A Convex Relaxation Approach to Generalization Analysis for Parallel Positively Homogeneous Networks

Uday Kiran Reddy Tadipatri
University of Pennsylvania

Benjamin D. Haeffele
University of Pennsylvania

Joshua Agterberg
University of Illinois Urbana-Champaign

René Vidal
University of Pennsylvania

Abstract

We propose a general framework for deriving generalization bounds for parallel positively homogeneous neural networks–a class of neural networks whose input-output map decomposes as the sum of positively homogeneous maps. Examples of such networks include matrix factorization and sensing, single-layer multi-head attention mechanisms, tensor factorization, deep linear and ReLU networks, and more. Our general framework is based on linking the non-convex empirical risk minimization (ERM) problem to a closely related convex optimization problem over prediction functions, which provides a global, achievable lower-bound to the ERM problem. We exploit this convex lower-bound to perform generalization analysis in the convex space while controlling the discrepancy between the convex model and its non-convex counterpart. We apply our general framework to a wide variety of models ranging from low-rank matrix sensing, to structured matrix sensing, two-layer linear networks, two-layer ReLU networks, and single-layer multi-head attention mechanisms, achieving generalization bounds with a sample complexity that scales almost linearly with the network width.

1 INTRODUCTION

Despite significant recent advances in the analysis of deep neural networks (DNNs), key gaps persist in establishing guaranteed performance of such models–particularly regarding theoretical guarantees on unseen data. This lack of performance guarantees is especially concerning for high-stakes applications such as autonomous vehicles, healthcare, or other high-consequence decision-making systems. To ensure the safe and reliable deployment of deep learning models, it is essential that generalization guarantees be established under reasonable data-generating mechanisms.

Related work. There is a broad literature on generalization theory. Classical approaches can be categorized along two separate (but related) lines:(i) data-dependent versus data-independent bounds, and (ii) uniform versus non-uniform concentration. Informally, *data-dependent* bounds take into account explicit data-generating assumptions, whereas data-independent bounds hold *regardless* of the underlying data distribution. Similarly, *uniform* concentration guarantees focus on obtaining concentration inequalities *simultaneously* for all functions in some function class (known as the *hypothesis space*). In contrast, non-uniform concentration inequalities focus on particular functions estimated from the data. Classical approaches marry these two separate types of analyses by introducing measures such as the VC-dimension (Vapnik and Chervonenkis, 1968) or the Rademacher Complexity (Bartlett and Mendelson, 2001). However, these classical measures are often difficult to compute and overly pessimistic, especially when applied to DNNs (Zhang et al., 2021). Consequently, many classical approaches may fail in the modern, more complex DNN setting.

Modern generalization frameworks for DNNs acknowledge that data often comes from structured distributions (e.g., with an intrinsic dimensionality significantly below that of the ambient space) and that optimization algorithms like Stochastic Gradient Descent (SGD) explores only a small portion of the hypothe-

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

sis space (Neyshabur et al., 2017). As a result, the effective hypothesis space is much smaller than what classical bounds account for based on the expressivity of the model alone. Consequently, modern bounds focus on data-dependent, non-uniform approaches. For instance, *margin bounds* (Neyshabur et al., 2018; Golowich et al., 2018; Barron and Klusowski, 2019) provide specific generalization error bounds for DNNs trained to minimize max-margin type loss functions for classification tasks. Another line of research (Dziugaite and Roy, 2017; Arora et al., 2018; Banerjee et al., 2020) exploits the sensitivity of the non-convex landscapes around learned weights; however, this approach requires the estimation of hard quantities like expected sharpness and KL divergence and questions remain regarding the extent to which quantities such as sharpness explain network generalization (Wen et al., 2023; Andriushchenko et al., 2023).

Recent work has observed that optimization methods such as SGD, even without explicit regularization, tend to yield solutions that generalize well, a notion known as *implicit bias* (Gunasekar et al., 2017, 2018a,b; Soudry et al., 2018; Li et al., 2020; HaoChen et al., 2021; Vardi, 2023). This stands in contrast to classical theory, which suggests that explicit regularization is necessary to avoid overfitting. For example, DNNs have been shown to converge toward maximum-margin solutions in classification tasks (Soudry et al., 2018), while solutions in regression tasks often exhibit low-rank structures (Li et al., 2020) that generalize well. Although these analyses provide valuable insights, they are generally limited to specific objectives and types of neural network architectures.

A key challenge in understanding the generalization properties of DNNs is their non-convex landscape. Indeed, convex landscapes are better understood, and numerous generalization bounds have already been derived (Shalev-Shwartz et al., 2009; Lugosi and Neu, 2022). We argue that bridging the gap between non-convex and convex landscapes could provide a pathway to understanding generalization better. Our key contribution is to propose a new generalization analysis framework for DNNs based on linking their non-convex landscape to a convex one. Our framework builds upon Haeffele and Vidal (2017) and Vidal et al. (2022), who connected certain non-convex optimization problems to closely related convex ones. However, their work focuses on characterizing the optimization properties of such problems and does not consider generalization.

Paper contributions. In this work, we use the idea of analyzing non-convex problems via a closely related convex problem to derive generalization bounds for a broad family of learning models, which take the form of sums of (*slightly generalized*) positively homogeneous functions whose parameters are regularized by sums of positively homogeneous functions of the same degree. This allows for a reinterpretation of the (empirical and expected) non-convex optimization problems as closely related to carefully constructed convex problems. We then apply concentration of measure techniques to the convexified version under reasonable data distributions and show that this also implies the concentration of the non-convex problem of interest. More specifically, we extend the finite-dimensional framework of Haeffele and Vidal (2017) and Vidal et al. (2022) to its infinite-dimensional counterpart, which allows us to derive generalization guarantees from a novel viewpoint by exploiting the connection between our problem of interest and a closely related convex problem. We note that other prior work (Bach, 2017) has also considered similar relationships between convex and non-convex problems for establishing generalization results. However, the generalization guarantees in Bach (2017) largely rely on Rademacher complexities, which results in a sample complexity that grows quadratically with the network width. In contrast, we exploit the relationship between the convex and non-convex problems more directly, which allows us to derive bounds with an improved sample complexity.

To be more precise, our main results can be stated informally as follows. Let N be the number of data points, R be the number of positively homogeneous functions (or the width of the network) whose predictions are summed together to form the output, and $\mathsf{dim}(\mathcal{W})$ be the dimension of the parameters in one of the functions. When $N \gtrsim \tilde{\mathcal{O}}(R \times \mathsf{dim}(\mathcal{W}))$, we show that the generalization error can be bounded with high probability by two terms: the first term, dubbed the optimization error, which vanishes at a globally optimal solution, and the second term, dubbed the statistical error, which depends on the ratio $\frac{R \times \mathsf{dim}(\mathcal{W})}{N}$, and hence vanishes only asymptotically.

Our results apply to a wide range of signal processing and DNN problems. The derived bounds achieve near state-of-the-art sample complexity for non-convex low-rank matrix sensing that match the lower bound provided by Candès and Plan (2011) for convex low-rank matrix sensing. By applying these general results to two-layer linear (and ReLU) neural networks with weight decay and multi-head attention models, a key component of transformer architecture (Vaswani et al., 2017), we obtain novel generalization bounds with "tight"[1] sample complexities for both problems.

[1] Our notion of "tight" bounds corresponds to cases where the sample complexity scales linearly or nearly linearly (up to logarithmic factors) with the number of model parameters.

Outline. The remainder of this paper is organized as follows. In §2, we formulate the learning problem and introduce our approach. In §3, we explore how learning problems can be bounded via convex surrogates. In §4, we present the statistical bounds through the master theorem that provides generalization error bounds. In §5, we apply the master theorem to various problems in signal processing and DNNs and compare our derived sample complexities with those in the existing literature. The supplementary material contains detailed proofs of the mathematical statements, validations of our framework's assumptions through simulations, and an additional survey of related works.

Notation. For two random variables (Z, W), drawn from a joint distribution q, we define $\langle Z, W\rangle_q = \mathbb{E}[\langle Z, W\rangle]$, where the expectation is with respect to the joint probability distribution q. For a generic function $f : \mathbb{R}^d \to \mathbb{R}$, we denote $\|f\|_{Lip}$ as its Lipschitz constant; i.e., the smallest number L_f such that $|f(x) - f(y)| \leq L_f\|x - y\|$. A function f is said to be integrable with respect to measure q, i.e., $f \in L^2(q)$, if $\left(\int_{x\in\mathcal{X}} \|f(x)\|^2 dq(x)\right)^{1/2} < \infty$. The inequality $f(x) \gtrsim g(x)$, means that there exists a constant $c > 0$ such that $f(x) \geq cg(x)$. We define the ReLU function as $[x]_+ = \max(x, 0)$. For the matrix U the variable $\mathbf{u}_j$ corresponds to jth column of U.

2 PROBLEM FORMULATION

Given a realization of a pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ from a distribution μ with $\mathcal{X} \subset \mathbb{R}^{n_X}$, $\mathcal{Y} \subset \mathbb{R}^{n_Y}$, we consider a (non)parametric regression problem of the form $Y = g(X, \epsilon)$, where ϵ is a source of additional noise (typically independent from X). We are interested in approximating g by the sum of r prediction functions, $\phi : \mathcal{W} \times \mathbb{R}^{n_X} \to \mathbb{R}^{n_Y}$, parameterized by $W \in \mathcal{W}$, i.e.,

$$\hat{Y} = \sum_{j=1}^{r} \phi(W_j)(X) = \Phi_r(\{W_j\})(X). \tag{1}$$

We will additionally refer to $\phi(W)(X)$ as the *factor map/sub-network* depending on the specific problem.

Our goal is to learn the parameters $\{W_j\}$[2] that minimize the regularized population risk defined as

$$\mathsf{NC}_\mu(\{W_j\}) := \underbrace{\mathbb{E}_{(X,Y)}\left[\ell(Y, \Phi_r(\{W_j\}_{j=1}^r)(X))\right]}_{=:\ell(g,\Phi_r(\{W_j\}_{j=1}^r))_\mu} + \lambda\Theta_r(\{W_j\}_{j=1}^r), \tag{2}$$

where $Y = g(X, \epsilon)$ is the target random variable, $\ell(\cdot, \cdot)$ is the *loss function*, typically convex in the second argument, and $\Theta_r(\{W_j\}_{j=1}^r)$ is an *explicit regularization* function which helps find structured parameters, such as minimum norm or sparse solutions. Specifically, the regularization term $\Theta_r(\{W_j\}_{j=1}^r)$ is defined as

$$\Theta_r(\{W_j\}_{j=1}^r) := \sum_{j=1}^{r} \theta(W_j), \tag{3}$$

where $\theta : \mathcal{W} \to \mathbb{R}^+$ is a regularization term for each factor map, and $\lambda \in \mathbb{R}^+$ is a regularization hyperparameter that controls the trade-off between loss reduction and inducing structure.

Notice that we will minimize the population risk $\mathsf{NC}_\mu(\{W_j\})$ over both r and $\{W_j\}_{j=1}^r$. More explicitly, we will allow for problems where, in addition to optimizing over the model parameters, one also optimizes over the number of prediction functions r (e.g., the network width) during training. However, our results will also apply to a value of r that is fixed *a priori*.

Estimating $\mathsf{NC}_\mu(\{W_j\})$ directly is challenging due to (i) the lack of access to the distribution μ, (ii) the fact that $(\{W_j\})$ (and potentially the number r) are random variables dependent on the training data $\{(X_i, Y_i)\}$, and (iii) the non-linearity and potential non-convexity of NC_μ. We address the first point (as is standard) via empirical minimization of $\mathsf{NC}_\mu(\cdot)$ using the *empirical risk* (or *training error*) defined via:

$$\mathsf{NC}_{\mu_N}(\{W_j\}) := \underbrace{\frac{1}{N}\sum_{i=1}^{N} \ell(Y_i, \Phi_r(\{W_j\}_{j=1}^r)(X_i))}_{=:\ell(g,\Phi_r(\{W_j\}_{j=1}^r))_{\mu_N}} + \lambda\Theta_r(\{W_j\}_{j=1}^r), \tag{4}$$

where μ_N denotes the empirical distribution of the samples $\{X_i, Y_i\}_{i=1}^N$. We define empirical risk minimization (ERM) via the arg min of $\mathsf{NC}_{\mu_N}(\{W_j\})$. For concreteness, recall we also allow for the minimization over r (provided r is bounded above by some quantity independent of the data), though our results hold for any fixed r.

Note that if we minimize the objective $\mathsf{NC}_{\mu_N}(\cdot)$, there is no guarantee that we will also minimize $\mathsf{NC}_\mu(\cdot)$. This discrepancy is quantified by the *Generalization Error:*

$$\begin{aligned} &|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})| \\ &= \left|\ell(g, \Phi_r(\{W_j\}_{j=1}^r))_\mu - \ell(g, \Phi_r(\{W_j\}_{j=1}^r))_{\mu_N}\right|. \end{aligned} \tag{5}$$

Note that the regularization terms containing Θ_r are the same between the two objectives, giving the typical difference between the empirical and population losses.

In this work, we compute an upper bound for the generalization error at any stationary point of the empirical problem, $\mathsf{NC}_{\mu_N}(\{W_j\})$, under certain technical assumptions. To build our main results, we relate these

[2] We occasionally notate $\{W_j\}_{i=1}^r$ as $\{W_j\}$ for brevity of notation, but the dependence on r is always implied.

non-convex objectives $\mathsf{NC}_\mu(\{W_j\})$ and $\mathsf{NC}_{\mu_N}(\{W_j\})$ to closely related *convex* objectives in the prediction space, respectively $\mathsf{C}_\mu(f_\mu)$ and $\mathsf{C}_{\mu_N}(f_{\mu_N})$, whose definitions will be introduced in §3. This allows us to decompose the generalization error in (5) as:

$$\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) = \underbrace{\Big[\mathsf{NC}_\mu(\{W_j\}) - \mathsf{C}_\mu(f_\mu)\Big]}_{\text{Population Gap}} - \underbrace{\Big[\mathsf{NC}_{\mu_N}(\{W_j\}) - \mathsf{C}_{\mu_N}(f_{\mu_N})\Big]}_{\text{Empirical Gap}} + \underbrace{\Big[\mathsf{C}_\mu(f_\mu) - \mathsf{C}_{\mu_N}(f_{\mu_N})\Big]}_{\text{Convex Generalization Gap}}. \tag{6}$$

Our Theorem 1 bounds the *Empirical Gap* and the *Population Gap*. With these bounds, we then apply concentration techniques to bound the *Convex Generalization Gap* and obtain our main Theorem 2 which gives bounds for the generalization error in (5).

3 CONVEX BOUNDS FOR LEARNING

In this section, we present bounds for the *Empirical Gap* and *Population Gap* through Theorem 1, linking our learning problem of interest to functions that are convex in the space of prediction functions. To begin, we state several requirements for our framework.

Assumption 1 (Regularization)**.** *The regularization function θ is positive semidefinite; i.e, $\theta(0) = 0$ and $\theta(W) \geq 0, \forall W \in \mathcal{W}$.*

This is a mild assumption; it only ensures we do not impose negative regularization on the parameters $\{W_j\}$. Our next assumption is our main functional assumption on ϕ and θ.

Assumption 2 (Balanced Homogeneity of ϕ and θ)**.** *The factor map ϕ and the regularization map θ can be scaled equally by non-negative scaling of (a subset of) the parameters. Formally, we assume that there exists sub-parameter spaces $(\mathcal{K}, \mathcal{H})$ from the parameter space $\mathcal{W}$ such that $\mathcal{K} \times \mathcal{H} = \mathcal{W}$, $\forall (\mathbf{k}, \mathbf{h}) \in (\mathcal{K}, \mathcal{H})$, and $\beta \geq 0$ we have $\phi((\beta\mathbf{k}, \mathbf{h})) = \beta^p \phi((\mathbf{k}, \mathbf{h}))$ and $\theta((\beta\mathbf{k}, \mathbf{h})) = \beta^p \theta((\mathbf{k}, \mathbf{h}))$ for some $p > 0$. Further, we assume that for bounded input X the set $\{\phi(W)(X) : \forall W \in \mathcal{W}$ s.t. $\theta(W) \leq 1\}$ is bounded.*

This is a slight generalization of positive homogeneity, which only requires positive homogeneity in a subset of parameters, provided the image of the factor map for parameters with $\theta(W) \leq 1$ is bounded [3].

[3]For example, $\phi(v)(X)$ can take the form $v^a g(X)$ for $a \geq 0$, and $\theta(v) = |v|^a$, where $g : \mathbb{R}^{n_x} \to \mathbb{R}$ is some fixed function. More generally, we can choose $\phi(v_1, v_2) = v_1^a g_{v_2}(X)$ and $\theta(v_1, v_2) = |v_1|^a + \delta_{\mathcal{V}_2}(v_2)$, where $g_{v_2} : \mathbb{R}^{n_x} \to \mathbb{R}$ is a function parameterized by $v_2 \in \mathcal{V}_2$ and has a bounded range for bounded inputs.

