_k L\tilde{r}_x \left(\frac{\Delta}{L\tilde{r}_x}\right)^{2k/(k+1)}$, we have $\phi_{\varkappa,\iota}\Big(\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\Big) = \phi_{\varkappa}\Big(\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\Big)$ for all $i$ and almost surely over $\boldsymbol{w} \sim \mu_1$ and $\boldsymbol{w} \sim \mu_2$. By the triangle inequality, we have

$$\begin{aligned}
\left|g(\boldsymbol{U}\boldsymbol{x}^{(i)}) - \hat{y}(\boldsymbol{x}^{(i)}; \mu)\right| \le & \left|\left\{\int \phi_{\kappa,\iota}\Big(\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\Big) - \phi_\infty\Big(\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\Big)\right\} \mathrm{d}\mu_1(\boldsymbol{w})\right| \\
& + \left|\left\{\int \phi_{\kappa,\iota}\Big(\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\Big) - \phi_\infty\Big(\langle \boldsymbol{w}, \tilde{\boldsymbol{x}} \rangle^{(i)}\Big)\right\} \mathrm{d}\mu_2(\boldsymbol{w})\right| \\
& + \left|g(\boldsymbol{U}\boldsymbol{x}^{(i)}) - \int \phi_\infty\Big(\left\langle \boldsymbol{w}, \tilde{\boldsymbol{x}}^{(i)} \right\rangle\Big)(\mathrm{d}\mu_1(\boldsymbol{w}) - \mathrm{d}\mu_2(\boldsymbol{w}))\right| \\
\le & \frac{2\ln 2}{\kappa} + C_k L\tilde{r}_x \big(\frac{\Delta}{L\tilde{r}_x}\big)^{\frac{-2}{k+1}} \ln\big(\frac{\Delta}{L\tilde{r}_x}\big),
\end{aligned}$$

which completes the proof. ■

Next, we control the effect of entropic regularization on the minimum of $\tilde{\mathcal{F}}_{\beta,\lambda}$ via the following lemma.

**Lemma 14** *Suppose $\rho$ is $C_\rho$ Lipschitz. For every $\mu^* \in \mathcal{P}(\mathbb{R}^{2d+2})$, we have*

$$\min_{\mu \in \mathcal{P}^{\mathrm{ac}}(\mathbb{R}^{2d+2})} \tilde{\mathcal{F}}_{\beta,\lambda}(\mu) \le \hat{\mathcal{J}}_0(\mu^*) + \frac{\lambda}{2}\mathcal{R}(\mu^*) + \frac{2\sqrt{2}C_\rho}{\sqrt{\pi\lambda\beta}} \mathbb{E}_{S_n}[\|\tilde{\boldsymbol{x}}\|].$$

**Proof** We will smooth $\mu^*$ by convoliving it with $\gamma$, i.e. we consider $\mu = \mu^* * \gamma$. Let $\boldsymbol{u} \sim \gamma$ independent of $\boldsymbol{w} \sim \mu^*$ and denote $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top)^\top$ with $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^{d+1}$. We first bound $\hat{\mathcal{J}}_0(\mu^* * \gamma)$. Using the Lipschitzness of the loss and of $\phi_{\kappa,\iota}$, we have

$$\begin{aligned}
\hat{\mathcal{J}}_0(\mu^* * \gamma) - \hat{\mathcal{J}}_0(\mu^*) = & \mathbb{E}_{S_n}\left[\ell\Big(\int \Psi(\boldsymbol{x}; \boldsymbol{w}) \mathrm{d}(\mu^* * \gamma)(\boldsymbol{w}) - y\Big) - \ell\Big(\int \Psi(\boldsymbol{x}; \boldsymbol{w}) \mathrm{d}\mu^*(\boldsymbol{w}) - y\Big)\right] \\
\le & C_\rho \mathbb{E}_{S_n}\left[\left|\int \Psi(\boldsymbol{x}; \boldsymbol{w}) \mathrm{d}(\mu^* * \gamma)(\boldsymbol{w}) - \int \Psi(\boldsymbol{x}; \boldsymbol{w}) \mathrm{d}\mu^*(\boldsymbol{w})\right|\right] \\
= & C_\rho \mathbb{E}_{S_n}\left[\left|\int (\mathbb{E}_{\boldsymbol{u}}[\Psi(\boldsymbol{x}; \boldsymbol{w} + \boldsymbol{u})] - \Psi(\boldsymbol{x}; \boldsymbol{w})) \mathrm{d}\mu^*(\boldsymbol{w})\right|\right] \\
\le & C_\rho \mathbb{E}_{S_n}\left[\int \mathbb{E}_{\boldsymbol{u}}[|\phi_{\kappa,\iota}(\langle \boldsymbol{\omega}_1 + \boldsymbol{u}_1, \tilde{\boldsymbol{x}} \rangle) - \phi_{\kappa,\iota}(\langle \boldsymbol{\omega}_1, \tilde{\boldsymbol{x}} \rangle)|] \mathrm{d}\mu^*(\boldsymbol{w})\right] \\
& + C_\rho \mathbb{E}_{S_n}\left[\int \mathbb{E}_{\boldsymbol{u}}[|\phi_{\kappa,\iota}(\langle \boldsymbol{\omega}_2 + \boldsymbol{u}_2, \tilde{\boldsymbol{x}} \rangle) - \phi_{\kappa,\iota}(\langle \boldsymbol{\omega}_2, \tilde{\boldsymbol{x}} \rangle)|] \mathrm{d}\mu^*(\boldsymbol{w})\right] \\
\le & C_\rho \mathbb{E}_{S_n}\left[\int \{\mathbb{E}_{\boldsymbol{u}_1}[|\langle \boldsymbol{u}_1, \tilde{\boldsymbol{x}} \rangle|] + \mathbb{E}_{\boldsymbol{u}_2}[|\langle \boldsymbol{u}_2, \tilde{\boldsymbol{x}} \rangle|]\} \mathrm{d}\mu^*(\boldsymbol{\omega})\right] \\
= & \frac{2\sqrt{2}C_\rho}{\sqrt{\pi\lambda\beta}} \mathbb{E}_{S_n}[\|\tilde{\boldsymbol{x}}\|].
\end{aligned}$$

Next, we bound the KL divergence via its convexity in the first argument,

$$\mathcal{H}(\mu^* * \gamma \,|\, \gamma) = \mathcal{H}\Big(\int \gamma(\cdot - \boldsymbol{w}') \mathrm{d}\mu^*(\boldsymbol{w}') \,|\, \gamma\Big) \le \int \mathcal{H}(\gamma(\cdot - \boldsymbol{w}') \,|\, \gamma(\cdot)) \mathrm{d}\mu^*(\boldsymbol{w}').$$

Furthermore,

$$\mathcal{H}(\gamma(\cdot - \boldsymbol{w}') \,|\, \gamma(\cdot)) = \int \frac{\lambda\beta}{2}\big( - \|\boldsymbol{w} - \boldsymbol{w}'\|^2 + \|\boldsymbol{w}\|^2\big)\gamma(\mathrm{d}\boldsymbol{w} - \boldsymbol{w}') = \frac{\lambda\beta\|\boldsymbol{w}'\|^2}{2}.$$

Consequently,

$$\mathcal{H}(\mu^* * \gamma \,|\, \gamma) \le \frac{\lambda\beta}{2}\mathcal{R}(\mu^*),$$

which finishes the proof. ■

Combining above results, we have the following statement.

**Corollary 15** *Suppose the event of Lemma 10 holds, $\rho$ is $C_\rho$ Lipschitz, and $\lambda \lesssim 1$. Then,*

$$\min_{\mu\in\mathcal{P}^{\mathrm{ac}}(\mathbb{R}^{2d+2})} \tilde{\mathcal{F}}_{\beta,\lambda}(\mu) - \mathbb{E}_{S_n}[\rho(\xi)] \lesssim C_\rho \frac{\tilde{r}_x}{r_x}\left(\frac{r_x\Delta}{\tilde{r}_x}\right)^{\frac{-2}{k+1}} \ln\left(\frac{r_x\Delta}{\tilde{r}_x}\right) + \frac{C_\rho}{\kappa} + \frac{\lambda\Delta^2}{\tilde{r}_x^2} + \frac{C_\rho(c_x + \tilde{r}_x)}{\sqrt{\lambda\beta}},$$

*for all $\Delta \ge C_k$, provided that $\iota \ge C_k\Delta^{2k/(k+1)}(r_x/\tilde{r}_x)^{(k-1)/(k+1)}$.*