Our next assumption concerns the loss function ℓ.

Assumption 3 (Convex Loss)**.** *The loss $\ell(Y, \hat{Y})$ is second-order differentiable (written $\ell \in \mathcal{C}^2$), and L-smooth w.r.t. $\hat{Y}$, i.e, for any $Y, \hat{Y} \in \mathbb{R}^{n_Y}$*

$$0 \preceq \nabla^2_{\hat{Y}} \ell(Y, \hat{Y}) \preceq L I_{n_Y}. \tag{7}$$

Additionally, the gradient of the loss is bi-Lipschitz smooth; that is, for all $Y_1, Y_2, \hat{Y}_1, \hat{Y}_2 \in \mathbb{R}^{n_Y}$

$$\|\nabla_{\hat{Y}} \ell(Y_2, \hat{Y}_2) - \nabla_{\hat{Y}} \ell(Y_1, \hat{Y}_1)\| \leq L\Big[\|Y_2 - Y_1\|_2 + \|\hat{Y}_2 - \hat{Y}_1\|_2\Big], \tag{8}$$

and the loss is constant if both the arguments are the same, i.e., for all $Y_1, Y_2 \in \mathbb{R}^{n_Y}$, $\ell(Y_1, Y_1) = \ell(Y_2, Y_2)$.

This ensures that the loss function is convex and smooth. Furthermore, if the loss is α-strongly convex, i.e, $0 \prec \alpha I_{n_Y} \preceq \nabla^2_{\hat{Y}} \ell(Y, \hat{Y}) \preceq L I_{n_Y}$ we have derived tighter results (see the Appendix).

We define the *induced regularization* function as

$$\begin{aligned} \Omega(f) := \inf_{r, \{W_j\}} \ & \Theta_r(\{W_j\}) \\ \text{s.t. } & f(X) = \Phi_r(\{W_j\})(X); \quad \forall X \in \mathcal{X}, \end{aligned} \tag{9}$$

with the function taking value infinity if $f(X)$ cannot be realized for some choice of the parameters $(r, \{W_j\}_{j=1}^r)$. Using similar arguments as in Haeffele and Vidal (2015) it can be shown that under assumptions 1–2, the function $\Omega(f)$ is convex in the space of prediction functions; see Proposition 1 in the Appendix. Moreover, by Assumption 3, the loss function is convex with respect to the model predictions, which allows us to define the following two *convex* optimization problems over the space of prediction functions:

$$\mathsf{C}_\mu(f) := \mathbb{E}_{(X,Y)}[\ell(Y, f(X))] + \lambda \Omega(f), \tag{10}$$

where $f \in L^2(\mu)$, and

$$\mathsf{C}_{\mu_N}(f) := \frac{1}{N} \sum_{i=1}^N \ell(Y_i, f(X_i)) + \lambda \Omega(f), \tag{11}$$

where $f \in L^2(\mu_N)$.

From the definition of $\Omega(f)$ we have that C_μ and C_{μ_N} are always lower bounds of NC_μ and NC_{μ_N}, respectively, for any $(f, \{W_j\})$ such that $f(X) = \Phi_r(\{W_j\})(X)$, which becomes a tight bound for any parametrization $(\{W_j\})$ of f which achieves the infimum. As a result, we can relate solutions of the non-convex problems to the corresponding convex problem via tools from convex analysis, as we establish in the following result.

Theorem 1 (Convex Bounds for Learning)**.** *Under assumptions 1–3, let $f^*_{\mu_N}$ (or f^*_μ) be the global minimizer for $\mathsf{C}_{\mu_N}(\cdot)$ (or $\mathsf{C}_\mu(\cdot)$). For any stationary points $(r, \{W_j\})$ of the function $\mathsf{NC}_{\mu_N}(\cdot)$ and any $f \in L^2(\mu) \cap L^2(\mu_N)$ the following are true:*

1. *Empirical optimality gap:*

$$\mathsf{C}_{\mu_N}(f^*_{\mu_N}) \leq \mathsf{NC}_{\mu_N}(\{W_j\}) \leq \mathsf{C}_{\mu_N}(f) + \lambda\Omega(f)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right], \tag{12}$$

2. *Population optimality gap:*

$$\begin{aligned}\mathsf{C}_\mu(f^*_\mu) \leq \mathsf{NC}_\mu(\{W_j\}) \leq \mathsf{C}_\mu(f) &+ \lambda\Omega(f)\left[\Omega^\circ_\mu\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \\ &+ \Big[\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu \\ &- \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}\Big],\end{aligned} \tag{13}$$

where $\Omega^\circ_q(\cdot)$ is referred to as polar in the measure q defined as

$$\Omega^\circ_q(g) := \sup_{\theta(W)\leq 1} \langle g, \phi(W)\rangle_q. \tag{14}$$

Readers are referred to Appendix A for the proof with extensions to strongly convex functions.

The population optimality gap is obtained by an infinite-dimensional extension of Proposition 3 in Haeffele and Vidal (2020). The additional term in the population optimality gap (13) arises from the fact that the stationary points of ERM, $\mathsf{NC}_{\mu_N}(\cdot)$, are not necessarily the same as those of $\mathsf{NC}_\mu(\cdot)$.

From equation (5) the goal is to bound the difference between the original non-convex formulations NC_{μ_N} and NC_μ. By Theorem 1, we established the optimality gaps for both empirical and population non-convex optimization problems, and by computing the difference between equation (12) and (13), with algebraic manipulation we arrive at the following quantities:

- *Convex Generalization Gap*: The convex generalization gap is defined as $|C_\mu(f) - C_{\mu_N}(f)|$.
- *Polar Gap*: By virtue of the fact that the loss functions each contain the respective polars, we define the Polar Gap as the quantity $|\Omega^\circ_\mu(\nabla_{\hat{Y}}\ell(g,f)) - \Omega^\circ_{\mu_N}(\nabla_{\hat{Y}}\ell(g,f))|$.
- *Equilibria Gap*: We define the Equilibria Gap via $\left|\langle\nabla_{\hat{Y}}\ell(g,f), f\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell(g,f), f\rangle_\mu\right|$.
- *Norm Gap*: The final remaining quantity is defined via $\left|\|f^*_\mu - f\|^2_{\mu_N} - \|f^*_\mu - f\|^2_\mu\right|$. This quantity applies only to strongly convex functions (see the Appendix).

A major technical contribution of this paper is to demonstrate that each of these quantities uniformly concentrates at a rate equal to or smaller than the "statistical error" under certain realistic assumptions that are discussed in §4. The only remaining term from Theorem 1 is the quantity $\Omega(f_\mu)[\Omega^\circ_{\mu_N}(\cdot) - 1]$, which bounds the sub-optimality (in objective value) of the current stationary point for the empirical optimization problem. This term approaches zero at the global optimum of NC_{μ_N} (see §A in the Appendix).

4 STATISTICAL BOUNDS

In Theorem 1, we established bounds for the *Empirical Gap* and *Population Gap*. Building on these results, we identified key quantities such as the *Convex Generalization Gap*, *Polar Gap*, *Equilibrium Gap*, and *Norm Gap*, all of which can be controlled under certain general conditions (Assumptions 1–6, along with Assumption 7' from the Appendix) that we state momentarily. In this section, we present Theorem 2, which consolidates these bounds to derive our main generalization error bound. For clarity and to minimize technical complexity, we present Theorem 2 with Assumption 7, which a stronger version of Assumption 7'.

To begin, we state our additional assumptions. We assume that ϕ is Lipschitz.

Assumption 4 (Lipschitz Continuity of ϕ)**.** *Let $\mathcal{B}$ be some compact subset of $\mathcal{W}$, and denote*

$$\mathcal{F}_\theta := \{W : \theta(W) \leq 1\} \cap \mathcal{B} \subseteq \mathbb{B}(r_\theta), \tag{15}$$

where $\mathbb{B}(r_\theta)$ is the L_2 ball with radius r_θ.[4] *The factor map ϕ is Lipschitz continuous with respective to inputs for any choice of parameters $W \in \mathcal{F}_\theta$, i.e,*

$$L_\phi := \sup_{W\in\mathcal{F}_\theta} \|\phi(W)\|_{Lip} < \infty. \tag{16}$$

Our next assumption imposes tail conditions on the random variables (X, Y).

Assumption 5 (Data Model)**.** *The input data $X \in \mathbb{R}^{n_X}$ is drawn from the 1-Lipschitz concentrated sub-Gaussian distribution with a proxy variance σ^2_X/n_X;*

[4] The radius r_θ can depend on the dimension of W. For instance, suppose $W \in \mathbb{R}^n$ and $\theta(W) = \|W\|_1$, as $\|W\|_1 \leq \sqrt{n}\|W\|_2$, then r_θ must be at least $\sqrt{n}$. On another instance, suppose $W = (\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^n)$, and $\theta(W) = \|\mathbf{u}\|_2\|\mathbf{v}\|_2$; this requires r_θ to be at least $1/2$.

i.e., for any 1-Lipschitz continuous function, $h : \mathbb{R}^{n_X} \to \mathbb{R}$ there exists $c > 0$ such that

$$P\left(|h(X) - \mathbb{E}_X[h(X)]| \geq \epsilon\right) \leq c \exp\left(-\frac{n_X \epsilon^2}{2\sigma_X^2}\right). \quad (17)$$

The target function Y takes the form $Y = g(X, \epsilon)$, where $g \in L^2(\mu)$ is bi-Lipschitz in X and ϵ; that is,

$$\|g(X_2, \epsilon_2) - g(X_1, \epsilon_1)\|_2 \leq \|g\|_{Lip}\Big[\|X_2 - X_1\|_2 + \|\epsilon_2 - \epsilon_1\|_2\Big], \quad (18)$$

and $\epsilon \sim \mathcal{N}(0, (\sigma^2_{Y|X}/n_E)I)$ in $\mathbb{R}^{n_E}$.

We note that the above assumption is mild. While extending our framework to heavy-tailed distributions are likely possible; it would require a more intricate analysis and may result in worse error rates and larger sample complexities.

Our next assumption concerns the possible functions learned via empirical risk minimization.

Assumption 6 (Hypothesis class)**.** *Stationary points of $NC_{\mu_N}(\cdot)$ have bounded regularization and bounded width, $r \leq R$, almost surely. The input-output map, $\Phi_r(\{W_j\})$ has Lipschitz constant at most γ, and the parameters are bounded. Let $\mathcal{B}_R \subseteq \mathcal{W}^R$ be some compact set; then the hypothesis class is defined as*

$$\mathcal{F}_{\mathcal{W}} := \left\{\{W_j\}_{j=1}^r : \|\Phi_r(\{W_j\})\|_{Lip} \leq \gamma\right\} \cap \mathcal{B}_R. \quad (19)$$

In words, the set of maps learned through ERM are essentially Lipschitz in the parameters $\{W_j\}$, and, furthermore, the $\{W_j\}$ are bounded (almost surely). Moreover, the assumption that $r \leq R$ ensures that at most R individual functions $\{W_j\}$ are needed, which implicitly imposes a "low-complexity" constraint on the learned function. Finally, note that we assume that γ does not depend on the width of the network. In practice, our empirical observations show that the Lipschitz constant does not increase with width, making it a realistic assumption. For further details, refer to the numerical simulations in §E of Appendix.

Our general master theorem, Theorem 4 in the Appendix, requires only Assumptions 1–6 and 7' (in the Appendix). For the sake of notational brevity, we state our main results with the slightly stronger Assumption 7 instead of Assumption 7'.

Assumption 7 (Boundedness)**.** *For all $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, and $\{W_j\} \in \mathcal{F}_{\mathcal{W}}$, the predictions, and gradients are bounded; i.e,*

$$\|\Phi_r(\{W_j\})(X)\| \leq B_\Phi, \ \|\nabla_{\hat{Y}} \ell(Y, \Phi_r(\{W_j\})(X))\| \leq B_\ell. \quad (20)$$

Further, for any $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, for any $\{W_j\}, \{\tilde{W}'_j\} \in \mathcal{F}_{\mathcal{W}}$, $W, \tilde{W} \in \mathcal{F}_\theta$, the network, ϕ and Φ_r, are Lipschitz in the parameters; i.e,

$$\|\Phi_r(\{W_j\})(X) - \Phi_r(\{\tilde{W}_j\})(X)\|_2 \leq \tilde{L}_\Phi \max_j \|W_j - \tilde{W}_j\|_2, \ \text{and} \quad (21)$$

$$\|\phi(W)(X) - \phi(\tilde{W})(X)\|_2 \leq \tilde{L}_\phi \|W - \tilde{W}\|_2. \quad (22)$$

Assumption 7 ensures that predictions and its gradients are bounded while the network being Lipschitz continuous on the parameter space for any inputs. Assumption 7 implicitly indicates that either the data points are uniformly bounded or the search space for the parameters is of small dimension, which can restrict the potential applications. However, as we demonstrate in the more general version (Theorem 4) in the Appendix, it suffices that the conditions above hold only for some convex set $\mathcal{C}$, though this extension requires significantly more notation and discussion, so we do not include it here.

Theorem 2 (Master Theorem)**.** *Suppose Assumptions 1–7 hold. Let $\delta \in (0, 1]$ be fixed, and let f^*_μ be the global optimum of C_μ. Suppose that $\gamma \geq \Omega(f^*_\mu)\tilde{L}_\phi$, and define*

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|^2_{Lip}}{\gamma^2}\left(1 + \frac{\sigma^2_{Y|X}}{\sigma_X^2}\right)\right]\right\}; \quad (23)$$

$$\epsilon_2 = 4\tilde{L}_\Phi B_\Phi \max\Big\{1, 2L + 2B_\ell/B_\Phi, 8\Omega(f^*_\mu)(B_\ell \tilde{L}_\phi)/(\tilde{L}_\Phi B_\Phi), 8L\Omega(f^*_\mu)\Big\}. \quad (24)$$

Let $\{W_j\}$ denote any stationary point of $\mathsf{NC}_{\mu_N}(\cdot)$. Then with probability at least $1 - \delta$, it holds that

$$\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \lesssim \quad (25)$$

$$\underbrace{\frac{\lambda}{n_Y}\Omega(f^*_\mu)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)_{\mu_N}\right) - 1\right]}_{\textit{Optimization Error}}$$

$$+ \underbrace{\epsilon_1\sqrt{\frac{R \cdot \mathsf{dim}(\mathcal{W}) \log\left(\frac{\gamma\epsilon_2 r_\theta}{L_\phi}\right)\log(N) + \log\left(\frac{1}{\delta}\right)}{N}}}_{\textit{Statistical Error}}\Bigg\}.$$

Remarks: The generalization error is upper bounded by two terms:

- the *Optimization Error*, which quantifies the distance to the globally optimal solution, and
- the *Statistical Error*, or the intrinsic error that depends on the sample complexity and the noise.

The optimization error diminishes as we approach a global optimum of the ERM problem NC_{μ_N} and vanishes at a global optimum, whereas the statistical error diminishes as the sample size increases relative to

the intrinsic dimension, i.e., when $N \gtrsim R \times \mathsf{dim}(\mathcal{W})$ (ignoring logarithmic factors). By a naive counting argument, there are $R \times \mathsf{dim}(\mathcal{W})$ many parameters in the underlying network. As we will see in subsequent sections, this sample complexity turns out to be optimal or nearly optimal for a number of reasonable statistical settings. The implicit constants appearing in the result are universal and are not problem dependent.

5 APPLICATIONS

In this section, we present applications of the Theorem 2 for low-rank matrix sensing, two-layer ReLU neural networks, and single-layer multi-head attention. To apply Theorem 2, we must compute the problem-specific quantities $\Omega(f_\mu^*)$, $\Omega^\circ_{\mu_N}(\cdot)$, L, $\|g\|_{\text{Lip}}$, σ_X, $\sigma_{Y|X}$, ϵ_1, ϵ_2, r_θ, γ, L_ϕ. For each application, we have estimated these quantities, with further details provided in the proofs located in Appendix C.1, C.4, and C.5, respectively. We summarize and compare the obtained sample complexities for the various applications with their state-of-the-art bounds in Table 1. The additional applications to structured matrix sensing and two-layer linear neural networks can be found in Appendix C.2 and Appendix C.3, respectively.

Low-rank matrix sensing: We first consider low-rank matrix sensing (Candès and Plan, 2011), which is a well-studied problem in the signal processing and statistics literature. Given a few linear measurements of an unknown low-rank matrix, the goal is to estimate the low-rank matrix in the presence of noise. One potential strategy is to define a convex program via nuclear-norm regularization (Candès and Recht, 2009). While recovery guarantees for this convex program are well-studied, solving it is a computationally intensive procedure involving computing a full singular value decomposition at each iteration. To address this issue, several authors have considered a non-convex variant that reparameterizes the low-rank matrix into its underlying left and right factors, which is known as the Burer-Monteiro factorization (Burer and Monteiro, 2003). While the new optimization problem runs faster in practice, it is also non-convex, and its properties can be difficult to analyze theoretically. Corollary (1) provides the bounds on the generalization error for this non-convex program.