**Proof** We will use Lemma 14 with $\mu^* \in \mathcal{P}(\mathbb{R}^{2d+2})$ constructed in Proposition 13. Then, for all $\Delta \ge C_k$,

$$\begin{aligned}
\hat{\mathcal{J}}_0(\mu^*) &= \mathbb{E}[\rho(\hat{y}(\boldsymbol{x};\mu^*) - y)] \\
&= \mathbb{E}_{S_n}[\rho(\hat{y}(\boldsymbol{x};\mu^*) - g(\boldsymbol{U}\boldsymbol{x}) - \xi)] \\
&\le \mathbb{E}_{S_n}[\rho(\xi)] + C_\rho\mathbb{E}_{S_n}[|\hat{y}(\boldsymbol{x};\mu^*) - g(\boldsymbol{U}\boldsymbol{x})|] \\
&\le \mathbb{E}_{S_n}[\rho(\xi)] + C_kC_\rho\frac{\tilde{r}_x}{r_x}\left(\frac{r_x\Delta}{\tilde{r}_x}\right)^{-\frac{2}{k+1}} \ln\left(\frac{r_x\Delta}{\tilde{r}_x}\right) + \frac{C_\rho\ln 4}{\kappa}.
\end{aligned}$$

Furthermore, Proposition 13 guarantees $\mathcal{R}(\mu^*) \le \Delta^2/\tilde{r}_x^2$. Combining these bounds with Lemma 14 completes the proof. ■

### B.3. Generalization Analysis

Let

$$\mu_\beta^* \coloneqq \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^{2d+2})} \mathcal{F}_\beta(\mu) = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^{2d+2})} \tilde{\mathcal{F}}_\beta(\mu).$$

Corollary 15 gives an upper bound on $\hat{\mathcal{J}}_0(\mu^*)$. In this section, we transfer the bound to $\mathcal{J}_0(\mu^*)$ via a Rademacher complexity analysis. Since Corollary 15 implies a bound on $\mathcal{H}(\mu \,|\, \gamma)$, we will control the following quantity,

$$\sup_{\mu:\mathcal{H}(\mu\,|\,\gamma)\le\Delta^2} \mathcal{J}_0(\mu) - \hat{\mathcal{J}}_0(\mu).$$

To be able to provide guarantees with high probability, we will prove uniform convergence over a truncated version of the risk instead, given by

$$\sup_{\mu:\mathcal{H}(\mu\,|\,\gamma)\le\Delta^2} \mathcal{J}_0^\varkappa(\mu) - \hat{\mathcal{J}}_0^\varkappa(\mu),$$

where

$$\mathcal{J}_0^{\varkappa}(\mu) \coloneqq \mathbb{E}[\rho_{\varkappa}(\hat{y}(\boldsymbol{x};\mu) - y)], \quad \hat{\mathcal{J}}_0^{\varkappa}(\mu) \coloneqq \mathbb{E}_{S_n}[\rho_{\varkappa}(\hat{y}(\boldsymbol{x};\mu) - y)],$$

and $\rho_{\varkappa}(\cdot) \coloneqq \rho(\cdot) \wedge \varkappa$. We will later specify the choice of $\varkappa$.

We are now ready to present the Rademacher complexity bound.

**Lemma 16 ([14, Lemma 5.5], [45, Lemma 1])** *Suppose $\rho$ is either a $C_\rho$-Lipschitz loss or the squared error loss. Let $\vartheta \coloneqq \sqrt{2\varkappa}$ for the squared error loss and $C_\rho$ for the Lipschitz loss. Recall $\gamma = \mathcal{N}(0, \frac{\mathbf{I}_{d+1}}{\lambda\beta})$. Then,*

$$\mathbb{E}\left[\sup_{\{\mu\in\mathcal{P}^{\mathrm{ac}}(\mathbb{R}^{2d+2}):\mathcal{H}(\mu\,|\,\gamma)\le M\}} \mathcal{J}_0^{\varkappa}(\mu) - \hat{\mathcal{J}}_0^{\varkappa}(\mu)\right] \le 4\vartheta\iota\sqrt{\frac{2M}{n}}.$$

**Proof** We repeat the proof here for the reader's convenience. Let $(\xi_i)_{i=1}^n$ denote i.i.d. Rademacher random variables. Notice that for the squared error loss, $\rho_{\varkappa}$ is $\sqrt{2\varkappa}$ Lipschitz. Then, by a standard symmetrization argument and Talagrand's contraction lemma, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\mu:\mathcal{H}(\mu\,|\,\gamma)\le M} \mathcal{J}_0(\mu) - \hat{\mathcal{J}}_0(\mu)\right] &\le 2\,\mathbb{E}\left[\sup_{\mu:H(\mu\,|\,\gamma)\le M} \frac{1}{n}\sum_{i=1}^n \xi_i \rho(\hat{y}(\boldsymbol{x}^{(i)};\mu) - y)\right] \\ &\le 2\vartheta\,\mathbb{E}\left[\sup_{\mu:H(\mu\,|\,\gamma)\le M} \frac{1}{n}\sum_{i=1}^n \xi_i \hat{y}(\boldsymbol{x}^{(i)};\mu)\right]\end{aligned}$$

Next, we proceed to bound the Rademacher complexity. Specifically,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\mu:H(\mu\,|\,\gamma)\le M} \frac{1}{n}\sum_{i=1}^n \xi_i \int \Psi(\boldsymbol{x}^{(i)};\boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})\right] &= \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{1}{\alpha}\sup_{\mu:H(\mu\,|\,\gamma)\le M} \int \frac{\alpha}{n}\sum_{i=1}^n \xi_i \Psi(\boldsymbol{x}^{(i)};\boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})\right] \\ &\le \frac{M}{\alpha} + \frac{1}{\alpha}\,\mathbb{E}_{\boldsymbol{\xi}}\left[\ln \int \exp\left(\frac{\alpha}{n}\sum_{i=1}^n \xi_i \Psi(\boldsymbol{x}^{(i)};\boldsymbol{w})\right)\mathrm{d}\gamma(\boldsymbol{w})\right] \\ &\le \frac{M}{\alpha} + \frac{1}{\alpha}\ln \int \mathbb{E}_{\xi}\left[\exp\left(\frac{\alpha}{n}\sum_{i=1}^n \xi_i \Psi(\boldsymbol{x}^{(i)};\boldsymbol{w})\right)\right]\mathrm{d}\gamma(\boldsymbol{w}),\end{aligned}$$

where the first inequality follows from the KL divergence lower bound of Lemma 11. Additionally, by sub-Gaussianity and independence of $(\xi_i)$ and Lipschitzness of $\phi_{\kappa,\iota}$, we have

$$\begin{aligned}\mathbb{E}_{\xi}\left[\exp\left(\frac{\alpha}{n}\sum_{i=1}^n \xi_i \Psi(\boldsymbol{x}^{(i)};\boldsymbol{w})\right)\right] &\le \exp\left(\frac{\alpha^2}{2n^2}\sum_{i=1}^n \Psi(\boldsymbol{x}^{(i)};\boldsymbol{w})^2\right) \\ &\le \exp\left(\frac{2\alpha^2\iota^2}{n}\right)\end{aligned}$$

Plugging this back into our original bound, we obtain

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\mu:H(\mu\,|\,\gamma)\le M} \frac{1}{n}\sum_{i=1}^n \xi_i \hat{y}(\boldsymbol{x};\mu)\right] \le \frac{M}{\alpha} + \frac{2\alpha\iota^2}{n}.$$

Choosing $\alpha = \sqrt{\frac{Mn}{2\iota^2}}$, we obtain

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{\mu:\mathcal{H}(\mu\,|\,\gamma)\le M}\frac{1}{n}\sum_{i=1}^{n}\xi_i\hat{y}(\boldsymbol{x};\mu)\right] \le 2\iota\sqrt{\frac{2M}{n}},$$

which completes the proof. ■

We can convert the above bound in expectation to a high-probability bound as follows.

**Lemma 17** *In the setting of Lemma 16, for any $\delta > 0$, we have*

$$\sup_{\mu\in\mathcal{P}^{\mathrm{ac}}(\mathbb{R}^{2d+2}):\mathcal{H}(\mu\,|\,\gamma)\le M}\mathcal{J}_0^{\varkappa}(\mu) - \hat{\mathcal{J}}_0^{\varkappa}(\mu) \lesssim \vartheta\iota\sqrt{\frac{M}{n}} + \varkappa\sqrt{\frac{\ln(1/\delta)}{n}},$$

*with probability at least $1-\delta$.*

**Proof** As the truncated loss is bounded by $\varkappa$, the result is an immediate consequence of McDiarmid's inequality. ■

Next, we control the effect of truncation by bounding $\mathcal{J}_0(\mu)$ via $\mathcal{J}_0^{\varkappa}(\mu)$, which is achieved via the following lemma.

**Lemma 18** *Suppose $\mathcal{H}(\mu\,|\,\gamma) \le M$. Then,*

$$\mathcal{J}_0(\mu) - \mathcal{J}_0^{\varkappa}(\mu) \lesssim \Big(\iota + \mathbb{E}\big[y^2\big]^{1/2}\Big)e^{-\Omega(\varkappa^2)}.$$

**Proof** Notice that since the loss is $C_\rho$-Lipschitz and $\rho(0) = 0$, we have $|\rho(\hat{y}-y)| \le C_\rho|\hat{y}-y|$. Recall that we use $L$ for the Lipschitz constant of $g$, and $|\hat{y}(\boldsymbol{x};\mu)| \le 2\iota$. Then,

$$\begin{aligned}
\mathcal{J}_0(\mu) - \mathcal{J}_0^{\varkappa}(\mu) &\le \mathbb{E}[\mathbb{1}(\rho(\hat{y}(\boldsymbol{x};\mu)-y)\ge\varkappa)\rho(\hat{y}(\boldsymbol{x};\mu)-y)]\\
&\le C_\rho\mathbb{P}(\rho(\hat{y}(\boldsymbol{x};\mu)-y)\ge\varkappa)^{1/2}\,\mathbb{E}\big[(\hat{y}(\boldsymbol{x};\mu)-y)^2\big]^{1/2}\\
&\le C_\rho\mathbb{P}(2\iota+|y|\ge\varkappa/C_\rho)^{1/2}\Big(\mathbb{E}\big[\hat{y}(\boldsymbol{x};\mu)^2\big]^{1/2}+\mathbb{E}\big[y^2\big]^{1/2}\Big)\\
&\le C_\rho\mathbb{P}(2\iota+|g(0)|+L\|\boldsymbol{U}\boldsymbol{x}\|+\xi\ge\varkappa/C_\rho)^{1/2}\Big(\mathbb{E}\big[\hat{y}(\boldsymbol{x};\mu)^2\big]^{1/2}+\mathbb{E}\big[y^2\big]^{1/2}\Big).
\end{aligned}$$

Let $\varkappa/C_\rho \ge 4\iota + 2|g(0)| + 2Lr_x$, and recall that $L = \mathcal{O}(1/r_x)$. Then, by a subGaussian concentration bound, we have

$$\mathbb{P}(2\iota+|g(0)|+L\|\boldsymbol{U}\boldsymbol{x}\|+\xi\ge\varkappa/C_\rho)^{1/2} \le e^{-\Omega\left(\frac{\varkappa^2}{\sigma_u^2C_\rho^2}\right)}.$$

We conclude the proof by remarking that by our assumptions, $\sigma_u$ and $C_\rho$ are absolute constants. ■

Finally, we combine the steps above to give an upper bound on $\mathcal{J}_0(\mu_\beta^*)$, stated in the following lemma.

**Lemma 19** *Suppose* $\lambda = \tilde{\lambda} r_x^2$ *and* $\beta = \frac{d_{\text{eff}} + \tilde{r}_x^2/r_x^2}{\varepsilon^2 \tilde{\lambda}}$ *for* $\varepsilon, \tilde{\lambda} \lesssim 1$. *Let* $\tilde{\varepsilon} \coloneqq \tilde{\mathcal{O}}(\tilde{\lambda}^{\frac{1}{k+2}}) + \varepsilon + \kappa^{-1}$. *Suppose* $n \gtrsim \frac{(d_{\text{eff}} + \tilde{r}_x^2/r_x^2)\iota^2}{\tilde{\lambda}\varepsilon^4}$ *and* $\iota \gtrsim \tilde{\lambda}^{-\frac{k}{k+2}} (\tilde{r}_x/r_x)^{\frac{2k^2+4k+2}{k+2}}$. *Then,*

$$\mathcal{J}_0(\mu_\beta^*) - \mathbb{E}[\rho(\xi)] \lesssim \tilde{\varepsilon}, \quad \textit{and} \quad \beta^{-1}\mathcal{H}(\mu_\beta^* \,|\, \gamma) \lesssim \mathbb{E}[\rho(\xi)] + \tilde{\varepsilon} \lesssim 1.$$

**Proof** By Corollary 15 and a standard concentration bound on $\mathbb{E}_{S_n}[\rho(\xi)]$ with sufficiently large $n$ to induce neglibile error in comaprison with the rest of the terms in the corollary, we have

$$\hat{\mathcal{J}}_0(\mu_\beta^*) + \beta^{-1}\mathcal{H}(\mu_\beta^* \,|\, \gamma) - \mathbb{E}[\rho(\xi)] \lesssim \frac{\tilde{r}_x}{r_x}\left(\frac{r_x\Delta}{\tilde{r}_x}\right)^{\frac{-2}{k+1}} \ln\left(\frac{r_x\Delta}{\tilde{r}_x}\right) + \frac{\lambda\Delta^2}{\tilde{r}_x^2} + \frac{(c_x + \tilde{r}_x)}{\sqrt{\lambda\beta}} + \frac{1}{\kappa}.$$

By choosing

$$\Delta = \left(\frac{r_x^2}{\lambda}\right)^{\frac{1}{2}\cdot\frac{k+1}{k+2}} \left(\frac{\tilde{r}_x}{r_x}\right)^{\frac{1}{2}\cdot\frac{3k+5}{k+2}},$$

and assuming $c_x \gtrsim \tilde{r}_x$,

$$\beta^{-1}\mathcal{H}(\mu_\beta^* \,|\, \gamma) \lesssim \mathbb{E}[\rho(\xi)] + \left(\frac{\lambda}{r_x^2}\right)^{\frac{1}{k+2}} \left(\frac{\tilde{r}_x}{r_x}\right)^{\frac{k+1}{k+2}} \ln\left(\frac{\tilde{r}_x r_x}{\lambda}\right) + \frac{c_x}{\sqrt{\lambda\beta}} + \frac{1}{\kappa}.$$

Note that the above choice on $\Delta$ translates to a lower bound on $\iota$ in Corollary 15, given by

$$\iota \gtrsim$$

By choosing $\lambda = \tilde{\lambda} r_x^2$ and using the fact that $\tilde{r}_x \leq \tilde{\mathcal{O}}(r_x)$ and $\beta = \frac{c_x^2}{r_x^2 \tilde{\lambda}\varepsilon^2}$, we have the simpification,

$$\beta^{-1}\mathcal{H}(\mu_\beta^* \,|\, \gamma) \lesssim \mathbb{E}[\rho(\xi)] + \tilde{\mathcal{O}}(\tilde{\lambda}^{\frac{1}{k+2}}) + \varepsilon + \frac{1}{\kappa} \lesssim 1,$$

and,

$$\hat{\mathcal{J}}_0(\mu_\beta^*) - \mathbb{E}[\rho(\xi)] \lesssim \tilde{\mathcal{O}}(\tilde{\lambda}^{\frac{1}{k+2}}) + \varepsilon + \frac{1}{\kappa} =: \tilde{\varepsilon}.$$

Note that $\hat{\mathcal{J}}_0^{\varkappa}(\mu_\beta^*) \leq \hat{\mathcal{J}}_0(\mu_\beta^*)$. Using the generalization bound of Lemma 17 with the choice of $\delta = n^{-q}$ for some constant $q > 0$, we have with probability $1 - \mathcal{O}(n^{-q})$,

$$\begin{aligned} \mathcal{J}_0^{\varkappa}(\mu_\beta^*) - \hat{\mathcal{J}}_0^{\varkappa}(\mu_\beta^*) &\lesssim \iota\sqrt{\frac{\beta}{n}} + \varkappa\sqrt{\frac{\ln n}{n}} \\ &\lesssim \iota\sqrt{\frac{d_{\text{eff}}}{n\tilde{\lambda}\varepsilon^2}} + \varkappa\sqrt{\frac{\ln n}{n}}. \end{aligned} \tag{B.5}$$

Furthermore, by Lemma 18 we have

$$\mathcal{J}_0(\mu_\beta^*) - \mathcal{J}_0^{\varkappa}(\mu_\beta^*) \lesssim \iota e^{-\Omega(\varkappa^2)}.$$

Combining the above with (B.5) and choosing on $\varkappa \asymp \sqrt{\ln n}$, we have

$$\mathcal{J}_0(\mu_\beta^*) - \mathbb{E}[\rho(\xi)] \lesssim \tilde{\varepsilon} + \iota\sqrt{\frac{d_{\text{eff}}}{n\tilde{\lambda}\varepsilon^2}} + \sqrt{\frac{\ln^2 n}{n}},$$

which holds with probability at least $1 - \mathcal{O}(n^{-q})$ over the randomness of $S_n$. ■

### B.4. Convergence Analysis

So far, our analysis has only proved properties of $\mu_\beta^*$. In this section, we relate these properties to $\mu_l^m$ via propagation of chaos. In particular, [44] showed that for $\boldsymbol{W} \sim \mu_l^m$, $\hat{y}(\boldsymbol{x}; \mu_l^m)$ converges to $\hat{y}(\boldsymbol{x}; \mu_\beta^*)$ in a suitable sense characterized shortly, as long as the objective over $\mu_l^m$ converges to $\mathcal{F}_{\beta,\lambda}(\mu_\beta^*)$. Notice that $\mu_\ell^m$ is a measure on $\mathcal{P}(\mathbb{R}^{(2d+2)m})$ instead of $\mathcal{P}(\mathbb{R}^{2d+2})$. Thus, we need to adjust the definition of objective by defining the following

$$\mathcal{F}_{\beta,\lambda}^m(\mu^m) \coloneqq \mathbb{E}_{\boldsymbol{W}\sim\mu^m}\left[\hat{J}_0(\boldsymbol{W}) + \frac{\lambda}{2}R(\boldsymbol{W})\right] + \frac{1}{m\beta}\mathcal{H}(\mu^m). \tag{B.6}$$

We can use the same reformulation introduced earlier in (B.1) to define

$$\tilde{\mathcal{F}}_{\beta,\lambda}^m(\mu^m) \coloneqq \mathbb{E}_{\boldsymbol{W}\sim\mu^m}\left[\hat{J}_0(\boldsymbol{W})\right] + \frac{1}{m\beta}\mathcal{H}(\mu^m \,|\, \gamma^{\otimes m}), \tag{B.7}$$

which is equivalent to $\mathcal{F}_{\beta,\lambda}^m$ up to an additive constant. With these definitions, we can now control $\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\mu_l^m)]$ via $\mathcal{J}_0(\mu_\beta^*)$. The following lemma is based on [44, Lemma 4], with a more careful analysis to obtain sharper constants.