Corollary 1 (Low-Rank Matrix Sensing)**.** *Consider the true model for (X, y), where $X \in \mathbb{R}^{m\times n}$ is a random matrix with i.i.d. entries $X_{lk} \sim \mathcal{N}(0, \frac{1}{mn})$ and $y = \langle M^*, X\rangle + \epsilon$, where $M^* \in \mathbb{R}^{m\times n}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent from X. For all $i \in [N]$, let (X_i, y_i) be i.i.d. samples from this true model. Consider the estimator $\hat{y} = \langle UV^T, X\rangle$, where $U \in \mathbb{R}^{m\times R}$ and $V \in \mathbb{R}^{n\times R}$. Let $\delta \in (0, 1]$ be fixed. Define the non-convex problem*

$$\mathsf{NC}^{\mathsf{MS}}_{\mu_N}((U,V)) := \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \langle UV^T, X_i\rangle\right)^2 + \lambda\sum_{j=1}^{R}\|\mathbf{u}_j\|_2\|\mathbf{v}_j\|_2, \tag{26}$$

and define $\mathsf{NC}^{MS}_{\mu}((U,V))$ similarly with the sum over i replaced by expectation taken over (X, y).

Let $(\hat{U}, \hat{V})$ be a stationary point of $\mathsf{NC}^{\mathsf{MS}}_{\mu_N}(\cdot)$. Suppose there exists $C_{UV}, B_u, B_v > 0$ such that $\|\hat{U}\hat{V}^T\|_2 \le C_{UV}\|M^*\|_*$, and for all $j \in [R]$, $\|\hat{\mathbf{u}}_j\|_2 \le B_u$, $\|\hat{\mathbf{v}}_j\|_2 \le B_v$. Then with probability at least $1-\delta$, it holds that

$$\left|\mathsf{NC}^{\mathsf{MS}}_{\mu}((\hat{U},\hat{V})) - \mathsf{NC}^{\mathsf{MS}}_{\mu_N}((\hat{U},\hat{V}))\right| \lesssim \tag{27}$$
$$\|M^*\|_*\left[\|\frac{1}{N}\sum_{i=1}^{N}(y_i - \langle\hat{U}\hat{V}^T, X_i\rangle)X_i\|_2 - \lambda\right] + C^2_{UV}\|M^*\|^2_*\times \sqrt{\frac{R\log\left(R(C_{UV}+B_uB_v)\right)(m+n)\log(N)+\log(1/\delta)}{N}}.$$

Remarks: Observe that at a global minimum, the right side tends to zero when $R(m+n)/N \to 0$, ignoring logarithmic terms. Existing literature on non-convex noisy low-rank matrix sensing typically requires knowledge of true $\text{rank}(M^*) = R^*$, and the state-of-the-art sample complexity for this setting is of order $R^*(m+n)$ in the un-regularized setting (Stöger and Zhu, 2024). In contrast, Corollary 1 does not require knowledge of the true rank. However, if the estimated rank R is too small ($R < R^*$), then the optimization error still persists. In contrast, if ($R \ge R^*$) then optimization error can vanish subject to the ability of the algorithm utilized to reach stationary points, Haeffele and Vidal (2015) provides such guarantees.

Two-layer ReLU Networks: Next, we move on to two-layer ReLU networks, which introduce an additional nonlinearity with respect to the inputs. ReLU networks are widely used and proven to be universal approximators (Huang, 2020). Prior work on generalization analysis for ReLU networks is based on classical measures, such as Rademacher complexity (Bartlett et al., 2019). The following result circumvents the difficulty in the estimate of such classical measures.

Corollary 2 (Two-Layer ReLU Neural Network)**.** *Consider the true model for $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} \sim \mathcal{N}(0, (1/n)I_n) \in \mathbb{R}^n$, $\mathbf{y} = U^*[V^{*T}\mathbf{x}]_+ + \epsilon$, where $U^* \in \mathbb{R}^{m\times R^*}$, $V^* \in \mathbb{R}^{n\times R^*}$, and $\epsilon \sim \mathcal{N}(0, (\sigma^2/m)I_m) \in \mathbb{R}^m$ independent from $\mathbf{x}$. For all $i \in [N]$, let $(\mathbf{x}_i, \mathbf{y}_i)$ be i.i.d. samples from this true model. Consider the estimator $\hat{\mathbf{y}} = U[V^T\mathbf{x}]_+$, where $U \in \mathbb{R}^{m\times R}, V \in \mathbb{R}^{n\times R}$.*

Table 1: Comparisons with the state-of-the-art sample complexities. N represents the number of data points.

Application	Our work, $N \gtrsim$	State-of-the-art, $N \gtrsim$
Low rank matrix sensing	$\tilde{\mathcal{O}}(R(m+n))$	$R^*(m+n)$, (Stöger and Zhu, 2024) (no regularization)
Structured matrix sensing		–
2-Layer linear NN		$R(m+n)$ (Kakade et al., 2008) (bounded data-points)
2-Layer ReLU NN		$R(m+n)\log(R(m+n))$, (Bartlett et al., 2019)
Multi-head attention		$R(m+n)$, (Trauger and Tewari, 2024) (bounded data-points)

Let $\delta \in (0,1]$ be fixed. Define the non-convex problem

$$\mathsf{NC}_{\mu_N}^{\mathsf{ReLU}}((U,V)) := \frac{1}{2N}\sum_{i=1}^{N}\|\mathbf{y}_i - U[V^T\mathbf{x}_i]_+\|_2^2 + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right), \tag{28}$$

and define $\mathsf{NC}_{\mu}^{\mathsf{ReLU}}((U,V))$ similarly with the sum over i replaced by expectation taken over $(\mathbf{x},\mathbf{y})$.

Let $(\hat{U},\hat{V})$ be a stationary point of $\mathsf{NC}_{\mu_N}^{\mathsf{ReLU}}(\cdot)$. Suppose there exists $C_{UV}, B_u, B_v > 0$ such that $\|\hat{U}\hat{V}^T\|_2 \le C_{UV}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]$, and for all $j \in [R]$, $\|\hat{\mathbf{u}}_j\|_2 \le B_u$, $\|\hat{\mathbf{v}}_j\|_2 \le B_v$. Then with probability at least $1-\delta$, it holds that

$$\frac{1}{m}\left|\mathsf{NC}_{\mu}^{\mathsf{ReLU}}((\hat{U},\hat{V})) - \mathsf{NC}_{\mu_N}^{\mathsf{ReLU}}((\hat{U},\hat{V}))\right| \lesssim \tag{29}$$
$$\frac{1}{2m}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\left[\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2\|\mathbf{x}_i\|_2 - \lambda\right]$$
$$+ C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\times$$
$$\left[\frac{R(m+n)\log\left(R(m+n)(C_{UV}+B_u^2+B_v^2)\right)\log(N)}{N} + \frac{\log(1/\delta)}{N}\right]^{1/2}.$$

Remarks: Analogous to matrix sensing, when $R(m+n)/N \to 0$, the right side tends to zero at global optimality (ignoring logarithmic terms). Furthermore, Corollary 2 recovers the state-of-the-art result by Bartlett et al. (2019).

Transformers: Finally, we move on to our last application (though of course, the applications are in fact myriad in principle) to a single layer multi-head attention, which are backbones for transformer-style architecture (Vaswani et al., 2017). In practice, transformers are shown to have remarkable generalization capabilities (Zhou et al., 2024). However, there is a lack of intensive theoretical analysis for this architecture. Few attempts on estimating the capacities of the attention mechanisms have been made in Edelman et al. (2022) and Trauger and Tewari (2024), among others. For our analysis, we consider the case where the output of the model is one particular token within the input (e.g., transformers use a dedicated class token for the output initialized as a constant vector). The output for one attention head is modeled as $VX\sigma((KX)^\top Q\mathbf{x}_{out})$ where $\mathbf{x}_{out}$ is the column of X corresponding to the transformer output. We then reparameterize $K^TQ\mathbf{x}_{out} = \mathbf{z}$ and present the following result.

Corollary 3 (Transformers)**.** *Consider the true model for $(X,\mathbf{y})$, where $X \in \mathbb{R}^{n\times T}$ is a random matrix with i.i.d. entries $X_{lk} \sim \mathcal{N}(0,1/(nT))$ and $\mathbf{y} = A^*X\mathbf{b}^* + \epsilon$, where $A^* \in \mathbb{R}^{m\times n}$, $\mathbf{b}^* \in \mathbb{S}^{T-1}$ and $\epsilon \sim \mathcal{N}(0,(\sigma^2/m)I_m)$ is independent from X. For all $i \in [N]$, let $(X_i,\mathbf{y}_i)$ be i.i.d. samples from this true model. Consider the estimator $\hat{\mathbf{y}} = \sum_{j=1}^{R} V_jX\sigma(X^T\mathbf{z}_j)$, $V_j \in \mathbb{R}^n, \mathbf{z}_j \in \mathbb{R}^n$. Let $\delta \in (0,1]$ be fixed. Define the non-convex problem*

$$\mathsf{NC}_{\mu_N}^{\mathsf{TF}}(\{(V_j,\mathbf{z}_j)\}) := \frac{1}{2N}\sum_{i=1}^{N}\|\mathbf{y}_i - \sum_{j=1}^{R}V_jX_i\sigma_t(X_i^T\mathbf{z}_j)\|_2^2 + \lambda\sum_{j=1}^{R}\left[\|V_j\|_F + \delta_{\{\mathbf{z}:\|\mathbf{z}\|_2\le 1\}}(\mathbf{z}_j)\right], \tag{30}$$

where, $\sigma_t(\cdot)$ is softmax function with temperature t, for $k \in [T]$ defined $\sigma_t(\mathbf{u})_k := \exp(tu_k)/\sum_{l=1}^{T}\exp(tu_l)$ and define $\mathsf{NC}_{\mu}^{\mathsf{TF}}(\{(V_j,\mathbf{z}_j)\})$ similarly with the sum over i replaced by expectation taken over $(X,\mathbf{y})$.

Let $\{(\hat{V}_j,\hat{\mathbf{z}}_j)\}$ be a stationary point of $\mathsf{NC}_{\mu_N}^{\mathsf{TF}}(\cdot)$. Suppose there exists $C_V, B_V > 0$ such that $\sum_{j=1}^{R}\|\hat{V}_j\|_F \le C_V\|A^*\|_F$, and for all $j \in [R]$, $\|\hat{V}_j\|_F \le B_V$. Then with probability at least $1-\delta$ it holds that

$$\frac{1}{m}\left|\mathsf{NC}_{\mu}^{\mathsf{TF}}(\{(\hat{V}_j,\hat{\mathbf{z}}_j)\}) - \mathsf{NC}_{\mu_N}^{\mathsf{TF}}(\{(\hat{V}_j,\hat{\mathbf{z}}_j)\})\right| \lesssim \tag{31}$$
$$\frac{1}{2m}\|A^*\|_F\left[\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2\|X_i\|_2 - \lambda\right]$$
$$+ C_V^2\|A^*\|_F^2\times$$
$$\sqrt{\frac{R(m+n)\log(R(m+n)(C_V+B_V))\log(N)+\log(1/\delta)}{N}}.$$

Remarks: The dependence on $\mathbf{b}^*$ is not explicitly reflected in Equation (31) because the ground truth

model is bilinear. Consequently, assuming $\mathbf{b}^*$ is unit-norm without loss of generality, as its norm can be absorbed into A^*. Thus, the dependence on $\mathbf{b}^*$ is implicitly captured by the norm of A^* in Equation (31). As in the previous two applications, we can achieve consistency at global optimality when $N \gtrsim R(m+n)$, ignoring logarithmic terms. Note that the sample complexity has no dependency on the number of tokens, T, which suggests an explanation for the success behind the prediction capabilities of transformers for longer length inputs (Zhou et al., 2024). Our sample complexity matches the state-of-the-art bounds on the transformers by Trauger and Tewari (2024).

6 CONCLUSIONS

In this work, we provide generalization bounds for non-convex problems of the form of sums of (*slightly generalized*) positively homogeneous functions with a general objective. Our bounds provide sample complexities that are near-optimal and applicable to various problems, such as low-rank matrix sensing, two-layer neural networks, and single-layer multi-head attention. The sample complexity of our bounds grows almost linear with the total number of parameters in the model, and for matrix sensing, this sample complexity is optimal, as demonstrated in Candès and Plan (2011). Our proofs are based on analyzing closely related convex programs in the prediction space; this perspective enabled us to provide near-optimal sample complexities due to existing results on generalization properties for convex functions. In future work, it would be interesting to sharpen the dependence of our bounds on all the relevant parameters and apply our techniques to other machine learning problems.

Acknowledgments

UKRT gratefully acknowledges Pratik Chaudhari, Hancheng Min, Kyle Poe and Ziqing Xu for their valuable discussions and constructive feedback. His research was supported by the Leggett Family Fellowship and the Dean's Fellowship programs. Other authors acknowledge the support of the Research Collaboration on the Mathematical and Scientific Foundations of Deep Learning (NSF grant 2031985 and Simons Foundation grant 814201).

References

Allen-Zhu, Z., Li, Y., and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32.

Andriushchenko, M., Croce, F., Müller, M., Hein, M., and Flammarion, N. (2023). A modern look at the relationship between sharpness and generalization. *International Conference on Learning Representations (ICLR)*.

Arora, S., Cohen, N., Golowich, N., and Hu, W. (2019). A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pages 254–263. PMLR.

Bach, F. (2013). Convex relaxations of structured matrix factorizations. *arXiv preprint arXiv:1309.311*.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53.

Banerjee, A., Chen, T., and Zhou, Y. (2020). Derandomized pac-bayes margin bounds: Applications to non-convex and non-smooth predictors. *arXiv preprint arXiv:2002.09956*.

Barron, A. R. and Klusowski, J. M. (2019). Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17.

Bartlett, P. L. and Mendelson, S. (2001). Rademacher and gaussian complexities: Risk bounds and structural results. In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, volume 2111, pages 224–240. Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science.

Bordelon, B., Chaudhry, H. T., and Pehlevan, C. (2024). Infinite limits of multi-head transformer dynamics. *arXiv preprint arXiv:2405.15712*.

Burer, S. and Monteiro, R. D. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357.

Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849.

Deora, P., Ghaderi, R., Taheri, H., and Thrampoulidis, C. (2024). On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*.

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence*.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. (2022). Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR.

Feldman, V. and Vondrak, J. (2019). High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR.

Ge, R., Jin, C., and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: a unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1233–1242. JMLR.org.

Giampouras, P., Vidal, R., Rontogiannis, A., and Haeffele, B. D. (2020). A novel variational form of the schatten-p quasi-norm. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Golowich, N., Rakhlin, A., and Shamir, O. (2018). Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. (2018a). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30.

Haeffele, B. D. and Vidal, R. (2015). Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*.

Haeffele, B. D. and Vidal, R. (2017). Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7331–7339.

Haeffele, B. D. and Vidal, R. (2020). Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1468–1482.

HaoChen, J. Z., Wei, C., Lee, J., and Ma, T. (2021). Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR.

Hendrickx, J. and Olshevsky, A. (2010). Matrix p-norms are np-hard to approximate if $p \neq 1, 2, \infty$. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2802.

Huang, C. (2020). ReLU networks are universal approximators via piecewise linear or constant functions. *Neural Computation*, 32(11):2249–2278.

Imaizumi, M. and Schmidt-Hieber, J. (2023). On generalization bounds for deep networks based on loss surface implicit regularization. *IEEE Transactions on Information Theory*, 69(2):1203–1223.

Jia, X., Wang, H., Peng, J., Feng, X., and Meng, D. (2023). Preconditioning matters: Fast global convergence of non-convex matrix factorization via scaled gradient descent. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 76202–76213. Curran Associates, Inc.

Jin, J., Li, Z., Lyu, K., Du, S. S., and Lee, J. D. (2023). Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15200–15238. PMLR. ISSN: 2640-3498.

Kakade, S. M., Sridharan, K., and Tewari, A. (2008). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329. Publisher: Institute of Mathematical Statistics.

Li, G. and Wei, Y. (2023). A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. (2023). Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19565–19594. PMLR.

Li, Z., Luo, Y., and Lyu, K. (2020). Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*.

Lugosi, G. and Neu, G. (2022). Generalization bounds via convex analysis. In *Conference on Learning Theory*, pages 3524–3546. PMLR.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632.

McAllester, D. A. (1999). PAC-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM.

Muthukumar, R. and Sulam, J. (2023). Sparsity-aware generalization theory for deep neural networks. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 5311–5342. PMLR. ISSN: 2640-3498.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069 – 1097.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. (2017). Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018). A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*.

Nichani, E., Damian, A., and Lee, J. D. (2024). How transformers learn causal structure with gradient descent. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38018–38070. PMLR.

Oymak, S. and Soltanolkotabi, M. (2019). Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *Proceedings of the 36th International Conference on Machine Learning*, pages 4951–4960. PMLR. ISSN: 2640-3498.

Recht, B., Xu, W., and Hassibi, B. (2008). Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization. In *2008 47th IEEE Conference on Decision and Control*. IEEE.

Reddy, T. U. K. and Vidyasagar, M. (2023). Convergence of momentum-based heavy ball method with batch updating and/or approximate gradients. In *2023 Ninth Indian Control Conference (ICC)*, pages 182–187.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009). Stochastic convex optimization. In *COLT*, volume 2, page 5.

Singh, S. S. (2023). Analyzing transformer dynamics as movement through embedding space. *arXiv preprint arXiv:2308.10874*.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57.

Stöger, D. and Soltanolkotabi, M. (2021). Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843.

Stöger, D. and Zhu, Y. (2024). Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity. *arXiv preprint arXiv:2408.13276*.

Team, G. (2024). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tian, Y., Wang, Y., Chen, B., and Du, S. S. (2023). Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947.

Trauger, J. and Tewari, A. (2024). Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. N. and Chervonenkis, A. Y. (1968). The uniform convergence of frequencies of the appearance of events to their probabilities. *Doklady Akademii Nauk SSSR*, 181(4):781–783.

Vardi, G. (2023). On the implicit bias in deep-learning algorithms. *Commun. ACM*, 66(6):86–93.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vidal, R., Zhu, Z., and Haeffele, B. D. (2022). *Optimization Landscape of Neural Networks*, page 200–228. Cambridge University Press.

Wen, K., Li, Z., and Ma, T. (2023). Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yang, Y., Wipf, D. P., et al. (2022). Transformers from an optimization perspective. *Advances in Neural Information Processing Systems*, 35:36958–36971.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Zhang, Z., Fang, J., Lin, J., Zhao, S., Xiao, F., and Wen, J. (2020). Improved upper bound on the complementary error function. *Electronics Letters*, 56(13):663–665.

Zhou, Y., Alon, U., Chen, X., Wang, X., Agarwal, R., and Zhou, D. (2024). Transformers can achieve length generalization but not robustly. In *International Conference on Learning Representations*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/**Not Applicable**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A Convex Relaxation Approach to Generalization Analysis for Parallel Positively Homogeneous Networks: Supplementary Materials

In this supplementary material, we provide a detailed discussion of the rigorous technical aspects omitted from the main text. Additionally, we present a comprehensive review of related works and include a few numerical experiments. Below is the table of contents for this appendix/supplementary material.

Contents

A CONVEX BOUNDS FOR LEARNING

In this section, we discuss the proof for Theorem 1 that establishes the optimality gaps in the empirical and population landscapes. First, we analyze the convexity of the induced regularizer, $\Omega(\cdot)$ and properties of the stationary points in non-convex landscape. These are the key components of our proof for Theorem 1. We state a more general version of Assumption 2 by having the flexibility of the loss being strongly convex to derive tighter results.

Assumption 2' (Convex Loss)**.** *The loss $\ell(Y, \hat{Y})$ is second-order differentiable (written $\ell \in \mathcal{C}^2$), α-strong and L-smooth w.r.t. $\hat{Y}$, i.e, for any $Y, \hat{Y} \in \mathbb{R}^{n_Y}$*

$$0 \preceq \alpha I_{n_Y} \preceq \nabla^2_{\hat{Y}} \ell(Y, \hat{Y}) \preceq L I_{n_Y}. \tag{32}$$

Additionally, the gradient of the loss is bi-Lipschitz; that is, for all $Y_1, Y_2, \hat{Y}_1, \hat{Y}_2 \in \mathbb{R}^{n_Y}$

$$\|\nabla_{\hat{Y}} \ell(Y_2, \hat{Y}_2) - \nabla_{\hat{Y}} \ell(Y_1, \hat{Y}_1)\| \leq L\Big[\|Y_2 - Y_1\|_2 + \|\hat{Y}_2 - \hat{Y}_1\|_2\Big], \tag{33}$$

and the loss is constant if both the arguments are the same, i.e., for all $Y_1, Y_2 \in \mathbb{R}^{n_Y}$, $\ell(Y_1, Y_1) = \ell(Y_2, Y_2)$.

Note that we allow $\alpha = 0$, in which case we recover Assumption 2.

A.1 Induced Regularizer in Convex Space

First, we show that the induced regularizer is convex in the function spaces through Proposition 1.

Proposition 1 (Convexity of induced regularizer)**.** *Suppose assumptions 1-2' hold. Then $\Omega(f)$ is convex in f in the space of functions $\mathbb{R}^{n_X} \to \mathbb{R}^{n_Y}$.*

Proof. This proof is infinite dimensional extension of Haeffele and Vidal (2015). Recall the definition of induced regularizer:

$$\Omega(f) := \inf_{r, \{W_j\}} \Theta_r(\{W_j\}) \text{ such that } f(X) = \Phi_r(\{W_j\}); \forall X \in \mathcal{X}. \tag{34}$$

Define the function class

$$\mathcal{F}_\Phi := \{\Phi_r(\{W_j\}) : r \in \mathbb{N}, W_j \in \mathcal{W}\}. \tag{35}$$

By definition if $f \notin \mathcal{F}_\Phi$ then $\Omega(f)$ evaluates to infinity. Now suppose that $\beta \geq 0$ and for any $f \in \mathcal{F}_\Phi$,

$$\Omega(\beta f) = \inf_{r, \{W_j\}} \Theta_r(\{W_j\}) \text{ such that } \beta f(X) = \Phi_r(\{W_j\}); \forall X \in \mathcal{X}, \tag{36}$$

Now by Assumption 2', there exists $\hat{\beta}$ such that $\beta \Phi_r(\{W_j\}) = \Phi_r(\{\hat{\beta} W_j\})$, and $\beta \Theta_r(\{W_j\}) = \Theta_r(\{\hat{\beta} W_j\})$ (throughout note that this scaling is applied only to the $\mathcal{W}_p$ subset of parameters from Assumption 2, but we do not notate this explicitly for brevity of notation). Now we perform a change of variables in the induced regularizer, obtaining

$$\Omega(\beta f) = \inf_{r, \{\hat{\beta} W_j\}} \Theta_r(\{\hat{\beta} W_j\}) \text{ such that } \beta f(X) = \Phi_r(\{\hat{\beta} W_j\}); \forall X \in \mathcal{X}. \tag{37}$$

Then we have that

$$\Omega(\beta f) = \inf_{r, \{W_j\}} \beta \Theta_r(\{W_j\}) \text{ such that } \beta f(X) = \beta \Phi_r(\{W_j\}); \forall X \in \mathcal{X} = \beta \Omega(f). \tag{38}$$

We have established that the function $\Omega(\cdot)$ is 1-degree homogeneous. Now we prove that the function $\Omega(\cdot)$ is sub-additive. Choose any $f_1, f_2 \in \mathcal{F}_\Phi$, because the case when either of them is not in $\mathcal{F}_\Phi$ is trivially sub-additive. Recall

$$\Omega(f_1) = \inf_{r,\{W_j\}} \Theta_r(\{W_j\}) \text{ such that } f_1(X) = \Phi_r(\{W_j\}); \forall X \in \mathcal{X}, \tag{39}$$

$$\Omega(f_2) = \inf_{r,\{W_j\}} \Theta_r(\{W_j\}) \text{ such that } f_2(X) = \Phi_r(\{W_j\}); \forall X \in \mathcal{X}, \tag{40}$$

$$\Omega(f_1 + f_2) = \inf_{r,\{W_j\}} \Theta_r(\{W_j\}) \text{ such that } f_1(X) + f_2(X) = \Phi_r(\{W_j\}); \forall X \in \mathcal{X}. \tag{41}$$

For any $\epsilon > 0$ let $(r_1, \{W_j^1\})$ and $(r_2, \{W_j^2\})$ be parameters which come within ϵ of the infimum in the optimization problems for $\Omega(f_1)$ and $\Omega(f_2)$ respectively. Then note that

$$\Omega(f_1 + f_2) \leq \Theta_{r_1}(\{W_j^1\}) + \Theta_{r_2}(\{W_j^2\}) \leq \Omega(f_1) + \Omega(f_2) + 2\epsilon. \tag{42}$$

Letting $\epsilon \to 0$ gives that $\Omega(f_1 + f_2) \leq \Omega(f_1) + \Omega(f_2)$. Thus, as $\Omega(\cdot)$ is both positively homogenous with degree one and sub-additive, it is convex. □

From the above proposition we have that $\Omega(\cdot)$ is a convex function, therefore we have that $C_\cdot(\cdot)$ is indeed a convex function in the prediction functions space. Our results primarily depend upon the optimal regularization of the globally optimal solution of a convex function, $C_\cdot(\cdot)$. As we operate in the space of functions, it is very unlikely that we have the knowledge of the global optima. Nevertheless, by exploiting the convexity of $C_\cdot(\cdot)$ we can upper bound the optimal regularization. Proposition 2 establishes the upper bound for the optimal regularization for regression loss.

Proposition 2. *Consider $\ell(Y_1, Y_2) = \frac{1}{2}\|Y_1 - Y_2\|_2^2$, $\{W_j\} \in \mathcal{F}_\mathcal{W}$. Suppose $X \sim \mu$, ϵ is random variable such that $\mathbb{E}[\epsilon] = 0$ and independent from x. Let $Y = \Phi_r(\{W_j\})(X) + \epsilon$, and suppose f_μ^* is the global optimal solution of $C_\mu(\cdot)$. Then we have*

$$\Omega(\Phi_r(\{W_j\})) \geq \Omega(f_\mu^*). \tag{43}$$

Proof. As f_μ^* is the global optimal solution, we have that

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}\|\Phi_r(\{W_j\})(X) + \epsilon - f_\mu^*(X)\|_2^2\right] + \lambda\Omega(f_\mu) \leq \mathbb{E}\left[\frac{1}{2}\|\Phi_r(\{W_j\})(X) + \epsilon - \Phi_r(\{W_j\})(X)\|_2^2\right] \\ + \lambda\Omega(\Phi_r(\{W_j\}))\end{aligned} \tag{44}$$

Now, by re-arranging the terms we obtain

$$\mathbb{E}\left[\frac{1}{2}\langle\Phi_r(\{W_j\})(X) - f_\mu^*(X), \epsilon\rangle\right] + \lambda\Omega(f_\mu) \leq \lambda\Omega(\Phi_{r^*}(\{W_j\})). \tag{45}$$

As ϵ is independent of X,

$$\frac{1}{2}\langle\mathbb{E}_X\left[\Phi_r(\{W_j\})(X) - f_\mu^*(X)\right], \mathbb{E}_\epsilon[\epsilon]\rangle + \lambda\Omega(f_\mu) \leq \lambda\Omega(\Phi_{r^*}(\{W_j\})). \tag{46}$$

Then we have

$$\Omega(f_\mu) \leq \Omega(\Phi_r(\{W_j\})). \tag{47}$$

□

A.2 Proof of Theorem 1

Optimization algorithms used to optimize DNNs try to find the set of parameters that are first-order optimal. However, we do not have a guarantee that these points are saddle/local minima/global minima. In proposition 3, we provide properties that any first-order optimal satisfies for positively homogeneous networks.

Proposition 3 (Stationary Points)**.** *Under assumption 2, if $\{W_j\}$ are stationary points of $\mathsf{NC}_\mu(\cdot)$, then for all $j \in [r]$,*

$$\langle -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g, \Phi_r(\{W_j\})), \phi(W_j)\rangle = \theta(W_j). \tag{48}$$

Proof. This proof is similar to that of Proposition 2 in Haeffele and Vidal (2020) but applied to a general class of (*slightly*) positively homogeneous functions (see assumption 2).

From assumption 2, there exists a subset of parameters where both θ and ϕ are positively homogeneous. Let $\mathbf{w}_i$ be the subset of parameters in $\mathcal{W}_p$ from assumption 2. Then we have

$$\langle \mathbf{w}_i, \partial_{\mathbf{w}_i}\theta(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n)\rangle = \lim_{\epsilon\to 0}\left[\frac{\theta(\mathbf{w}_1,\dots,(1+\epsilon)\mathbf{w}_i,\dots,\mathbf{w}_n)}{\epsilon} - \frac{\theta(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n)}{\epsilon}\right]. \tag{49}$$

Let p_i be the homogeneous degree of the parameters $\mathbf{w}_i$. Note that $\partial_{\mathbf{w}_i}\theta(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n) \in \mathbb{R}^{dim(\mathbf{w}_i)\times 1}$, $\partial_{\mathbf{w}_i}\phi(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n) \in \mathbb{R}^{dim(\mathbf{w}_i)\times n_Y}$. Then

$$\langle \mathbf{w}_i, \partial_{\mathbf{w}_i}\theta(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n)\rangle = \theta(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n)\lim_{\epsilon\to 0}\frac{(1+\epsilon)^{p_i}-1}{\epsilon}, \tag{50}$$

$$= p_i\theta(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n). \tag{51}$$

Similarly, following a similar argument for ϕ we obtain

$$\langle \partial_{\mathbf{w}_i}\phi(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n), \mathbf{w}_i\rangle = p_i\phi(\mathbf{w}_1,\dots,\mathbf{w}_i,\dots,\mathbf{w}_n). \tag{52}$$

As W_j are the stationary points we have that

$$0 \in \partial_{W_j}\Phi_r(\{W_j\})\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_\mu + \lambda\partial_{W_j}\Theta_r(\{W_j\}). \tag{53}$$

Since $\Phi_r(\{W_j\}) = \sum_{j=1}^r \phi(W_j)$, we have that $\partial_{W_j}\Phi_r(\{W_j\}) = \partial_{W_j}\phi(W_j)$. Similarly, $\partial_{W_j}\Theta_r(\{W_j\}) = \partial_{W_j}\theta(W_j)$ holds true. Consequently,

$$0 \in \partial_{W_j}\phi(W_j)\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_\mu + \lambda\partial_{W_j}\theta(W_j). \tag{54}$$

Letting $W_j = \begin{bmatrix}\mathbf{w}_1 & \dots & \mathbf{w}_n\end{bmatrix}$, for all $\mathbf{w}_i$ it holds that

$$0 \in \partial_{\mathbf{w}_i}\phi(W_j)\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_\mu + \lambda\partial_{\mathbf{w}_i}\theta(W_j). \tag{55}$$

Taking the inner product of the above equation with $\mathbf{w}_i$, when $p_i \neq 0$ we have that

$$0 \in {\mathbf{w}_i}^T\partial_{\mathbf{w}_i}\phi(W_j)\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_\mu + \lambda{\mathbf{w}_i}^T\partial_{\mathbf{w}_i}\theta(W_j). \tag{56}$$

From (51) we have that

$$0 = p_i\phi(W_j)^T\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_\mu + \lambda p_i\theta(W_j). \tag{57}$$

Rearranging, we obtain

$$\langle -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\})), \phi(W_j)\rangle = \theta(W_j), \tag{58}$$

which holds for all $j \in [r]$. □

Proposition 3 establishes that at any stationary point, the inner product between the prediction errors and the predictions equates to the the current regularization. Next, we exploit this property of stationary points that enable us to tie the non-convex landscape to its convex counterpart. Lemma 1 establishes the difference between the non-convex and convex objective values at stationary points.