**Lemma 20** *Let $\bar{r}_x \coloneqq \|\boldsymbol{\Sigma}\|^{1/2} \vee \tilde{r}_x$, and suppose $\rho$ is $C_\rho \lesssim 1$-Lipschitz. Then,*

$$\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\boldsymbol{W})] - \mathcal{J}_0(\mu_\beta^*) \lesssim \sqrt{\frac{\bar{r}_x^2 W_2^2\big(\mu_l^m, \mu_\beta^{*\otimes m}\big) + \iota^2}{m}}. \tag{B.8}$$

*In particular, combined with [44, Lemma 3], the above implies*

$$\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\boldsymbol{W})] - \mathcal{J}_0(\mu_\beta^*) \lesssim \sqrt{\frac{\bar{r}_x^2\beta C_{\mathrm{LSI}}}{m}\big(\tilde{\mathcal{F}}_{\beta,\lambda}^m(\mu_l^m) - \tilde{F}_{\beta,\lambda}(\mu_\beta^*)\big) + \frac{\iota^2}{m}}. \tag{B.9}$$

**Proof** Notice that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\boldsymbol{W})] &= \mathbb{E}_{\boldsymbol{W}}\left[\mathbb{E}_{\boldsymbol{x}}\left[\rho(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu_\beta^*) + \hat{y}(\boldsymbol{x};\mu_\beta^*) - y)\right]\right] \\
&\le \mathbb{E}_{\boldsymbol{x}}\left[\rho(\hat{y}(\boldsymbol{x};\mu_\beta^*) - y)\right] + C_\rho\,\mathbb{E}_{\boldsymbol{W}}\left[\mathbb{E}_{\boldsymbol{x}}\left[\left|\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu_\beta^*)\right|\right]\right] \\
&\le \mathcal{J}_0(\mu_\beta^*) + C_\rho\sqrt{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{W}}\left[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu_\beta^*))^2\right]\right]}
\end{aligned}$$

Suppose $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_m) \sim \mu_l^m$ and $\boldsymbol{W}' = (\boldsymbol{w}_1', \dots, \boldsymbol{w}_m') \sim \mu_\beta^{*\otimes m}$. Let $\Gamma$ denote the optimal $W_2$ coupling between $\boldsymbol{W}$ and $\boldsymbol{W}'$, and assume $\boldsymbol{W}, \boldsymbol{W}' \sim \Gamma$. Then,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{W}}\left[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu^*))^2\right] &= \mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\left[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'}) + \hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'}) - \hat{y}(\boldsymbol{x};\mu_\beta^*))^2\right] \\
&\le 2\,\mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\left[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'})^2\right] + 2\,\mathbb{E}_{\boldsymbol{W}'}\left[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'}) - \hat{y}(\boldsymbol{x};\mu_\beta^*))^2\right]
\end{aligned}$$

Moreover, by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\left[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}}) - \hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'})^2\right] &\le \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\left[(\Psi(\boldsymbol{x};\boldsymbol{w}_i) - \Psi(\boldsymbol{x};\boldsymbol{w}_i'))^2\right] \\
&\le \frac{2}{m}\sum_{i=1}^m \mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\left[\left\langle \boldsymbol{\omega}_{i1} - \boldsymbol{\omega}_{i1}', \tilde{\boldsymbol{x}}\right\rangle^2\right] + \frac{2}{m}\sum_{i=1}^m \mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\left[\left\langle \boldsymbol{\omega}_{i2} - \boldsymbol{\omega}_{i2}', \tilde{\boldsymbol{x}}\right\rangle^2\right].
\end{aligned}$$

Hence,

$$\mathbb{E}_{\boldsymbol{x}}\big[\mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\big[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}})-\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'}))^2\big]\big] \le \frac{2\left\|\tilde{\boldsymbol{\Sigma}}\right\|}{m}\mathbb{E}_{\boldsymbol{W},\boldsymbol{W}'}\Big[\big\|\boldsymbol{W}-\boldsymbol{W}'\big\|_{\mathrm{F}}^2\Big] = \frac{2\left\|\tilde{\boldsymbol{\Sigma}}\right\|}{m}W_2^2\big(\mu_t^m,\mu_\beta^{*\otimes m}\big).$$

For the second term, notice that $\hat{y}(\boldsymbol{x};\mu_\beta^*)=\mathbb{E}_{\boldsymbol{W}'}[\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'})]=\mathbb{E}_{\boldsymbol{w}_i'}[\Psi(\boldsymbol{x};\boldsymbol{w}_i')]$ for all $1\le i\le m$. By independence of $(\boldsymbol{w}_i')$ and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{W}'}\big[(\hat{y}(\boldsymbol{x};\mu_{\boldsymbol{W}'})-\hat{y}(\boldsymbol{x};\mu_\beta^*))^2\big] &= \frac{1}{m}\mathbb{E}_{\boldsymbol{w}'}\big[(\Psi(\boldsymbol{x};\boldsymbol{w}')-\hat{y}(\boldsymbol{x};\mu^*))^2\big]\\ &= \frac{1}{m}\mathbb{E}_{\boldsymbol{w}'}\left[\left(\int(\Psi(\boldsymbol{x};\boldsymbol{w}')-\Psi(\boldsymbol{x};\boldsymbol{w}))\mathrm{d}\mu_\beta^*(\boldsymbol{w})\right)^2\right]\\ &\lesssim \frac{\iota^2}{m}.\end{aligned}$$

■

Thus, the rest of this section deals with establishing convergence rates for $\mathcal{F}_{\beta,\lambda}^m(\mu_l^m)\to\mathcal{F}_{\beta,\lambda}(\mu_\beta^*)$. To use the one-step decay of optimality gap provided by [44], we depend on the following assumption.

**Assumption 3** *Suppose there exist constants $L$, $C_L$, and $R$, such that*

1. **(Lipschitz gradients of the Gibbs potential)** *For all $\mu,\mu'\in\mathcal{P}_2(\mathbb{R}^{2d+2})$ and $\boldsymbol{w},\boldsymbol{w}'\in\mathbb{R}^{2d+2}$,*

$$\left\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w})-\nabla\hat{\mathcal{J}}_0'[\mu'](\boldsymbol{w}')\right\|\le L(W_2(\mu,\mu')+\big\|\boldsymbol{w}-\boldsymbol{w}'\big\|), \tag{B.10}$$

   *where $W_2$ is the 2-Wasserstein distance.*

2. **(Bounded gradients of the Gibbs potential)** *For all $\mu\in\mathcal{P}_2(\mathbb{R}^{2d+2})$ and $\boldsymbol{w}\in\mathbb{R}^{2d+2}$, we have $\left\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w})\right\|\le R$.*

3. **(Bounded second variation)** *Denote the second variation of $\hat{\mathcal{J}}_0(\mu)$ at $\boldsymbol{w}$ via $\hat{\mathcal{J}}_0''[\mu](\boldsymbol{w},\boldsymbol{w}')$, which is defined as the first variation of $\mu\mapsto\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w})$. Then, for all $\mu\in\mathcal{P}_2(\mathbb{R}^{2d+2})$ and $\boldsymbol{w},\boldsymbol{w}'\in\mathbb{R}^{2d+2}$,*

$$\left|\hat{\mathcal{J}}_0''[\mu](\boldsymbol{w},\boldsymbol{w}')\right|\le L(1+C_L(\|\boldsymbol{w}\|^2+\big\|\boldsymbol{w}'\big\|^2)). \tag{B.11}$$

We can now state the one-step bound.