Lemma 1 (Optimality Gap)**.** *Let $\ell(\cdot,\cdot)$ denote any L-smooth, and α-strongly convex loss function, let q be some measure, and suppose that $\{W_j\}$ is a stationary point of $\mathsf{NC}_q(\{W_j\})$. Let f_q^* denote the global minimizer of $\mathsf{C}_q(\cdot)$. Then for any $f \in L^2(q)$, we have that*

$$\mathsf{C}_q(f_q^*) \leq \mathsf{NC}_q(\{W_j\}) \leq \mathsf{C}_q(f) + \lambda\Omega_q(f)\left[\Omega_q^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f - \Phi_r(\{W_j\})\|_q^2 \tag{59}$$

Proof. The loss $\ell(Y, \hat{Y})$ is (L, λ)-convex in $\hat{Y}$. Therefore, for any functions $g : \mathcal{X} \times E \to \mathcal{Y}$, and $f_1, f_2 \in L^2(q)$ we have that

$$\ell\left(g(X,\epsilon), f(X)\right) \geq \ell\left(g(X,\epsilon), \Phi_r(\{W_j\})(X)\right) \tag{60}$$
$$+ \left\langle \nabla_{\hat{Y}} \ell\left(g(X,\epsilon), \Phi_r(\{W_j\})(X)\right), f(X) - \Phi_r(\{W_j\})(X) \right\rangle_{\mathcal{Y}} \tag{61}$$
$$+ \frac{\alpha}{2} \|f(X) - \Phi_r(\{W_j\})(X)\|_{\mathcal{Y}}^2. \tag{62}$$

Taking expectations of both sides with respect to the probability measure q, we have that

$$\ell(g, f)_q \geq \ell\left(g, \Phi_r(\{W_j\})\right)_q + \left\langle \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), f - \Phi_r(\{W_j\}) \right\rangle_q + \frac{\alpha}{2} \|f - \Phi_r(\{W_j\})\|_q^2. \tag{63}$$

As the $\{W_j\}$ are the stationary points of $\mathsf{NC}_q(\{W_j\})$ from Proposition 3 we have that for all $j \in [r]$

$$\langle -\frac{1}{\lambda} \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), \phi(W_j) \rangle_q = \theta(W_j). \tag{64}$$

Summing the above identity up overall j, it holds that

$$\langle -\frac{1}{\lambda} \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\}) \rangle_q = \Theta_r(\{W_j\}). \tag{65}$$

Therefore, plugging this identity into the inequality (63), we have that

$$\ell(g, f)_q \geq \underbrace{\ell\left(g, \Phi_r(\{W_j\})\right)_q + \lambda \Theta_r(\{W_j\})}_{\mathsf{NC}_q(\{W_j\})} + \langle \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), f \rangle_q + \frac{\alpha}{2} \|f - \Phi_r(\{W_j\})\|_q^2, \tag{66}$$

which implies that

$$\ell(g, f)_q + \lambda \langle -\frac{1}{\lambda} \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), f \rangle_q \geq NC_q(\{W_j\}) + \frac{\alpha}{2} \|f - \Phi_r(\{W_j\})\|_q^2. \tag{67}$$

We have established from Proposition 1 that Ω is a convex function. As a well-known result from convex analysis (see Proposition 5) we have that for any convex function Ω and any $f, g \in L^2(q)$, it holds that $\langle f, g \rangle_q \leq \Omega_q(f) \Omega_q^\circ(g)$. Consequently,

$$\ell(g, f)_q + \lambda \Omega_q(f) \Omega_q^\circ\left(-\frac{1}{\lambda} \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right)\right) \geq NC_q(\{W_j\}) + \frac{\alpha}{2} \|f - \Phi_r(\{W_j\})\|_q^2. \tag{68}$$

Therefore, rearranging,

$$\underbrace{\ell(g, f)_q + \lambda \Omega_q(f)}_{\mathsf{C}_q(f)} + \lambda \Omega_q(f) \left[\Omega_q^\circ\left(-\frac{1}{\lambda} \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \geq NC_q(\{W_j\}) + \frac{\alpha}{2} \|f - \Phi_r(\{W_j\})\|_q^2, \tag{69}$$

and, as a result,

$$\mathsf{NC}_q(\{W_j\}) \leq \mathsf{C}_q(f) + \lambda \Omega_q(f) \left[\Omega_q^\circ\left(-\frac{1}{\lambda} \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2} \|f - \Phi_r(\{W_j\})\|_q^2. \tag{70}$$

Let $f_q^* = \arg\min_f \mathsf{C}_q(f)$, and $(r^*, \{W_j^*\}) = \arg\min_{r, \{W_j\}} NC_q(\{W_j\})$. Since f_q^* is the minimizer of $C_q(f)$, it holds that

$$\mathsf{C}_q(f_q^*) \leq \mathsf{C}_q(\Phi_{r^*}(\{W_j^*\})) = \ell(g, \Phi_{r^*}(\{W_j^*\}))_q + \lambda \Omega_q(\Phi_{r^*}(\{W_j^*\})). \tag{71}$$

Therefore, we obtain

$$\mathsf{C}_q(f_q^*) \leq \ell(g, \Phi_{r^*}(\{W_j^*\}))_q + \lambda \Omega_q(\Phi_{r^*}(\{W_j^*\})) \tag{72}$$
$$\leq \ell(g, \Phi_{r^*}(\{W_j^*\}))_q + \lambda \Theta_{r^*}(\{W_j^*\}) \tag{73}$$
$$= \mathsf{NC}_q(\{W_j^*\}) \tag{74}$$
$$\leq \mathsf{NC}_q(\{W_j\}). \tag{75}$$

Therefore, combining Equations (75) and (70), we obtain the bound

$$\mathsf{C}_q(f_q^*) \leq \mathsf{NC}_q(\{W_j\}) \leq \mathsf{C}_q(f) + \lambda\Omega_q(f)\left[\Omega_q^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f - \Phi_r(\{W_j\})\|_q^2. \tag{76}$$

□

Lemma 1 has established that the non-convex objective, $\mathsf{NC}_q(\cdot)$ is both upper and lower bounded by the convex function, $\mathsf{C}_q(\cdot)$. Now, we utilize this result to compute the empirical gap with the measure, μ_N for the stationary points obtained from the ERM. On these stationary points, we bound the optimality gap by changing the measure to μ, i.e., the behavior of ERM's first-order points on population landscape.

Theorem 3 (Global Optimality)**.** *Under assumptions 1, 2', 3. Let $f_{\mu_N}^*$ (or f_μ^*) be the global minimizer for $\mathsf{C}_{\mu_N}(\cdot)$ (or $\mathsf{C}_\mu(\cdot)$). For any stationary points, $(r, \{W_j\})$ of the function $\mathsf{NC}_{\mu_N}(\cdot)$ and any $f \in L^2(\mu) \cap L^2(\mu_N)$ the following items are true:*

1. *Empirical optimality gap:*

$$\mathsf{C}_{\mu_N}(f_{\mu_N}^*) \leq \mathsf{NC}_{\mu_N}(\{W_j\}) \leq \mathsf{C}_{\mu_N}(f) + \lambda\Omega(f)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f - \Phi_r(\{W_j\})\|_{\mu_N}^2, \tag{77}$$

2. *Population optimality gap:*

$$\begin{aligned}\mathsf{C}_\mu(f_\mu^*) \leq \mathsf{NC}_\mu(\{W_j\}) \leq \mathsf{C}_\mu(f) + \lambda\Omega(f)\left[\Omega_\mu^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f - \Phi_r(\{W_j\})\|_\mu^2 \\ + \left[\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}\right].\end{aligned} \tag{78}$$

where $\Omega_q^\circ(\cdot)$ is the polar in the measure q defined as

$$\Omega_q^\circ(g) := \sup_{\theta(W)\leq 1} \langle g, \phi(W)\rangle_q \tag{79}$$

Remarks: Setting $f = f_{\mu_N}^*$ in Equation 77 and taking $(r, \{W_j\})$ to be any stationary point of $\mathsf{NC}_{\mu_N}(\cdot)$ gives a means to verify if $\{W_j\}$ is a globally optimal solution. We see that it suffices to check if $\Phi_r(\{W_j\})$ is a first-order stationary point of $\mathsf{C}_{\mu_N}(\cdot)$, which is a necessary condition for a local minimum of convex functions.

From convex analysis, if a function $f \in L^2(\mu_N)$ is a first-order solution of C_{μ_N} then we have that 0 belongs to the sub-gradient of $\mathsf{C}_\mu(\cdot)$ at f. As the loss ℓ is first-order differentiable (by Assumption 3 or 2') we have that

$$0 \in \partial\mathsf{C}_{\mu_N}(f) \iff -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g, f)_{\mu_N} \in \partial\Omega(f), \tag{80}$$

where $\partial\mathsf{C}_{\mu_N}(f)$ denotes the subgradient of C (viewed as a function of f). The above condition for f can also be verified by a dual notion known as the polar condition, Definition 6 (Rockafellar, 1970). The sub-gradient of a convex function can be defined through the notion of it's polar via

$$\partial\Omega_{\mu_N}(f) = \left\{g \in L_2(\mu_N) : \langle g, f\rangle_{\mu_N} = \Omega_{\mu_N}(f), \Omega_{\mu_N}^\circ(g) \leq 1\right\}. \tag{81}$$

From Lemma 1 in the supplement of Haeffele and Vidal (2017) the following statements are equivalent:

1. $\{W_j\}$ is an optimal factorization of f; i.e, $\Theta_r(\{W_j\}) = \Omega_{\mu_N}(f)$.
2. $\exists h \in L^2(\mu_N)$ such that $\Omega_{\mu_N}^\circ(h) \leq 1$ and $\langle h, \Phi_r(\{W_j\})\rangle_{\mu_N} = \Theta_r(\{W_j\})$.
3. $\exists h \in L^2(\mu_N)$ such that $\Omega_{\mu_N}^\circ(h) \leq 1$ and $\langle h, \phi(W_j)\rangle_{\mu_N} = \theta(W_j); \forall i \in [r]$.

Further, if (2) or (3) above is satisfied then we have that $h \in \partial\Omega_{\mu_N}(f)$. From Proposition 3 we have that for any stationary point $(r, \{W_j\})$ of NC_{μ_N},

$$\langle -\frac{1}{\lambda}\nabla\ell_{\mu_N}(g, \Phi_r(\{W_j\})), \Phi_r(\{W_j\})\rangle_{\mu_N} = \Theta_r(\{W_j\}). \tag{82}$$

Consequently, to check if a stationary point is globally optimal, it then suffices to check whether the polar condition $\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla\ell(g,\Phi_r(\{W_j\}))_{\mu_N}\right) \leq 1$ holds at the stationary point, $(r,\{W_j\})$. In the case when the polar condition holds true, the upper bound evaluates to $\mathsf{C}_{\mu_N}(f^*_\mu)$ matching the lower bound of $\mathsf{NC}_{\mu_N}(\cdot)$, which in turn implies global optimality.

Then, we can claim the following:

> At a stationary point $\{W_j\}$, if $\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla\ell(g,\Phi_r(\{W_j\}))_{\mu_N}\right) \leq 1$, then $\{W_j\}$ is globally optimal.

Now we prove Theorem 1.

Proof. The proof sketch is similar to Proposition 4 from Haeffele and Vidal (2020). Equation (12) can be obtained from the Lemma 1, for any stationary points, $(r,\{W_j\})$ of $\mathsf{NC}_{\mu_N}(\cdot)$.

Since, $f \in L^2(\mu)\cap L^2(\mu_N) \subseteq L^2(\mu_n)$, and the parameters satisfy the equality in Lemma 1, we can conclude that Equation (12) holds. The local minima of $\mathsf{NC}_{\mu_N}(\cdot)$ need not be local minima of $\mathsf{NC}_{\mu}(\cdot)$, therefore we shall obtain an discrepency term. From the fact that ℓ is a α-strongly convex function we have the inequality

$$\ell(g,f)_\mu \geq \ell_\mu\left(g,\Phi_r(\{W_j\})\right) + \langle\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right), f-\Phi_r(\{W_j\})\rangle_\mu + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|^2_\mu. \tag{83}$$

Adding $\lambda\Theta_r(\{W_j\})$ on both sides we obtain the inequality

$$\begin{aligned}\ell(g,f)_\mu + \lambda\Theta_r(\{W_j\}) &\geq \ell_\mu\left(g,\Phi_r(\{W_j\})\right) + \lambda\Theta_r(\{W_j\}) &(84)\\ &\quad + \langle\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right), f-\Phi_r(\{W_j\})\rangle_\mu + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|^2_\mu. &(85)\end{aligned}$$

Now replacing the first term term on the side with $\mathsf{NC}_\mu(\{W_j\})$ we obtain

$$\ell(g,f)_\mu + \lambda\Theta_r(\{W_j\}) \geq \mathsf{NC}_\mu(\{W_j\}) + \langle\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right), f-\Phi_r(\{W_j\})\rangle_\mu + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|^2_\mu. \tag{86}$$

From Proposition 3 we have that for stationary points $\{W_j\}$, it holds that $\Theta_r(\{W_j\}) = \langle -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell_{\mu_N}(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_{\mu_N}$. Therefore, by plugging this into the inequality above, we obtain that

$$\begin{aligned}\ell(g,f)_\mu + \lambda\langle -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell_{\mu_N}(g,\Phi_r(\{W_j\}))_{\mu_N}, \Phi_r(\{W_j\})\rangle &&(87)\\ \geq \mathsf{NC}_\mu(\{W_j\}) + \langle\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right), f-\Phi_r(\{W_j\})\rangle_\mu + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|^2_\mu. &&(88)\end{aligned}$$

Rearranging the terms we have

$$\begin{aligned}\ell(g,f)_\mu + \lambda\langle -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right), f\rangle_\mu &&(89)\\ + \langle\nabla_{\hat{Y}}\ell_\mu(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell_{\mu_N}(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_{\mu_N} &&(90)\\ \geq \mathsf{NC}_\mu(\{W_j\}) + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|^2_\mu. &&(91)\end{aligned}$$

Next, the following inequality always holds:

$$\langle f, -\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\rangle_\mu \leq \Omega_\mu(f)\Omega^{\circ}_\mu\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\right). \tag{92}$$

Rearranging (98) and plugging in (92), we obtain the inequality

$$\begin{aligned}\ell(g,f)_\mu + \lambda\Omega_\mu(f)\Omega^{\circ}_\mu\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\right) &&(93)\\ + \langle\nabla_{\hat{Y}}\ell_\mu(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell_{\mu_N}(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_{\mu_N} &&(94)\\ \geq \mathsf{NC}_\mu(\{W_j\}) + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|^2_\mu. &&(95)\end{aligned}$$

We add and subtract $\Omega(f)$ to obtain

$$\ell(g,f)_\mu + \lambda\Omega_\mu(f) + \lambda\Omega_\mu(f)\left[\Omega_\mu^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\right)-1\right] \tag{96}$$

$$+\langle\nabla_{\hat{Y}}\ell_\mu(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell_{\mu_N}(g,\Phi_r(\{W_j\})),\Phi_r(\{W_j\})\rangle_{\mu_N} \tag{97}$$

$$\geq \mathsf{NC}_\mu(\{W_j\}) + \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|_\mu^2. \tag{98}$$

Rearranging the right most term and using the definition of $C_\mu(f)$ we obtain,

$$\mathsf{NC}_\mu(\{W_j\}) \leq \mathsf{C}_\mu(f) + \lambda\Omega_\mu(f)\left[\Omega_\mu^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\right)-1\right] - \frac{\alpha}{2}\|f-\Phi_r(\{W_j\})\|_\mu^2 \tag{99}$$

$$+\left[\langle\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_{\mu_N},\Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))_{\mu_N},\Phi_r(\{W_j\})\rangle_{\mu_N}\right]. \tag{100}$$

This yields the right hand side of (13). As for the left hand side, by definition we have that for any $(r,\{W_j\})$, $\mathsf{C}_\mu(f_\mu^*) \leq \mathsf{NC}_\mu(\{W_j\})$. This completes the proof. □

Theorem 1 provides the behavior of ERM solutions in the population landscape. This paves a path to bound the empirical and population objectives at these stationary points.

B STATISTICAL BOUNDS

This section provides a more general version of Theorem 2 that does not need Assumption 7 to hold uniformly for all the data points, (X,Y). Rather, we relax the assumption to the following.

Assumption 7' (Probabilistic boundedness)**.** *There exists a convex set, $\mathcal{C} \subseteq \mathbb{R}^{n_X}\times\mathbb{R}^{n_Y}$ such that*

$$P(\cap_{i=1}^N (X_i,\epsilon_i)\in\mathcal{C}) \geq 1-\delta_{\mathcal{C}}. \tag{101}$$

For all $(X,\epsilon)\in\mathcal{C}$, and $\{W_j\}\in\mathcal{F}_{\mathcal{W}}$ the predictions and gradients are bounded; i.e.,

$$\|\Phi_r(\{W_j\})(X)\| \leq B_\Phi,\ \|\nabla_{\hat{Y}}\ell(g(X,\epsilon),\Phi_r(\{W_j\})(X)\| \leq B_\ell. \tag{102}$$

Further, for any $\{W_j\},\{\tilde{W}_j'\}\in\mathcal{F}_{\mathcal{W}}$, $W,\tilde{W}\in\mathcal{F}_\theta$, $(X,\epsilon)\in\mathcal{C}$, the network ϕ and Φ_r are Lipschitz in the parameters; i.e,

$$\|\Phi_r(\{W_j\})(X)-\Phi_r(\{\tilde{W}_j\})(X)\| \leq \tilde{L}_\Phi \max_j \|W_j-\tilde{W}_j\|_2, \tag{103}$$

and

$$\|\phi(W)(X)-\phi(\tilde{W})(X)\| \leq \tilde{L}_\phi\|W-\tilde{W}\|_2. \tag{104}$$

Additionally, define the quantity

$$B(\mathcal{C}) := \Big[(1+\alpha)\sup_{\{W_j\}\in\mathcal{F}_{\mathcal{W}}}\left|\|f_\mu^*\circ\mathcal{P}_{\mathcal{C}}-\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\|_\mu^2 - \|f_\mu^*-\Phi_r(\{W_j\})\|_\mu^2\right| \tag{105}$$

$$+\sup_{\{W_j\}\in\mathcal{F}_{\mathcal{W}},W'\in\mathcal{F}_\theta}\left|\langle\nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_{\mathcal{C}},\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\right),\phi(W')\circ\mathcal{P}_{\mathcal{C}}\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right),\phi(W')\rangle_\mu\right| \tag{106}$$

$$+\sup_{\{W_j\}\in\mathcal{F}_{\mathcal{W}}}\left|\langle\nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_{\mathcal{C}},\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\right),\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right),\Phi_r(\{W_j\})\rangle_\mu\right|\Big], \tag{107}$$

where $\mathcal{P}_{\mathcal{C}}(\cdot)$ is the Euclidean projection to the set $\mathcal{C}$.

Comparison with Assumption 7: Unlike in Assumption 7, we do not require the equations (102), (103), and (104) to hold for all the inputs. However, we relax this restriction by assuming that there exists a convex set, $\mathcal{C}$ which consists of the data points with probability at least $1-\delta_{\mathcal{C}}$. For well-behaved probability distributions like sub-Gaussian distributions (see Assumption 2'), such a convex exists with very high probability; i.e., very small $\delta_{\mathcal{C}}$.

Now we state the general master theorem that relies on Assumptions 1, 2', 3, 4, 5, 6 and 7' (but not on Assumption 7).

Theorem 4 (General Master Theorem). *Suppose Assumptions 1, 2', 3, 4, 5, 6 and 7' hold. Let $\delta \in (0,1]$ be fixed, and let f_μ^* be the global optimum of C_μ. Suppose that $\gamma \geq \Omega(f_\mu^*)L_\phi$, and define*

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{Lip}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}; \tag{108}$$

$$\epsilon_2 = 4\tilde{L}_\Phi B_\Phi \max\left\{1, 2L + \frac{2B_\ell}{B_\Phi}, 8\Omega(f_\mu^*)\frac{B_\ell \tilde{L}_\phi}{\tilde{L}_\Phi B_\Phi}, 8L\Omega(f_\mu^*)\right\}. \tag{109}$$

Let $\{W_j\}$ denote any stationary point of $\mathsf{NC}_{\mu_N}(\cdot)$. Then with probability at least $1-(\delta+\delta_{\mathcal{C}})$, it holds that

$$\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \tag{110}$$

$$\lesssim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2n_Y}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{111}$$

$$+ \frac{B(\mathcal{C})}{n_Y} + (1+\alpha)\epsilon_1\sqrt{\frac{R\cdot \mathsf{dim}(\mathcal{W})\log\left(\gamma\epsilon_2 r_\theta/L_\phi\right)\log(N) + \log(1/\delta)}{N}}. \tag{112}$$

Additional Remarks: In addition to the discussion in Section 4, the general version mentioned above (i) takes into account unbounded sub-Gaussian distributions, and (ii) imposes a weaker notion of Lipschitz continuity on the parameters. For sub-Gaussian inputs, one may choose the convex set $\mathcal{C}$ to be a ball with radius $\mathbb{B}(g)$. As we grow g, the term $B(\mathcal{C})$ decays exponentially, and ϵ_2 only grows in the order of polynomial. This fast decay allows us to keep the statistical error under control while pertaining to the optimal sample complexity. Theorem 4 reduces to Theorem 2 by setting $\alpha = 0$ and $\mathcal{C} = \mathsf{conv}(\mathcal{X})$. Under this choice, Assumption 7' coincides with Assumption 7.

We discuss the proof in section B.2. Before diving into the proof, we discuss a few preliminaries on the covering number essential to estimate the capacity of the hypotheses class.

B.1 Computing Function Class Capacities

Lemma 2 (Covering number of $\mathcal{F}_\mathcal{W}$). *Under assumption 6, and 7' the ν-net covering number of the set $\mathcal{F}_\mathcal{W}$ on the metric, $\|.\|_{\infty,d}$ is upper bounded via*

$$\mathcal{C}_{\mathcal{F}_\mathcal{W}}(\nu) \leq \left(\mathcal{C}_\theta(L_\phi\nu/\gamma)\right)^R, \tag{113}$$

where $\mathcal{C}_\theta(\nu) := \mathcal{N}(\{W : \theta(W) \leq 1\}, d(.,.), \nu)$.

Proof. Recall that

$$\mathcal{C}_{\mathcal{F}_\mathcal{W}}(\nu) := \mathcal{N}(\mathcal{F}_\mathcal{W}, \max_i d(.,.), \nu); \tag{114}$$

$$\mathcal{C}_\theta(\nu) := \mathcal{N}(\mathcal{F}_\theta, d(.,.), \nu). \tag{115}$$

By the definition of ν covering number,

$$\mathcal{N}(\mathcal{F}_\mathcal{W}, \|.\|_{\infty,d}, \nu) := \inf\left|\left\{\{W_j^0\} \in \mathcal{F}_\mathcal{W} : \forall\{W_j\} \in \mathcal{F}_\mathcal{W} : \|\{W_j\} - \{W_j^0\}\|_{\infty,d} = \max_j d(W_j, W_j^0) \leq \nu\right\}\right|; \tag{116}$$

$$\mathcal{N}(\mathcal{F}_\theta, d(\cdot,\cdot), \nu) := \inf\left|\left\{W^0 \in \mathcal{F}_\theta : \forall W \in \mathcal{F}_\theta : d(W, W^0) \leq \nu\right\}\right|. \tag{117}$$

Therefore we can upper bound $\mathcal{N}(\mathcal{F}_\mathcal{W}, \|.\|_{\infty,d}, \nu)$ witht he product of $\mathcal{N}(\mathcal{F}_\theta, d(\cdot,\cdot), \nu)$ R times. We have

$$\mathcal{N}(\mathcal{F}_\mathcal{W}, \|.\|_{\infty,d}, \nu) \leq \left[\mathcal{N}\left(\frac{\gamma}{L_\phi}\mathcal{F}_\theta, d(.,.), \nu\right)\right]^R, \tag{118}$$

Rewriting the above for appropriately chosen ν we get

$$\mathcal{N}(\mathcal{F}_\mathcal{W}, \|.\|_{\infty,d}, \nu) \leq \left[\mathcal{N}\left(\mathcal{F}_\theta, d(.,.), \frac{L_\phi\nu}{\gamma}\right)\right]^R. \tag{119}$$

This concludes our proof. □

Lemma 3 (Bounding covering number)**.** *Consider a metric space, $(\mathcal{W} \subseteq \mathbb{R}^n, \|\cdot\|_2)$ and a compact set, $\mathcal{F}_{\mathcal{W}} \subseteq \mathcal{W}$. Suppose that there exist $r < \infty$ such that $\mathcal{F}_{\mathcal{W}} \subseteq \mathbb{B}(r)$. Then we have*

$$\mathcal{N}(\mathcal{F}_{\mathcal{W}}, \|\cdot\|_2, \nu) \leq \left(1 + \frac{2r}{\nu}\right)^n. \tag{120}$$

Proof. We have that $\mathcal{F}_{\mathcal{W}} \subseteq \mathbb{B}(r)$. By monotonicity of covering numbers, we have that

$$\mathcal{N}(\mathcal{F}_{\mathcal{W}}, \|\cdot\|_2, \nu) \leq \mathcal{N}(\mathbb{B}(r), \|\cdot\|_2, \nu). \tag{121}$$

From Corollary 4.2.13 in Vershynin (2018) we have that,

$$\mathcal{N}(\mathcal{F}_{\mathcal{W}}, \|\cdot\|_2, \nu) \leq \mathcal{N}(\mathbb{B}(r), \|\cdot\|_2, \nu) \leq \left(1 + \frac{2r}{\nu}\right)^n. \tag{122}$$

□

B.2 Proof of Theorem 2

This section discusses the proof of Theorem 4. We extensively use concentration results from Section D that are preliminaries for the upcoming technical details.

Proof. First, we recall the definition of generalization error:

$$\text{Generalization Error} := \left|\mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right|. \tag{123}$$

We can bound the above from the optimality gaps obtained in Theorem 1 via the following decomposition:

$$\mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) = \underbrace{\left[\mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{C}_{\mu}(f_{\mu})\right]}_{\text{Population Gap}} - \underbrace{\left[\mathsf{NC}_{\mu_N}(\{W_j\}) - \mathsf{C}_{\mu_N}(f_{\mu_N})\right]}_{\text{Empirical Gap}} \tag{124}$$

$$+ \underbrace{\left[\mathsf{C}_{\mu}(f_{\mu}) - \mathsf{C}_{\mu_N}(f_{\mu_N})\right]}_{\text{Convex Gap}}. \tag{125}$$

From Theorem 1 we have that for any $f_{\mu}, f_{\mu_N} \in L^2(\mu) \cap L^2(\mu_N)$ and stationary points $(r, \{W_j\})$, the empirical gap is bounded by

$$\mathsf{C}_{\mu_N}(f^*_{\mu_N}) - \mathsf{C}_{\mu_N}(f_{\mu_N}) \leq \mathsf{NC}_{\mu_N}(\{W_j\}) - \mathsf{C}_{\mu_N}(f_{\mu_N}) \tag{126}$$

$$\leq \lambda\Omega(f_{\mu_N}) \left[\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_{\mu_N} - \Phi_r(\{W_j\})\|^2_{\mu_N}, \tag{127}$$

and the population gap is bounded by

$$\mathsf{C}_{\mu}(f^*_{\mu}) - \mathsf{C}_{\mu}(f_{\mu}) \leq \mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{C}_{\mu}(f_{\mu}) \tag{128}$$

$$\leq \lambda\Omega(f_{\mu}) \left[\Omega^{\circ}_{\mu}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_{\mu} - \Phi_r(\{W_j\})\|^2_{\mu} \tag{129}$$

$$+ \left[\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu} - \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}\right]. \tag{130}$$

For any f_{μ}, f_{μ_N}, subtracting the above two equations we obtain

$$\mathsf{C}_{\mu}(f^*_{\mu}) - \mathsf{C}_{\mu_N}(f_{\mu_N}) - \lambda\Omega(f_{\mu_N}) \left[\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] + \frac{\alpha}{2}\|f_{\mu_N} - \Phi_r(\{W_j\})\|^2_{\mu_N} \tag{131}$$

$$\leq \mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) \tag{132}$$

$$\leq \lambda\Omega(f_{\mu}) \left[\Omega^{\circ}_{\mu}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_{\mu} - \Phi_r(\{W_j\})\|^2_{\mu} \tag{133}$$

$$+ \left[\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu} - \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}\right] \tag{134}$$

$$+ \left[\mathsf{C}_{\mu}(f_{\mu}) - \mathsf{C}_{\mu_N}(f^*_{\mu_N})\right]. \tag{135}$$

By choosing $f_\mu = f_{\mu_N} = f_\mu^*$ (as $f_\mu^* \in L^2(\mu) \cap L^2(\mu_N)$) and noting that f_μ^* is not a random variable unlike $f_{\mu_N}^*$ (which depends on the data points) we get

$$C_\mu(f_\mu^*) - C_{\mu_N}(f_\mu^*) - \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] + \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{136}$$

$$\leq \mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) \tag{137}$$

$$\leq \lambda\Omega(f_\mu^*)\left[\Omega_\mu^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \tag{138}$$

$$+ \left[\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}\right] \tag{139}$$

$$+ \left[\mathsf{C}_\mu(f_\mu^*) - \mathsf{C}_{\mu_N}(f_{\mu_N}^*)\right]. \tag{140}$$

Since f_μ^* is the global minimizer of $C_\mu(\cdot)$, it always holds that $\mathsf{C}_\mu(f_\mu^*) \leq C_\mu(f_{\mu_N}^*)$. We use this fact to upper bound the right side term, upon which we obtain the bound

$$\mathsf{C}_\mu(f_\mu^*) - \mathsf{C}_{\mu_N}(f_\mu^*) - \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] + \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{141}$$

$$\leq \mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) \tag{142}$$

$$\leq \lambda\Omega(f_\mu^*)\left[\Omega_\mu^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \tag{143}$$

$$+ \left[\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}\right] \tag{144}$$

$$+ \left[\mathsf{C}_\mu(f_{\mu_N}^*) - \mathsf{C}_{\mu_N}(f_{\mu_N}^*)\right]. \tag{145}$$

Now we add and subtract $\Omega_\mu^\circ(\cdot)$[5] and $\frac{\alpha}{2}\|f - \Phi_r(\{W_j\})\|_\mu^2$ on the right side. We then have that

$$\underbrace{\mathsf{C}_\mu(f_\mu^*) - \mathsf{C}_{\mu_N}(f_\mu^*)}_{=:T_1} - \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \tag{146}$$

$$+ \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{147}$$

$$\leq \mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) \tag{148}$$

$$\leq \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{149}$$

$$+ \frac{\alpha}{2}\left[\underbrace{\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2}_{=:T_2}\right] \tag{150}$$

$$+ \lambda\Omega(f_\mu^*)\left[\underbrace{\Omega_\mu^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - \Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right)}_{=:T_3}\right] \tag{151}$$

$$+ \left[\underbrace{\langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_{\mu_N}}_{=:T_5}\right] \tag{152}$$

$$+ \left[\underbrace{\mathsf{C}_\mu(f_{\mu_N}^*) - \mathsf{C}_{\mu_N}(f_{\mu_N}^*)}_{=:T_5}\right]. \tag{153}$$

Now we apply uniform concentration on the quantities T_1, T_2, T_3, T_4, and T_5 to get bound the statistical error terms.

From assumption 7' we assume that $\mathcal{C}$ is some convex set in $\mathbb{R}^{n_X} \times \mathbb{R}^{n_Y}$ such that the following hold true:

[5] We are ignoring the input arguments for brevity.

1. For any i.i.d. samples $\{X_i, \epsilon_i\}$ the $P(\bigcap_{i=1}^{N}(X_i, \epsilon_i) \in \mathcal{C}) \geq 1 - \delta_{\mathcal{C}}$.
2. For all $(X, \epsilon) \in \mathcal{C}$ and $\forall \zeta \in \mathcal{F}_{\mathcal{W}} : \|f_\zeta(X)\| \leq B_\Phi$.
3. For all $(X, \epsilon) \in \mathcal{C}$ we have $\forall \zeta \in \mathcal{F}_{\mathcal{W}} : \|\nabla_{\hat{Y}} \ell(g(X, \epsilon), f_\zeta(X))\| \leq B_\ell$.
4. For all $(X, \epsilon) \in \mathcal{C}$ and $\forall \zeta, \zeta' \in \mathcal{F}_{\mathcal{W}} : \|f_\zeta(X) - f_{\zeta'}(X)\| \leq \tilde{L}_\Phi \|\zeta - \zeta'\|_{\infty,2}$.
5. For all $(X, \epsilon) \in \mathcal{C}$ and $\forall \zeta, \zeta' \in \mathcal{F}_\theta : \|f_\zeta(X) - f_{\zeta'}(Z)\| \leq \tilde{L}_\phi \|\zeta - \zeta'\|_2$.
6. For any $\hat{Y}_1, \hat{Y}_2 \in \mathbb{R}^{n_Y}$ we have $\|\nabla_{\hat{Y}} \ell(Y, \hat{Y}_1) - \nabla_{\hat{Y}} \ell(Y, \hat{Y}_2)\| \leq L \|\hat{Y}_1 - \hat{Y}_2\|$.
7. $B_{nrm}(\mathcal{C}) := \sup_{\zeta \in \mathcal{F}_{\mathcal{W}}} \left| \|f_\mu^* \circ \mathcal{P}_{\mathcal{C}} - f_\zeta \circ \mathcal{P}_{\mathcal{C}}\|_\mu^2 - \|f_\mu^* - f_\zeta\|_\mu^2 \right| < \infty$.
8. $B_{plr}(\mathcal{C}) := \sup_{\zeta \in \mathcal{F}_{\mathcal{W}}, \zeta' \in \mathcal{F}_\theta} \left| \langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_{\mathcal{C}}, f_\zeta \circ \mathcal{P}_{\mathcal{C}}\right), f_{\zeta'} \circ \mathcal{P}_{\mathcal{C}} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, f_\zeta\right), f_{\zeta'} \rangle_\mu \right| < \infty$.
9. $B_{eql}(\mathcal{C}) := \sup_{\zeta \in \mathcal{F}_{\mathcal{W}}} \left| \langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_{\mathcal{C}}, f_\zeta \circ \mathcal{P}_{\mathcal{C}}\right), f_\zeta \circ \mathcal{P}_{\mathcal{C}} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, f_\zeta\right), f_\zeta \rangle_\mu \right| < \infty$.