**Theorem 21 ([44, Theorem 2])** *Suppose $\hat{\mathcal{J}}_0$ satisfies Assumption 3. Assume $\lambda\lesssim 1$, $\beta,L,R\gtrsim 1$, and the initialization satisfies $\mathbb{E}\left[\left\|\boldsymbol{w}_0^i\right\|^2\right]\lesssim R^2$ for all $1\le i\le m$. Then, for all $\eta\le 1/4$,*

$$\mathcal{F}_{\beta,\lambda}^m(\mu_{l+1}^m)-\mathcal{F}_{\beta,\lambda}(\mu_\beta^*)\le\exp\big(\frac{-\eta}{2\beta C_{\mathrm{LSI}}}\big)\big(\mathcal{F}_{\beta,\lambda}^m(\mu_l^m)-\mathcal{F}_{\beta,\lambda}(\mu_\beta^*)\big)+\eta A_{m,\beta,\lambda,\eta}, \tag{B.12}$$

*where*

$$A_{m,\beta,\lambda,\eta} \coloneqq C\left(L^2\big(d+\frac{R^2}{\lambda}\big)\big(\eta^2+\frac{\eta}{\beta}\big)+\frac{L}{m\beta}\Big(\frac{1}{C_{\mathrm{LSI}}}+\big(\frac{R^2}{\lambda^2}+\frac{d}{\lambda\beta}\big)\big(\frac{C_L}{C_{\mathrm{LSI}}}+\frac{L}{\beta}\big)\Big)\right) \tag{B.13}$$

*for some absolute constant $C > 0$.*

We now focus on bounding the constants that appear in Assumption 3.

**Lemma 22 (Lipschitzness of $\nabla\hat{\mathcal{J}}_0'$)** *Suppose $\rho$ is either the squared error loss or is $C_\rho$ Lipschitz and has a $C_\rho'$ Lipschitz derivative. Assume $\kappa \gtrsim 1$. Notice that for the squared error loss, $C_\rho' = 1$. Then, for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^{2d+2})$ and $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^{2d+2}$, we have*

$$\Big\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) - \hat{J}_n'[\mu'](\boldsymbol{w}')\Big\| \lesssim \kappa C_\rho\Big\|\mathbb{E}_{S_n}\Big[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|\|\boldsymbol{w}-\boldsymbol{w}'\| + C_\rho'\Big\|\mathbb{E}_{S_n}\Big[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|W_2(\mu,\mu'),$$

*for the Lipschitz loss, and*

$$\Big\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) - \hat{J}_n'[\mu'](\boldsymbol{w}')\Big\| \lesssim \kappa\sqrt{\hat{\mathcal{J}}_0(\mu)\big\|\mathbb{E}_{S_n}\big[\tilde{\boldsymbol{x}}^{\otimes 4}\big]\big\|_{2\to 2}}\|\boldsymbol{w}-\boldsymbol{w}'\| + \Big\|\mathbb{E}_{S_n}\Big[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|W_2(\mu,\mu'),$$

*for the squared error loss, where $\big\|\mathbb{E}_{S_n}\big[\tilde{\boldsymbol{x}}^{\otimes 4}\big]\big\|_{2\to 2} \coloneqq \sup_{\|\boldsymbol{v}\|\le 1}\Big\|\mathbb{E}_{S_n}\Big[\langle\tilde{\boldsymbol{x}},\boldsymbol{v}\rangle^2\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|$.*

**Proof** Recall that $\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) = \mathbb{E}_{S_n}[\rho'(\hat{y}(\boldsymbol{x};\mu)-y)\Psi(\boldsymbol{x};\boldsymbol{w})]$, where $\Psi(\boldsymbol{x};\boldsymbol{w}) = \phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle) - \phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_2,\tilde{\boldsymbol{x}}\rangle)$. We start with the triangle inequality,

$$\Big\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) - \nabla\hat{\mathcal{J}}_0'[\mu'](\boldsymbol{w}')\Big\| \le \Big\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) - \nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}')\Big\| + \Big\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}') - \nabla\hat{\mathcal{J}}_0'[\mu'](\boldsymbol{w}')\Big\|.$$

We now focus on the first term. For the Lipschitz loss,

$$\begin{aligned}
\Big\|\nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) - \nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}')\Big\| &= \big\|\mathbb{E}_{S_n}\big[\rho'(\hat{y}(\boldsymbol{x};\mu)-y)(\phi_\kappa'(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle) - \phi_\kappa'(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle)\tilde{\boldsymbol{x}}\big]\big\| \\
&\le C_\rho\mathbb{E}_{S_n}\big[(\phi_\kappa'(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle) - \phi_\kappa'(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle))^2\big]^{1/2}\Big\|\mathbb{E}_{S_n}\Big[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|^{1/2} \\
&\le C_\rho\kappa\mathbb{E}_{S_n}\Big[\langle\boldsymbol{\omega}_1-\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle^2\Big]^{1/2}\Big\|\mathbb{E}_{S_n}\Big[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|^{1/2} \\
&\le C_\rho\kappa\Big\|\mathbb{E}_{S_n}\Big[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|\|\boldsymbol{\omega}_1-\boldsymbol{\omega}_1'\|,
\end{aligned}$$

where the first inequality follows from Lemma 12, and the second inequality follows from the fact that $|\phi_\kappa''| \le \kappa$. For the squared error loss, we have

$$\begin{aligned}
\Big\|\nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}) - \nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w}')\Big\| &= \big\|\mathbb{E}_{S_n}\big[(\hat{y}(\boldsymbol{x};\mu)-y)(\phi_\kappa'(\langle\boldsymbol{w},\tilde{\boldsymbol{x}}\rangle) - \phi_\kappa'(\langle\boldsymbol{w}',\tilde{\boldsymbol{x}}\rangle)\tilde{\boldsymbol{x}}\big]\big\| \\
&= \sup_{\|\boldsymbol{v}\|\le 1}\mathbb{E}_{S_n}\big[(\hat{y}(\boldsymbol{x};\mu)-y)(\phi_\kappa'(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle) - \phi_\kappa'(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle)\langle\boldsymbol{v},\tilde{\boldsymbol{x}}\rangle\big] \\
&\le \sup_{\|\boldsymbol{v}\|\le 1}\sqrt{\mathbb{E}_{S_n}[(\hat{y}(\boldsymbol{x};\mu)-y)^2]\mathbb{E}_{S_n}\Big[(\phi_\kappa'(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle) - \phi_\kappa'(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle))^2\langle\boldsymbol{v},\tilde{\boldsymbol{x}}\rangle^2\Big]} \\
&\le \kappa\sqrt{\hat{\mathcal{J}}_0(\mu)\sup_{\|\boldsymbol{v}\|\le 1}\Big\langle\boldsymbol{v},\mathbb{E}_{S_n}\Big[\langle\boldsymbol{\omega}_1-\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle^2\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\boldsymbol{v}\Big\rangle} \\
&\le \kappa\sqrt{\hat{\mathcal{J}}_0(\mu)\Big\|\mathbb{E}_{S_n}\Big[\langle\boldsymbol{\omega}_1-\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle^2\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\Big]\Big\|} \\
&\le \kappa\sqrt{\hat{\mathcal{J}}_0(\mu)\big\|\mathbb{E}_{S_n}\big[\tilde{\boldsymbol{x}}^{\otimes 4}\big]\big\|_{2\to 2}}\|\boldsymbol{\omega}_1-\boldsymbol{\omega}_1'\|.
\end{aligned}$$

Similar bounds apply to the gradient with respect to $\boldsymbol{\omega}_2$, which completes the bound on the first term of the triangle inequality.

We now consider the second term of the triangle inequality. Here we consider Lipschitz losses and the squared error loss at the same time since both have a Lipschitz derivative.

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{\omega}')-\nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{\omega}')\right\| &= \left\|\left(\rho'(\hat{y}(\boldsymbol{x};\mu)-y)-\rho'(\hat{y}(\boldsymbol{x};\mu')-y)\right)\phi_\kappa'(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle)\tilde{\boldsymbol{x}}\right\| \\ &\le \mathbb{E}_{S_n}\left[\left(\rho'(\hat{y}(\boldsymbol{x};\mu)-y)-\rho'(\hat{y}(\boldsymbol{x};\mu')-y)\right)^2\right]^{1/2}\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|^{1/2} \\ &\le C_\rho'\mathbb{E}_{S_n}\left[(\hat{y}(\boldsymbol{x};\mu)-\hat{y}(\boldsymbol{x};\mu'))^2\right]^{1/2}\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|^{1/2},\end{aligned} \tag{B.14}$$

where the first inequality follows from Lemma 12. Let $\gamma\in\mathcal{P}_2(\mathbb{R}^{2d+2}\times\mathbb{R}^{2d+2})$ be a coupling of $\mu$ and $\mu'$ (i.e. the first and second marginals of $\gamma$ are equal to $\mu$ and $\mu'$ respectively). Recall that,

$$\hat{y}(\boldsymbol{x};\mu)-\hat{y}(\boldsymbol{x};\mu')=\int\left(\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle)-\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_2,\tilde{\boldsymbol{x}}\rangle)-\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle)+\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_2',\tilde{\boldsymbol{x}}\rangle)\right)\mathrm{d}\gamma(\boldsymbol{w},\boldsymbol{w}').$$