Next, we define the events

$$\mathcal{E}_{cvx}(\epsilon) := \{\forall \zeta \in \mathcal{F}_{\mathcal{W}} : |\mathsf{C}_{\mu_N}(f_\zeta) - \mathsf{C}_\mu(f_\zeta)| \leq \epsilon + B_{nrm}(\mathcal{C})\};$$
$$\mathcal{E}_{eql}(\epsilon) := \{\forall \zeta \in \mathcal{F}_{\mathcal{W}} : \left| \langle \nabla_{\hat{Y}} \ell \left(g, f_\zeta\right), f_\zeta \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, f_\zeta\right), f_\zeta \rangle_{\mu_N} \right| \leq \epsilon + B_{eql}(\mathcal{C})\}.$$

Since $\Omega^\circ(\cdot)$ is positively homogeneous function we can ignore the scalar $-\frac{1}{\lambda}$ while defining the events below:

$$\mathcal{E}_{plr}(\epsilon) := \{\forall \zeta \in \mathcal{F}_{\mathcal{W}} : \left| \Omega_{\mu_N}^\circ \left(\nabla_{\hat{Y}} \ell \left(g, f_\zeta\right)\right) - \Omega_\mu^\circ \left(\nabla_{\hat{Y}} \ell \left(g, f_\zeta\right)\right) \right| \leq \epsilon + B_{plr}(\mathcal{C})\}; \tag{154}$$
$$\mathcal{E}_{nrm}(\epsilon) := \{\forall \zeta \in \mathcal{F}_{\mathcal{W}} : \left| \|f_\mu^* - f_\zeta\|_{\mu_N}^2 - \|f_\mu^* - f_\zeta\|_\mu^2 \right| \leq \epsilon + B_{nrm}(\mathcal{C})\}. \tag{155}$$

Finally, define the following good event:

$$\mathcal{E}_{good}(\epsilon) := \mathcal{E}_{cvx}\left(\frac{\epsilon}{4}\right) \cap \mathcal{E}_{eql}\left(\frac{\epsilon}{4}\right) \cap \mathcal{E}_{plr}\left(\frac{\epsilon}{4\Omega(f_\mu^*)}\right) \cap \mathcal{E}_{nrm}\left(\frac{\epsilon}{2}\right). \tag{156}$$

When the event $\mathcal{E}_{good}(\epsilon)$ holds then we obtain following from the inequality (153),

$$-\epsilon/4 - B_{cvx}(\mathcal{C}) - \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] + \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{157}$$
$$\leq \mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\}) \tag{158}$$
$$\leq \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] - \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 \tag{159}$$
$$+ \frac{\alpha}{2}\left[\epsilon/2 + B_{nrm}(\mathcal{C})\right] + \lambda\Omega(f_\mu^*)\left[\epsilon/(4\lambda\Omega(f_\mu^*)) + B_{plr}(\mathcal{C})\right] \tag{160}$$
$$+ \left[\epsilon/4 + B_{eql}(\mathcal{C})\right] + \left[\epsilon/4 + B_{nrm}(\mathcal{C})\right]. \tag{161}$$

For $\alpha \geq 0$, these inequalities imply that

$$|NC_\mu(\{W_j\}) - NC_{\mu_N}(\{W_j\})| \leq \lambda\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \tag{162}$$
$$- \frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 + (1+\alpha)\epsilon + (1+\alpha)B_{nrm}(\mathcal{C}) + B_{eql}(\mathcal{C}) + \lambda\Omega(f_\mu^*)B_{plr}(\mathcal{C}). \tag{163}$$

Equation (163) holds with probability $\mathbb{P}(\mathcal{E}_{good}(\epsilon))$. We can bound the good event with union bound via

$$\mathbb{P}(\mathcal{E}_{good}(\epsilon)) \geq 1 - \underbrace{\mathbb{P}\left(\mathcal{E}_{nrm}^c\left(\frac{\epsilon}{2}\right)\right)}_{\text{Lemma 8}} - \underbrace{\mathbb{P}\left(\mathcal{E}_{cvx}^c\left(\frac{\epsilon}{4}\right)\right)}_{\text{Lemma 9}} - \underbrace{\mathbb{P}\left(\mathcal{E}_{eql}^c\left(\frac{\epsilon}{4}\right)\right)}_{\text{Lemma 10}} - \underbrace{\mathbb{P}\left(\mathcal{E}_{plr}^c\left(\frac{\epsilon}{4\lambda\Omega(f_\mu^*)}\right)\right)}_{\text{Lemma 11}} \tag{164}$$

Under Assumptions 1-6 and 7' we can apply Lemma 8 , 9, 10, and 11 to bound the probability of the occurrence of the events, $\mathcal{E}_{cvx}(\cdot), \mathcal{E}_{eql}(\cdot), \mathcal{E}_{plr}(\cdot)$, and $\mathcal{E}_{nrm}(\cdot)$.

Define the constants

$$B_1 := 4n_Y L\left[(\gamma^2 + \|g\|^2_{\text{Lip}})\sigma_X^2 + \|g\|^2_{\text{Lip}}\sigma^2_{Y|X}\right]; \tag{165}$$

$$B_2 := 16n_y\gamma\|\nabla_{\hat{Y}}\ell\|_{\text{Lip}}\sigma_X\sqrt{(\gamma^2 + \|g\|^2_{\text{Lip}})\sigma_X^2 + \|g\|^2_{\text{Lip}}\sigma^2_{E|X}}; \tag{166}$$

$$B_3 := 16n_y\Omega(f^*_\mu)L_\phi\|\nabla_{\hat{Y}}\ell\|_{\text{Lip}}\sigma_X\sqrt{(\gamma^2 + \|g\|^2_{\text{Lip}})\sigma_X^2 + \|g\|^2_{\text{Lip}}\sigma^2_{E|X}}; \tag{167}$$

$$B_4 := 128n_Y\gamma^2\sigma_X^2; \tag{168}$$

$$\epsilon_0 := \max\{B_1, B_2, B_3, B_4\}; \tag{169}$$

$$\epsilon_1 := \max\{B_1, B_2, B_3, B_4\}; \tag{170}$$

$$b_1 := 8B_\ell\tilde{L}_\Phi; \tag{171}$$

$$b_2 := 8\tilde{L}_\Phi\left[B_\ell + B_\Phi L\right]; \tag{172}$$

$$b_3 := 32\Omega(f^*_\mu)\max\{\tilde{L}_\phi B_\ell, L\tilde{L}_\Phi B_\Phi\}; \tag{173}$$

$$b_4 := 4\tilde{L}_\Phi B_\Phi; \tag{174}$$

$$\epsilon_2 := \max\{b_1, b_2, b_3, b_4\}. \tag{175}$$

Under the above conditions by Lemma 8, for any $\epsilon \in [0, B_4]$ we have that

$$\mathbb{P}\left(\mathcal{E}^c_{nrm}\left(\frac{\epsilon}{2}\right)\right) \leq \delta_{\mathcal{C}} + c_4\exp\left(\log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{b_4}\right)\right) - N\left(\frac{\epsilon}{B_4}\right)^2\right). \tag{176}$$

By Lemma 9, for any $\epsilon \in [0, B_1]$, we have

$$\mathbb{P}\left(\mathcal{E}^c_{cvx}\left(\frac{\epsilon}{4}\right)\right) \leq \delta_{\mathcal{C}} + 2\exp\left(\log\left(C_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{b_1}\right)\right) - c_1 N\left(\frac{\epsilon}{B_1}\right)^2\right), \tag{177}$$

for some positive constant, c_1.

Additionally by Lemma 10, for any $\epsilon \in [0, B_2]$ we have that

$$\mathbb{P}\left(\mathcal{E}^c_{eql}\left(\frac{\epsilon}{4}\right)\right) \leq \delta_{\mathcal{C}} + c_1\exp\left(\log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{b_2}\right)\right) - N\left(\frac{\epsilon}{B_2}\right)^2\right), \tag{178}$$

for some positive constant, c_2. Furthermore, by Lemma 11, for any $\epsilon \in [0, B_3]$ we have that

$$\mathbb{P}\left(\mathcal{E}^c_{plr}\left(\frac{\epsilon}{4\Omega(f^*_\mu)}\right)\right) \leq \delta_{\mathcal{C}} + c_3\exp\left(\log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{b_3}\right)\right) + \log\left(\mathcal{C}_{\mathcal{F}_\theta}\left(\frac{\epsilon}{b_3}\right)\right) - N\left(\frac{\epsilon}{B_3}\right)^2\right), \tag{179}$$

for some positive constant, c_3.

For the inequalities (177), (178), (179), and (176) to all hold we choose $\epsilon \in [0, \epsilon_0]$ and we upper bound the covering numbers $C_{\mathcal{F}_{\mathcal{W}}}(\nu)$ as they are strictly decreasing in ν by definition. Therefore, we have that

$$\max\left\{\log\left(C_{\mathcal{F}_{\mathcal{W}}}\left(\tfrac{\epsilon}{b_1}\right)\right), \log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\tfrac{\epsilon}{b_2}\right)\right), \log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\tfrac{\epsilon}{b_3}\right)\right), \log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\tfrac{\epsilon}{b_4}\right)\right)\right\} \leq \log\left(C_{\mathcal{F}_{\mathcal{W}}}\left(\tfrac{\epsilon}{\epsilon_2}\right)\right), \tag{180}$$

and

$$\log\left(\mathcal{C}_{\mathcal{F}_\theta}\left(\frac{\epsilon}{b_3}\right)\right) \leq \log\left(C_{\mathcal{F}_\theta}\left(\frac{\epsilon}{\epsilon_2}\right)\right). \tag{181}$$

Now we plug in inequalities (177), (178), (179), and (176) in the inequality (164). Denote

$$c_5 := \max\{2, c_2, c_3, c_4\}. \tag{182}$$

Then

$$\mathbb{P}(\mathcal{E}_{good}(\epsilon)) \geq 1 - c_5 \exp(\log(\mathcal{C}_{\mathcal{F}_\mathcal{W}}(\epsilon/\epsilon_3))) \times \left[\exp\left(-c_1 N \left(\frac{\epsilon}{B_1}\right)^2\right) + \exp\left(-N\left(\frac{\epsilon}{B_2}\right)^2\right)\right. \tag{183}$$

$$\left. + \exp\left(\log\left(C_{\mathcal{F}_\theta}\left(\frac{\epsilon}{\epsilon_3}\right)\right) - N\left(\frac{\epsilon}{B_3}\right)^2\right) + \exp\left(-N\left(\frac{\epsilon}{B_4}\right)^2\right)\right] - 4\delta_\mathcal{C}. \tag{184}$$

Now we lower bound the right side by replacing B_1, B_2, B_3, B_4 with the upper bound ϵ_1 yielding

$$\mathbb{P}(\mathcal{E}_{good}(\epsilon)) \geq 1 - c_6 \exp\left(\log\left(\mathcal{C}_{\mathcal{F}_\mathcal{W}}\left(\epsilon/\epsilon_2\right)\right) + \log\left(C_{\mathcal{F}_\theta}\left(\frac{\epsilon}{\epsilon_3}\right)\right) - c_7 N\left(\frac{\epsilon}{\epsilon_1}\right)^2\right) - 4\delta_\mathcal{C},$$

for some positive constants, c_6, c_7.

From Lemma 2 we have that for any $\nu > 0$ it holds that

$$\log(\mathcal{C}_{\mathcal{F}_\mathcal{W}}(\nu)) \leq R \log(\mathcal{C}_{\mathcal{F}_\theta}\left(L_\phi \nu/\gamma\right)). \tag{185}$$

Then we obtain

$$\mathbb{P}(\mathcal{E}_{good}(\epsilon)) \geq 1 - c_8 \exp\left(R \log\left(C_{\mathcal{F}_\theta}\left(\frac{L_\phi \epsilon}{\gamma \epsilon_2}\right)\right) - c_9 N\left(\frac{\epsilon}{\epsilon_1}\right)^2\right) - 4\delta_\mathcal{C}, \tag{186}$$

for some positive constants c_8, c_9.

From inequality (163), and (186) for any $\epsilon \in [0, \epsilon_0]$ we have that

$$\mathbb{P}\left(|NC_\mu(\{W_j\}) - NC_{\mu_N}(\{W_j\})| \geq \lambda\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right]\right.$$

$$\left. -\frac{\alpha}{2}\|f_\mu^* - \Phi_r(\{W_j\})\|^2_{\mu_N} + (1+\alpha)\epsilon + B_{eql}(\mathcal{C}) + B_{plr}(\mathcal{C}) + (1+\alpha)B_{nrm}(\mathcal{C})\right)$$

$$\leq c_8 \exp\left(R\log\left(C_{\mathcal{F}_\theta}\left(\frac{L_\phi\epsilon}{\gamma\epsilon_2}\right)\right) - c_9 N\left(\frac{\epsilon}{\epsilon_1}\right)^2\right) + 4\delta_\mathcal{C}. \tag{187}$$

Next, we derive the operation conditions for ϵ in terms of B_2, B_3, and B_4.

- $B_2 \geq B_3$: observe that

$$B_2 \geq B_3 \iff \gamma \geq \Omega(f_\mu^*)L_\phi, \tag{188}$$

  which establishes an upper bound on the regularization parameter.

- To establish a lower bound on regularization, we will require that $\min\{B_2, B_3\} \geq B_4$: We have that

$$\min\{B_2, B_3\} \geq B_4 \iff \tag{189}$$

$$\left(4\min\left\{1, \frac{\Omega(f_\mu^*)L_\phi}{\gamma}\right\}\right)4\gamma^2\sigma_X^2\sqrt{\left(1 + \|g\|^2_{\text{Lip}}/\gamma^2\right) + \frac{\|g\|^2_{\text{Lip}}\sigma^2_{Y|X}}{\gamma^2\sigma_X^2}} \geq 16\gamma^2\sigma_X^2. \tag{190}$$

  It is sufficient to have the below inequality to hold:

$$\min\left\{1, \frac{\Omega(f_\mu^*)L_\phi}{\gamma}\right\} \geq \frac{\gamma}{\sqrt{\gamma^2 + \|g\|^2_{\text{Lip}}}} \implies \min\{B_2, B_3\} \geq B_4. \tag{191}$$

Therefore, $\gamma \geq \Omega(f_\mu^*)L_\phi$ is sufficient condition for $B_2 \geq B_3 \geq B_4$.

Then we have that

$$\epsilon_0 = \min\{B_1, B_4\} = 16 n_Y \gamma^2 \sigma_X^2 \min\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|^2_{\text{Lip}}}{\gamma^2}\left(1 + \frac{\sigma^2_{Y|X}}{\sigma_X^2}\right)\right]\right\}, \tag{192}$$

and

$$\epsilon_1 = 16 n_Y \gamma^2 \sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}. \tag{193}$$

Now rescale the quantities $\epsilon \leftarrow n_Y \frac{\epsilon}{(1+\alpha)}$, $\epsilon_0 \leftarrow n_Y \epsilon_0$, and $\epsilon_1 \leftarrow n_Y \epsilon_1$. Then we have

$$\epsilon_0 = 16\gamma^2\sigma_X^2 \min\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}; \tag{194}$$

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}; \tag{195}$$

$$\epsilon_2 = \max\{8B_\ell \tilde{L}_\Phi, 8[\tilde{L}_\Phi B_\ell + \tilde{L}_\Phi B_\Phi L], 32\Omega(f_\mu^*)\tilde{L}_\Phi \max\{\tilde{L}_\phi B_\ell / \tilde{L}_\Phi, L B_\Phi\}, 4\tilde{L}_\Phi B_\Phi\}, \tag{196}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \geq \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right]\right. \\
&\left. -\frac{\alpha}{2n_Y}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 + \frac{1}{n_Y}\left[B_{eql}(\mathcal{C}) + \lambda\Omega(f_\mu^*)B_{plr}(\mathcal{C}) + (1+\alpha)B_{nrm}(\mathcal{C})\right] + \epsilon\right) \\
&\qquad \leq c_8 \exp\left(R\log\left(C_{\mathcal{F}_\theta}\left(\frac{L_\phi \epsilon}{\gamma \epsilon_2}\right)\right) - c_9 N\left(\frac{\epsilon}{(1+\alpha)\epsilon_1}\right)^2\right) + 4\delta_{\mathcal{C}}.
\end{aligned} \tag{197}$$

Now we bound the covering number under Assumption 4 and Lemma 2 via

$$\log\left(C_{\mathcal{F}_\theta}\left(\frac{L_\phi \epsilon}{\gamma \epsilon_2}\right)\right) \leq \mathsf{dim}(\mathcal{W})\log(1 + 2\gamma\epsilon_2 r_\theta/(L_\phi\epsilon)) \leq c_{11}\mathsf{dim}(\mathcal{W})\log(\gamma\epsilon_2 r_\theta/(L_\phi\epsilon)) \tag{198}$$

for some positive constant c_{11}.