Therefore by the triangle inequality for the $L_2$ norm $\mathbb{E}_{S_n}\left[(\cdot)^2\right]^{1/2}$ and Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{S_n}\left[(\hat{y}(\boldsymbol{x};\mu)-\hat{y}(\boldsymbol{x};\mu'))^2\right]^{1/2} \le& \mathbb{E}_{S_n}\left[\int\left(\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle)-\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle)\right)^2\mathrm{d}\gamma\right]^{1/2} \\ &+\mathbb{E}_{S_n}\left[\int\left(\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_2,\tilde{\boldsymbol{x}}\rangle)-\phi_{\kappa,\iota}(\langle\boldsymbol{\omega}_2',\tilde{\boldsymbol{x}}\rangle)\right)^2\mathrm{d}\gamma\right]^{1/2} \\ \le& \int\mathbb{E}_{S_n}\left[\langle\boldsymbol{\omega}_1-\boldsymbol{\omega}_1',\tilde{\boldsymbol{x}}\rangle^2\right]^{1/2}\mathrm{d}\gamma+\int\mathbb{E}_{S_n}\left[\langle\boldsymbol{\omega}_2-\boldsymbol{\omega}_2',\tilde{\boldsymbol{x}}\rangle^2\right]^{1/2}\mathrm{d}\gamma \\ \le& \left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|^{1/2}\int\left(\|\boldsymbol{\omega}_1-\boldsymbol{\omega}_1'\|+\|\boldsymbol{\omega}_2-\boldsymbol{\omega}_2'\|\right)\mathrm{d}\gamma(\boldsymbol{w}_1,\boldsymbol{w}_2) \\ \le& \sqrt{2\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|\int\|\boldsymbol{w}-\boldsymbol{w}'\|^2\mathrm{d}\gamma(\boldsymbol{w},\boldsymbol{w}')}.\end{aligned}$$

By choosing $\gamma$ whose transport cost attains (or converges to) the optimal cost, we have

$$\mathbb{E}_{S_n}\left[(\hat{y}(\boldsymbol{x};\mu)-\hat{y}(\boldsymbol{x};\mu'))^2\right]^{1/2}\le\sqrt{2\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|}W_2(\mu,\mu').$$

Plugging the above result into (B.14), we have

$$\left\|\nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{\omega}')-\nabla_{\boldsymbol{\omega}_2}\hat{\mathcal{J}}_0'[\mu'](\boldsymbol{\omega}')\right\|\le\sqrt{2}C_\rho'\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|W_2(\mu,\mu').$$

Notice that the same bound holds for gradients with respect to $\boldsymbol{\omega}_2$. Thus the bound of the second term in the triangle inequality and the proof is complete. ■

**Lemma 23 (Boundedness of $\nabla\hat{\mathcal{J}}_0'$)** *In the same setting as Lemma 22, for all $\mu\in\mathcal{P}_2(\mathbb{R}^{2d+2})$ and $\boldsymbol{w}\in\mathbb{R}^{2d+2}$, we have*

$$\left\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w})\right\|\le\sqrt{2}\tilde{C}_\rho\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|^{1/2},$$

*where $\tilde{C}_\rho=C_\rho$ when $\rho$ is Lipschitz and $\tilde{C}_\rho=\sqrt{2\hat{\mathcal{J}}_0(\mu)}$ when $\rho$ is the squared error loss.*

**Proof** Notice that $|\phi_\kappa'| \leq 1$. Therefore,

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\omega}_1}\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w})\right\| &= \left\|\mathbb{E}_{S_n}\left[\rho'(\hat{y}-y)\phi_\kappa'(\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle)\tilde{\boldsymbol{x}}\right]\right\| \\ &\leq \sqrt{\mathbb{E}_{S_n}[\rho'(\hat{y}-y)^2]\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|} \\ &\leq \tilde{C}_\ell\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|^{1/2},\end{aligned}$$

where the first inequality follows from Lemma 12. ∎

**Lemma 24 (Boundedness of $\hat{\mathcal{J}}_0''$)** *In the same setting as Lemma 22, for all $\mu \in \mathcal{P}_2(\mathbb{R}^{2d+2})$ and $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^{2d+2}$, we have*

$$\left|\hat{\mathcal{J}}_0''[\mu](\boldsymbol{w},\boldsymbol{w}')\right| \leq C_\rho'\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|\left(\|\boldsymbol{w}\|^2 + \|\boldsymbol{w}'\|^2\right),$$

*where we recall that $C_\rho' = 1$ for the squared error loss.*

**Proof** It is straightforward to show that

$$\hat{\mathcal{J}}_0''[\mu](\boldsymbol{w},\boldsymbol{w}') = \mathbb{E}_{S_n}\left[\rho''(\hat{y}(\boldsymbol{x};\mu)-y)\Psi(\boldsymbol{x};\boldsymbol{w})\Psi(\boldsymbol{x};\boldsymbol{w}')\right].$$

Then, by the Cauchy-Schwartz inequality,

$$\hat{\mathcal{J}}_0''[\mu](\boldsymbol{w},\boldsymbol{w}') \leq C_\rho'\mathbb{E}_{S_n}\left[\Psi(\boldsymbol{x};\boldsymbol{w})^2\right]^{1/2}\mathbb{E}_{S_n}\left[\Psi(\boldsymbol{x};\boldsymbol{w}')^2\right]^{1/2}.$$

Moreover, by the Lipschitzness of $\phi_{\kappa,\iota}$,

$$\begin{aligned}\mathbb{E}_{S_n}\left[\Psi(\boldsymbol{x};\boldsymbol{w})^2\right]^{1/2} &\leq \mathbb{E}_{S_n}\left[\langle\boldsymbol{\omega}_1,\tilde{\boldsymbol{x}}\rangle^2\right]^{1/2} + \mathbb{E}_{S_n}\left[\langle\boldsymbol{\omega}_2,\tilde{\boldsymbol{x}}\rangle^2\right]^{1/2} \\ &\leq \sqrt{2}\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|^{1/2}\|\boldsymbol{w}\|\end{aligned}$$

We can similarly bound the expression for $\boldsymbol{w}'$, and arrive at the statement of the lemma via Young's inequality,

$$\hat{\mathcal{J}}_0''[\mu](w,w') \leq 2\left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|C_\rho'\|\boldsymbol{w}\|\|\boldsymbol{w}'\| \leq \left\|\mathbb{E}_{S_n}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top\right]\right\|\left(\|\boldsymbol{w}\|^2 + \|\boldsymbol{w}'\|^2\right).$$

∎

In particular, the gradient bound of Lemma 23 implies the following LSI estimate, which follows from [44, Theorem 1].

**Proposition 25 ([44, Theorem 1])** *Suppose $\left\|\nabla\hat{\mathcal{J}}_0'[\mu](\boldsymbol{w})\right\| \leq R$ for all $\mu \in \mathcal{P}_2$ and $\boldsymbol{w} \in \mathbb{R}^{2d+2}$. Then, the family of probability measures $\nu_\mu \propto \exp(-\beta G'[\mu])$ for $\mu \in \mathcal{P}_2(\mathbb{R}^{2d+2})$ satisfy a uniform LSI with constant*

$$C_{\mathrm{LSI}} \lesssim \frac{1}{\beta\lambda}\exp\left(\frac{4\beta R^2\sqrt{2d/\pi}}{\lambda}\right) \wedge \left\{\frac{1}{\beta\lambda} + \exp\left(\frac{\beta R^2}{2\lambda}\right)\left(\frac{R^2}{\lambda^2} + \frac{1}{\beta\lambda}\right)\left(d + \frac{\beta R^2}{\lambda}\right)\right\}. \quad \text{(B.15)}$$

We collect the smoothness estimates and simplify them under the event of Lemma 10 in the following Corollary.

**Corollary 26** *Suppose $\rho$ and $\rho'$ are $C_\rho$ and $C'_\rho$ Lipschitz respectively, with $C_\rho, C'_\rho \lesssim 1$. Recall that $\mathbf{\Sigma} := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$. On the event of Lemma 10, we have $\|\mathbb{E}_{S_n}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^\top]\| \lesssim \|\mathbf{\Sigma}\| \vee \tilde{r}_x^2$, and consequently, $\hat{\mathcal{J}}'_0$ satisfies Assumption 3 with constants $L \lesssim \kappa(\|\mathbf{\Sigma}\| \vee \tilde{r}_x^2)$, $R \lesssim \|\mathbf{\Sigma}\|^{1/2} \vee \tilde{r}_x$, and $C_L = \kappa^{-1}$.*

Using the estimates above, we can present the following convergence bound $\mathcal{F}^m_{\beta,\lambda}(\mu^*_\beta) - \mathcal{F}_{\beta,\lambda}(\mu^*_\beta)$.

**Proposition 27** *Let $\bar{r}_x := \|\mathbf{\Sigma}\| \vee \tilde{r}_x$, and for simplcity assume $C_{\mathrm{LSI}} \geq \beta$. For any $\varepsilon \lesssim 1$, suppose the step size satisfies*

$$\eta \lesssim \frac{\varepsilon}{C_{\mathrm{LSI}}\kappa^2\bar{r}_x^4(d + \bar{r}_x^2/\lambda)},$$

*the width of the network satisfies,*

$$m \gtrsim \frac{\kappa\bar{r}_x^2\Big(1 + \big(\frac{\bar{r}_x^2}{\lambda^2} + \frac{d}{\lambda\beta}\big)\big(\frac{1}{\kappa} + \frac{\kappa\bar{r}_x^2 C_{\mathrm{LSI}}}{\beta}\big)\Big)}{\varepsilon},$$

*and the number of iterations satisfies*

$$l \gtrsim \frac{\beta C_{\mathrm{LSI}}}{\eta} \ln\big(\frac{\mathcal{F}^m_{\beta,\lambda}(\mu^m_0) - \mathcal{F}^*_\beta}{\varepsilon}\big).$$

*Then, we have $\mathcal{F}^m_{\beta,\lambda}(\mu^m_l) - \mathcal{F}_{\beta,\lambda}(\mu^*_\beta) \leq \varepsilon$.*

**Proof** Throughout the proof, we will assume the event of Lemma 10 holds. Let $\mathcal{F}^*_{\beta,\lambda} := \mathcal{F}_{\beta,\lambda}(\mu^*_\beta)$. Notice that by iterating the bound of Theorem 21, we have

$$\begin{aligned}
\mathcal{F}^m_{\beta,\lambda}(\mu^m_l) - \mathcal{F}^*_{\beta,\lambda} &\leq \exp\big(\frac{-l\eta}{2\beta C_{\mathrm{LSI}}}\big)(\mathcal{F}^m_{\beta,\lambda}(\mu^m_0) - \mathcal{F}^*_{\beta,\lambda}) + \frac{\eta A_{m,\beta,\lambda,\eta}}{1 - \exp\big(\frac{-\eta}{2\beta C_{\mathrm{LSI}}}\big)} \\
&\leq \exp\big(\frac{-l\eta}{2\beta C_{\mathrm{LSI}}}\big)(\mathcal{F}^m_{\beta,\lambda}(\mu^m_0) - \mathcal{F}^*_{\beta,\lambda}) + 4\beta C_{\mathrm{LSI}} A_{m,\beta,\lambda,\eta},
\end{aligned}$$

where the second inequality holds for $\eta \leq 2\beta C_{\mathrm{LSI}}$ since $1 - e^{-x} \geq x/2$ for $x \in [0,1]$. We now bound $A_{m,\beta,\lambda,\eta}$ so that the RHS of the above is less than $\mathcal{O}(\varepsilon)$ by choosing a sufficiently large $m$ and a sufficiently small $\eta$. Recall that given constants $L$ and $R$ from Assumption 3,

$$A_{m,\beta,\lambda,\eta} \asymp L^2\big(d + \frac{R^2}{\lambda}\big)\big(\eta^2 + \frac{\eta}{\beta}\big) + \frac{L}{m\beta}\Big(\frac{1}{C_{\mathrm{LSI}}} + \big(\frac{R^2}{\lambda^2} + \frac{d}{\lambda\beta}\big)\big(\frac{C_L}{C_{\mathrm{LSI}}} + \frac{L}{\beta}\big)\Big).$$

From Corollary 26, $L \asymp \kappa(\|\mathbf{\Sigma}\| \vee \tilde{r}_x^2)$, $R \asymp \|\mathbf{\Sigma}\|^{1/2} \vee \tilde{r}_x$, and $C_L = \kappa^{-1}$. To avoid notational clutter, let $\bar{r}_x^2 := \|\mathbf{\Sigma}\| \vee \tilde{r}_x^2$. Then, to control the terms containing $\eta$, it suffices to choose

$$\eta \lesssim \sqrt{\frac{\varepsilon}{\beta C_{\mathrm{LSI}}\kappa^2\bar{r}_x^4(d + \bar{r}_x^2/\lambda)}} \wedge \frac{\varepsilon}{C_{\mathrm{LSI}}\kappa^2\bar{r}_x^4(d + \bar{r}_x^2/\lambda)},$$

for which we can simply choose

$$\eta \lesssim \frac{\varepsilon}{C_{\mathrm{LSI}}\kappa^2\bar{r}_x^4(d + \bar{r}_x^2/\lambda)}.$$

Further, to control the term containing the number of particles $m$, we need

$$m \gtrsim \frac{\kappa \bar{r}_x^2 \Big(1 + \big(\frac{\bar{r}_x^2}{\lambda^2} + \frac{d}{\lambda\beta}\big)\big(\frac{1}{\kappa} + \frac{\kappa \bar{r}_x^2 C_{\mathrm{LSI}}}{\beta}\big)\Big)}{\varepsilon}.$$

To drive the suboptimality bound below $\varepsilon$, we also need to let the number of iterations $l$ satisfy

$$l \gtrsim \frac{\beta C_{\mathrm{LSI}}}{\eta} \ln\big(\frac{\mathcal{F}^m_{\beta,\lambda}(\mu_0^m) - \mathcal{F}^*_{\beta,\lambda}}{\varepsilon}\big).$$

With the above conditions, we can guarantee

$$\mathcal{F}^m_{\beta,\lambda}(\mu_l^m) - \mathcal{F}^*_{\beta,\lambda} \lesssim \varepsilon,$$

which finishes the proof. ■

Finally, we are ready to present the proof of Theorem. Specifically, we will prove the following theorem which is the more detailed version of Theorem 2. To do so, we introduce the following assumption, which can be verified by Proposition 25.

**Assumption 4** *Let $\boldsymbol{W}^l = (\boldsymbol{w}_1^l, \dots, \boldsymbol{w}_m^l)$ denote the trajectory of MFLA. Then, the probability measure $\nu_{\mu_{\boldsymbol{W}^l}} \propto \exp(-\beta \hat{\mathcal{J}}'_\lambda[\mu_{\boldsymbol{W}^l}])$ satisfies the LSI (A.6) with constant $C_{\mathrm{LSI}}$ for all $l \geq 0$. For simplicity, assume $C_{\mathrm{LSI}} \geq \beta$.*

With the above assumption, we can state the detailed version of Theorem 2.

**Theorem 28** *Under Assumptions 1, 2 and 4, consider MFLA (2.3) with parameters $\beta = \tilde{\Theta}(d_{\mathrm{eff}})$, $\lambda = \tilde{\lambda} r_x^2$ and $\eta \leq \tilde{\mathcal{O}}\big(\frac{1}{C_{\mathrm{LSI}}\kappa^2 d}\big)$. The algorithm is initialized with the weights sampled i.i.d. from some distribution $\boldsymbol{w}_j^0 \sim \mu_0$. Suppose $\tilde{\lambda}, \kappa^{-1} = o_n(1)$, and $\iota$ is chosen via the lower bound of Lemma 19. Then, with the number of samples $n$, the number of neurons $m$, and the number of iterations $l$ that can respectively be bounded by*

$$n \leq \tilde{\mathcal{O}}(d_{\mathrm{eff}}), \quad m \leq \tilde{\mathcal{O}}\big(\frac{C_{\mathrm{LSI}}\kappa^2 d}{\beta^2 \lambda}\big), \quad l \leq \tilde{\mathcal{O}}\big(\frac{C_{\mathrm{LSI}}\beta}{\eta}\big), \tag{B.16}$$

*with probability at least $1 - \mathcal{O}(n^{-q})$ for some $q > 0$, the excess risk satisfies*

$$\mathbb{E}_{\boldsymbol{W}^l}\, \mathbb{E}_{y,\boldsymbol{x}}[\rho(y - \hat{y}_m(\boldsymbol{x}; \boldsymbol{W}^l))] - \mathbb{E}_\xi[\rho(\xi)] \leq o_n(1). \tag{B.17}$$

*For the statement of Theorem 2, we can choose $\tilde{\lambda}^{-1}, \kappa = \mathrm{polylog}(n)$.*

### B.5. Proof of Theorem 2

Recall that $\lambda = \tilde{\lambda} r_x^2$, and let $\beta = \frac{d_{\mathrm{eff}} + \tilde{r}_x^2/r_x^2}{\varepsilon^2 \tilde{\lambda}}$ and $n \geq \frac{(d_{\mathrm{eff}} + \tilde{r}_x^2/r_x^2)\iota^2}{\tilde{\lambda}\varepsilon^4}$ for some $\varepsilon \lesssim 1$, where $\tilde{\varepsilon} := \tilde{\mathcal{O}}(\tilde{\lambda}^{\frac{1}{k+2}} + \varepsilon + \kappa^{-1})$. Then, from Lemma 19, we have $\mathcal{J}_0(\mu_\beta^*) - \mathbb{E}[\rho(\xi)] \lesssim \tilde{\varepsilon}$. On the other hand, given the step size $\eta$, width $m$, and number of iterations $l$ by Proposition 27, we have $\mathcal{F}^m_{\beta,\lambda}(\mu_l^m) - \mathcal{F}_{\beta,\lambda}(\mu_\beta^*) \leq \varepsilon$. Therefore,

$$\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\boldsymbol{W})] - \mathcal{J}_0(\mu_\beta^*) \lesssim \sqrt{\frac{\bar{r}_x^2 \beta C_{\mathrm{LSI}}\varepsilon}{m} + \frac{\iota^2}{m}}.$$

Additionally, from Lemma 19, we have

$$\beta^{-1}\mathcal{H}(\mu_\beta^* \,|\, \gamma) \lesssim \mathbb{E}[\rho(\xi)] + \tilde{\mathcal{O}}(\tilde{\lambda}^{\frac{1}{k+2}}) + \varepsilon + \kappa^{-1} \lesssim 1.$$

Consequently, for $m \geq \frac{\bar{r}_x^2(d_{\text{eff}}+\tilde{r}_x^2/r_x^2)C_{\text{LSI}}}{\tilde{\lambda}\varepsilon^3} \vee \frac{\iota}{\varepsilon^2}$, we have $\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\boldsymbol{W})] - \mathcal{J}_0(\mu_\beta^*) \leq \varepsilon$. Therefore, combining the bounds above, we have

$$\mathbb{E}_{\boldsymbol{W}\sim\mu_l^m}[J_0(\boldsymbol{W})] - \mathbb{E}[\rho(\xi)] \lesssim \tilde{\mathcal{O}}(\tilde{\lambda}^{\frac{1}{k+2}}) + \varepsilon + \kappa^{-1}.$$

Consequently, we can take $\tilde{\lambda} = o_n(1)$, $\varepsilon = o_n(1)$, $\kappa^{-1} = o_n(1)$ to finish the proof.

## Appendix C. Comparisons with the Formulation of [39]

Here, we provide a number of comparisons with results of [39]. In Section C.1, we show that the statistical model (2.1) is more general than their formulation, even for parity learning problems. In Section C.2, we provide an informal comparison of their effective dimension to our setting, exhibiting the improvement in our definition of effective dimension.

### C.1. Generality of the Formulations

We begin by pointing out that the formulation of $k$-index model of (2.1) is strictly more general than that of [39], even for learning $k$-sparse parities. Recall that in their setting, they consider inputs of the type $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}$ for some positive definite $\boldsymbol{\Sigma}$, where $\boldsymbol{z} \sim \text{Unif}(\{\pm 1\}^d)$ (their original formulation uses $\boldsymbol{z} \sim \text{Unif}(\{\pm 1/\sqrt{d}\}^d)$, but we rescale the input to be consistent with the notation of this paper). The labels are given by

$$y = \text{sign}\Big(\prod_{i=1}^k \langle \tilde{\boldsymbol{u}}_i, \boldsymbol{z}\rangle\Big) = \text{sign}\Big(\prod_{i=1}^k \Big\langle \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_i, \boldsymbol{x}\Big\rangle\Big), \tag{C.1}$$

where $\{\tilde{\boldsymbol{u}}_i\}_{i=1}^k$ are orthonormal vectors. Then, we can define an orthonormal set of vectors $\{\boldsymbol{u}_i\}_{i=1}^k$ such that $\text{span}(\boldsymbol{u}_1, \dots, \boldsymbol{u}_k) = \text{span}(\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_1, \dots, \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_k)$, and define $g$ such that

$$g\Big(\frac{\langle \boldsymbol{u}_1, \boldsymbol{x}\rangle}{\sqrt{k}}, \dots, \frac{\langle \boldsymbol{u}_k, \boldsymbol{x}\rangle}{\sqrt{k}}\Big) = g\left(\frac{\left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_1, \boldsymbol{z}\right\rangle}{\sqrt{k}}, \dots, \frac{\left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_k, \boldsymbol{z}\right\rangle}{\sqrt{k}}\right) = \text{sign}\Big(\prod_{i=1}^k \langle \tilde{\boldsymbol{u}}_i, \boldsymbol{z}\rangle\Big),$$

for all $\boldsymbol{z} \in \{\pm 1\}^d$. Therefore, the parity formulation of (C.1) can be seen as a special case of the $k$-index model (2.1). Note that $g$ is only defined on $2^d$ points, and we can extend it to all of $\mathbb{R}^k$ such that $g : \mathbb{R}^k \to \mathbb{R}$ is Lipschitz.

In contrast, the $k$-index model can represent parity problems that cannot be represented by (C.1). Starting from an orthonormal set of vectors $\{\boldsymbol{u}_i\}_{i=1}^k$ in $\mathbb{R}^d$, let

$$y = g\Big(\frac{\langle \boldsymbol{u}_1, \boldsymbol{x}\rangle}{\sqrt{k}}, \dots, \frac{\langle \boldsymbol{u}_k, \boldsymbol{x}\rangle}{\sqrt{k}}\Big) = \text{sign}\Big(\prod_{i=1}^k \langle \boldsymbol{u}_i, \boldsymbol{x}\rangle\Big). \tag{C.2}$$

Consider the case where $k = 2$, then $y = \text{sign}\left(\left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_1, \boldsymbol{z}\right\rangle\left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_2, \boldsymbol{z}\right\rangle\right)$. To be able to reformulate this to (C.1), we need to be find orthonromal $\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2 \in \mathbb{R}^d$ such that

$$\text{sign}\left(\left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_1, \boldsymbol{z}\right\rangle\left\langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_2, \boldsymbol{z}\right\rangle\right) = \text{sign}(\langle \tilde{\boldsymbol{u}}_1, \boldsymbol{z}\rangle\langle \tilde{\boldsymbol{u}}_2, \boldsymbol{z}\rangle), \quad \forall z \in \{\pm 1\}^d.$$

If $\boldsymbol{\Sigma}$ has rank less than $d$ such that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_1 = \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_2$, then the above implies $\text{sign}(\langle \tilde{\boldsymbol{u}}_1, \boldsymbol{z}\rangle\langle \tilde{\boldsymbol{u}}_2, \boldsymbol{z}\rangle) \geq 0$ for all $z \in \{\pm 1\}^d$. In particular, we must have some $\boldsymbol{z}$ where $\text{sign}(\langle \tilde{\boldsymbol{u}}_1, \boldsymbol{z}\rangle\langle \tilde{\boldsymbol{u}}_2, \boldsymbol{z}\rangle) > 0$, which implies that

$$\sum_{i=1}^{2^d} \langle \tilde{\boldsymbol{u}}_1, \boldsymbol{z}_i\rangle\langle \tilde{\boldsymbol{u}}_2, \boldsymbol{z}_i\rangle = \left\langle \tilde{\boldsymbol{u}}_1, \sum_{i=1}^{2^d} \boldsymbol{z}_i\boldsymbol{z}_i^\top \tilde{\boldsymbol{u}}_2 \right\rangle = 2^d \langle \tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\rangle > 0,$$

which is in contradiction with $\langle \tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\rangle = 0$. Therefore, for such $\boldsymbol{\Sigma}$, we cannot formulate (C.2) as a special case of (C.1). This argument is robust with respect to small perturbations of $\boldsymbol{\Sigma}$ which make it full-rank, implying that (C.2) is strictly more general than (C.1), even when only considering full-rank covariance matrices.

### C.2. Comparison with the Effective Dimension of [39]

A close inspection of the proofs in [39] demonstrates that one can define their effective dimension in a scale invariant manner as $\tilde{d}_{\text{eff}} := \text{tr}(\boldsymbol{\Sigma})\left\|\sum_{i=1}^k \boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_i\right\|^2$. From the previous section, we observed that to reduce their setting to ours, we need to choose a set $\{\boldsymbol{u}_i\}_{i=1}^k$ of normalized vectors that spans the set of vectors $\{\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_i\}_{i=1}^k$. In particular, we can choose $\boldsymbol{u}_i = \frac{\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_i}{\|\boldsymbol{\Sigma}^{-1/2}\tilde{\boldsymbol{u}}_i\|}$, or equivalently write $\tilde{\boldsymbol{u}}_i = \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i\|}$. While $\{\boldsymbol{u}_i\}_{i=1}^k$ are not orthogonal, our proofs do not rely on the orthogonality assumption and it is only made for simplicity. Hence, we have

$$\tilde{d}_{\text{eff}} = \text{tr}(\boldsymbol{\Sigma})\left\|\sum_{i=1}^k \frac{\boldsymbol{u}_i}{\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i\right\|}\right\|^2 \leq k\,\text{tr}(\boldsymbol{\Sigma})\sum_{i=1}^k \left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i\right\|^{-2}.$$

Note that the above upper bound is sharp when $k = 1$, and is lower bounded by our definition of effective dimension stated in Definition 1.