Define

$$B(\mathcal{C}) := B_{eql}(\mathcal{C}) + \lambda\Omega(f_\mu^*)B_{plr}(\mathcal{C}) + (1+\alpha)B_{nrm}(\mathcal{C}). \tag{199}$$

Then we have that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \geq \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega_{\mu_N}^\circ\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right]\right. \\
&\qquad \left. -\frac{\alpha}{2n_Y}\|f_\mu^* - \Phi_r(\{W_j\})\|_{\mu_N}^2 + \frac{1}{n_Y}B(\mathcal{C}) + \epsilon\right) \\
&\leq c_{12}\exp\left(c_{11}R\mathsf{dim}(\mathcal{W})\log(\gamma\epsilon_2 r_\theta/(L_\phi\epsilon)) - c_{12}N\left(\frac{\epsilon}{(1+\alpha)\epsilon_1}\right)^2\right) + 4\delta_{\mathcal{C}}.
\end{aligned} \tag{200}$$

For some fixed $\delta \in (0, 1]$ choose

$$\epsilon = \tilde{\Theta}\left((1+\alpha)\epsilon_1\sqrt{\frac{R\mathsf{dim}(\mathcal{W})\log(\gamma\epsilon_2 r_\theta/L_\phi)\log(N) + \log(1/\delta)}{N}}\right). \tag{201}$$

Then the right side term of inequality (200) will be

$$\exp\left(R\mathsf{dim}(\mathcal{W})\log(\gamma\epsilon_2 r_\theta/(L_\phi\epsilon)) - c_{12}N\left(\frac{\epsilon}{(1+\alpha)\epsilon_1}\right)^2\right) = \tilde{\mathcal{O}}(\delta). \tag{202}$$

Rewriting the equation (200), we have

$$\mathbb{P}\left(\frac{1}{n_Y}\left|\mathsf{NC}_{\mu}(\{W_j\})-\mathsf{NC}_{\mu_N}(\{W_j\})\right| \gtrsim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right)\right)-1\right]\right.$$
$$-\frac{\alpha}{2n_Y}\|f_\mu^*-\Phi_r(\{W_j\})\|^2_{\mu_N}+\frac{1}{n_Y}B(\mathcal{C})$$
$$\left.+(1+\alpha)\epsilon_1\sqrt{\frac{R\mathsf{dim}(\mathcal{W})\log(\gamma\epsilon_2 r_\theta/L_\phi)\log(N)+\log(1/\delta)}{N}}\right)\lesssim \delta+\delta_{\mathcal{C}}. \tag{203}$$

□

Theorem 4 has established generalization error for a generic parallel positively homogeneous network. Theorem 2 mentioned in the main text is a special case of Theorem 4, with the choice of convex set $\mathcal{C}=\mathsf{conv}(\mathcal{X}\times\mathbb{R}^{n_Y})$ by changing Assumption 7' to Assumption 7. Further, $B(\mathcal{X}\times\mathbb{R}^{n_Y})$ will evaluate to 0, as $\mathcal{P}_{\mathsf{conv}(\mathcal{X}\times\mathbb{R}^{n_Y})}(\cdot)$ is just an identity operator.

C APPLICATIONS

In this section, we apply our Theorem 4 for various applications. We apply our general theorem to low-rank matrix sensing, structured matrix sensing, two-layer linear neural network, two-layer ReLU neural network, and multi-head attention.

C.1 Low-Rank Matrix Sensing

In this section, we state the corollary and its proof for matrix sensing, which is a direct consequence of Theorem 4. Firstly, we need to choose a convex set, $\mathcal{C}$, such that the Assumption 7' is satisfied. For matrix sensing we choose, $\mathcal{C}=\{(X,\epsilon):\|X\|_F\le g,\|\epsilon\|_F\le g\}$ to verify Assumption 7'. We need to compute, $B(\mathcal{C})$. This involves computing the expectation over the projection. Lemma 4 is pivotal for estimating $B(\mathcal{C})$ in all the applications that are going to be discussed here.

Lemma 4 (Projection of Gaussian vector on balls)**.** *Consider a n-dimensional Gaussian vector* $\mathbf{x}\sim\mathcal{N}(0,(1/n)I_n)$*. Let* M *be a fixed matrix in* $\mathbb{R}^{n\times n}$ *and* $\mathcal{A}$ *be any set. Then*

$$\left|\langle M,\mathbb{E}\left[\left[\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right]\right]\mathbf{1}_{\mathcal{A}}(\mathbf{x})\rangle\right|\le\begin{cases}ge^{-g^2/2}\|M\|_2 & \text{if } g\ge 1\\ \frac{1}{g}e^{-g^2/2}\|M\|_2 & \text{otherwise}\end{cases} \tag{204}$$

where $\mathcal{P}_{\mathbb{B}(g)}(\cdot)$ *is Euclidean projection onto the ball* $\mathbb{B}(g):=\{\mathbf{x}:\|\mathbf{x}\|_2\le g\}$.

Proof. Define an event $\mathcal{E}:=\{\mathbf{x}\in\mathcal{B}(g)\}$. When $\mathcal{E}$ holds the function evaluates to zero,

$$\mathbb{E}\left[\left[\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right]\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]=\mathbb{E}\left[\left(\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right)\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right], \tag{205}$$

so it suffices to consider the complement of the event $\mathcal{E}$. Now we take the inner product with M yielding

$$\begin{aligned}\left|\langle M,\mathbb{E}\left[\left[\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right]\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\rangle\right| &= \left|\langle M,\mathbb{E}\left[\left(\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right)\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\rangle\right|\\ &\overset{(a)}{\le}\|M\|_2\|\mathbb{E}\left[\left(\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right)\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\|\\ &\overset{(b)}{\le}\|M\|_2\mathbb{E}\left[\|\left(\mathbf{x}\mathbf{x}^T-\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right)\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\|\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\\ &\overset{(c)}{=}\|M\|_2\mathbb{E}\left[\|\mathbf{x}\|_2^2-g^2|\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\\ &\overset{(d)}{=}\|M\|_2\mathbb{E}\left[\left|\|\mathbf{x}\|_2^2-g^2\right||\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\\ &\overset{(e)}{\le}\|M\|_2\mathbb{E}\left[\left|\|\mathbf{x}\|_2^2-g^2\right||\mathbf{1}_{\mathcal{E}^c}(\mathbf{x})\right]\\ &\overset{(f)}{=}\|M\|_2\int_{\mathbf{x}\in\mathcal{E}^c}\left|\|\mathbf{x}\|_2^2-g^2\right|\frac{1}{\sqrt{2\pi}}e^{-\frac{\|\mathbf{x}\|_2^2}{2}}d\mathbf{x}\\ &\overset{(g)}{=}\|M\|_2\left[ge^{-g^2/2}-\sqrt{\frac{\pi}{2}}(g^2-1)\mathsf{erfc}(g/\sqrt{2})\right].\end{aligned}$$

The aforementioned computations involves (a) Cauchy-Schwartz inequality, (b) Jensen's inequality, (c) the norm of $\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})$ when $\mathbf{x} in \mathcal{E}^c$ is g, (d) conditioning on indicator functions, (e) removing the conditioning increases the expectation over non-negative terms, (f) we apply the density of Gaussian, (g) standard normal integral. As a consequence of Theorem 1 from Zhang et al. (2020) we bound the complement error function,

$$\frac{e^{-z^2}}{\sqrt{\pi}z} \geq \mathsf{erfc}(z) \geq \frac{2}{\sqrt{\pi}} \frac{e^{-z^2}}{z + \sqrt{z^2+2}}. \tag{206}$$

Then we have that

$$\left|\langle M, \mathbb{E}\left[\left[\mathbf{x}\mathbf{x}^T - \mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})\mathcal{P}_{\mathcal{B}(g)}(\mathbf{x})^T\right]\mathbf{1}_{\mathcal{A}}(\mathbf{x})\right]\rangle\right| \leq \begin{cases} g e^{-g^2/2}\|M\|_2 & \text{if } g \geq 1 \\ \frac{1}{g} e^{-g^2/2}\|M\|_2 & \text{otherwise.} \end{cases} \tag{207}$$

□

Now, we state the generalization bound for the low-rank matrix sensing followed by its proof.

Corollary 4 (Low-Rank Matrix Sensing)**.** *Consider the true model for (X, y), where $X \in \mathbb{R}^{m\times n}$ is a random matrix with i.i.d. entries $X_{lk} \sim \mathcal{N}(0, \frac{1}{mn})$ and $y = \langle M^*, X\rangle + \epsilon$, where $M^* \in \mathbb{R}^{m\times n}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent from X. For all $i \in [N]$, let (X_i, y_i) be i.i.d. samples from this true model. Consider the estimator $\hat{y} = \langle UV^T, X\rangle$, where $U \in \mathbb{R}^{m\times R}$ and $V \in \mathbb{R}^{n\times R}$. Let $\delta \in (0, 1]$ be fixed. Define the non-convex problem*

$$\mathsf{NC}^{\mathsf{MS}}_{\mu_N}((U, V)) := \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \langle UV^T, X_i\rangle\right)^2 + \lambda\sum_{j=1}^{R}\|\mathbf{u}_j\|_2\|\mathbf{v}_j\|_2, \tag{208}$$

and define $\mathsf{NC}^{MS}_{\mu}((U, V))$ similarly with the sum over i replaced by expectation taken over (X, y).

Let (U, V) be a stationary point of $\mathsf{NC}^{\mathsf{MS}}_{\mu_N}((U, V))$. Suppose there exists $C_{UV}, B_u, B_v > 0$ such that $\|UV^T\|_2 \leq C_{UV}\|M^*\|_*$, and for all $j \in [R]$, $\|\mathbf{u}_j\|_2 \leq B_u$, $\|\mathbf{v}_j\|_2 \leq B_v$. Then, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}\left|\mathsf{NC}^{\mathsf{MS}}_{\mu}((U, V))) - \mathsf{NC}^{\mathsf{MS}}_{\mu_N}((U, V))\right| \lesssim \|M^*\|_* &\left[\|\frac{1}{N}\sum_{i=1}^{N}(y_i - \langle UV^T, X_i\rangle)X_i\|_2 - \lambda\right] \\ &+ C^2_{UV}\|M^*\|^2_* \times \sqrt{\frac{R\log\left(R(C_{UV}+B_uB_v)\right)(m+n)\log(N)+\log(1/\delta)}{N}}.\end{aligned}$$

Proof. We set the following to obtain a generalization bound from Theorem 4 for the case of matrix sensing. First,

$$\ell(Y, \hat{Y}) = \frac{1}{2}\|Y - \hat{Y}\|^2_2 \implies (\alpha, L) = (0, 1); \tag{209}$$

$$\phi(W) = \langle \mathbf{u}\mathbf{v}^T, X\rangle; \tag{210}$$

$$\theta(W) = \|\mathbf{u}\|_2\|\mathbf{v}\|_2. \tag{211}$$

Estimating $\Omega(f^*_\mu)$: Since, M^* is the true matrix the regularizer at globally optimal solution can be upper bounded by Proposition 2,

$$\Omega(f^*_\mu) \leq \|M^*\|_*. \tag{212}$$

Estimating $\Omega^{\circ}_{\mu_N}(\cdot)$: Now we move on to compute the polar. We have

$$\begin{aligned}
\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))\right) &= \Omega^{\circ}_{\mu_N}\left(\frac{1}{\lambda}(g-\Phi_r(\{W_j\}))\right) \\
&= \sup_{\|\mathbf{u}\|\leq 1;\|\mathbf{v}\|\leq 1}\frac{1}{N\lambda}\sum_{i=1}^{N}\langle Y_i-\langle UV^T,X_i\rangle,\mathbf{u}^TX_i\mathbf{v}\rangle \\
&= \sup_{\|\mathbf{u}\|\leq 1;\|\mathbf{v}\|\leq 1}\frac{1}{N\lambda}\langle \mathbf{v},\sum_{i=1}^{N}(Y_i-\langle UV^T,X_i\rangle)^T\mathbf{u}^TX_i\rangle \\
&= \sup_{\|\mathbf{u}\|\leq 1}\frac{1}{N\lambda}\|\sum_{i=1}^{N}(Y_i-\langle UV^T,X_i\rangle)\mathbf{u}^TX_i\| \\
&= \sup_{\|\mathbf{u}\|\leq 1}\frac{1}{N\lambda}\|\sum_{i=1}^{N}(Y_i-\langle UV^T,X_i\rangle)X_i^T\mathbf{u}\| \\
&= \frac{1}{\lambda}\|\frac{1}{N}\sum_{i=1}^{N}(Y_i-\langle UV^T,X_i\rangle)X_i\|_2.
\end{aligned}$$

Defining $\mathcal{F}_\theta$: Next, we estimate the relevant constants. First we estimate the constants from Assumption 4, suppose that $\mathcal{B}:=\{(\mathbf{u},\mathbf{v}):\|\mathbf{u}\|_2\leq 1,\|\mathbf{v}\|_2\leq 1\}$

$$\mathcal{F}_\theta := \{\mathbf{u}\mathbf{v}^T:\|\mathbf{u}\|_2\|\mathbf{v}\|_2\leq 1\}\cap\mathcal{B} \tag{213}$$

Estimating L_ϕ: The Lipschitz constant L_ϕ in the function $\mathcal{F}_\theta$ is $L_\phi=\sup_{(\mathbf{u},\mathbf{v})\in\mathcal{F}_\theta}\|\langle\mathbf{u}\mathbf{v}^T,.\rangle\|_{\text{Lip}}=1$.

Estimating r_θ: We have that from A.M-G.M inequality,

$$\|\mathbf{u}\|_2\|\mathbf{v}\|_2\leq\frac{1}{2}[\|\mathbf{u}\|_2^2+\|\mathbf{v}\|_2^2]. \tag{214}$$

Now for any $(\mathbf{u},\mathbf{v})\in\mathcal{F}_\theta$ we have that $0.5[\|\mathbf{u}\|_2+\|\mathbf{v}\|_2]\leq 1$. Therefore, $\forall(\mathbf{u},\mathbf{v})\in\mathcal{F}_\theta$ we have

$$\|\mathbf{u}\|_2\|\mathbf{v}\|_2\leq\frac{1}{2}[\|\mathbf{u}\|_2^2+\|\mathbf{v}\|_2^2]\leq\sqrt{\frac{1}{2}[\|\mathbf{u}\|_2^2+\|\mathbf{v}\|_2^2]} \tag{215}$$

Then we need that $\mathcal{F}_\theta\subseteq\mathbb{B}(r_\theta)$, then must be $r_\theta=\frac{1}{\sqrt{2}}$.

Defining $\mathcal{F}_\mathcal{W}$: From the corollary's assumptions we have that, $\mathcal{B}_R:=\{(\mathbf{u},\mathbf{v}):\|\mathbf{u}\|_2\leq B_u,\|\mathbf{v}\|\leq B_v\}$; our hypothesis class is defined as

$$\mathcal{F}_\mathcal{W}:=\left\{\{(\mathbf{u}_j,\mathbf{v}_j)\}:\|\langle UV^T,\cdot\rangle\|_{\text{Lip}}=\|UV^T\|_2\leq\gamma\right\}\cap\mathcal{B}_R. \tag{216}$$

As $\gamma\geq\Omega(f_\mu^*)L_\phi$, we choose $\gamma=C_{UV}\|M^*\|_*$ for $C_{UV}\geq 1$. We have that

$$\mathcal{F}_\mathcal{W}=\left\{\{(\mathbf{u}_j,\mathbf{v}_j)\}:\|\langle UV^T,\cdot\rangle\|_{\text{Lip}}=\|UV^T\|_2\leq C_{UV}\|M^*\|_*,\|\mathbf{u}_j\|\leq B_u,\|\mathbf{v}_j\|\leq B_v\right\}. \tag{217}$$

Estimating ϵ_0**:** From the data generating mechanism we have $\|g\|_{\text{Lip}}=\|M^*\|_2$, $\sigma_X=1$, $\sigma_{Y|X}=\sigma$. Then we have the following constants from Theorem 4:

$$\epsilon_0=16\gamma^2\sigma_X^2\min\left\{1,\frac{L}{4}\left[1+\frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1+\frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}, \tag{218}$$

which evaluates to

$$\epsilon_0=16C_{UV}^2\|M^*\|_*^2\min\left\{1,\frac{1+\sigma^2}{4C_{UV}^2}\right\}. \tag{219}$$

From the corollary assumption we have that $C_{UV} \leq 0.5\sqrt{1+\sigma^2}$, which implies that

$$\epsilon_0 = 4(1+\sigma^2)\|M^*\|_*^2. \tag{220}$$

Estimating ϵ_1: Similarly, we evaluate

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}, \tag{221}$$

obtaining

$$\epsilon_1 = 16C_{UV}^2\|M^*\|_*^2 \max\left\{1, \frac{1+\sigma^2}{4C_{UV}^2}\right\}. \tag{222}$$

From corollary assumption we have that $C_{UV} \leq 0.5\sqrt{1+\sigma^2}$ which gives

$$\epsilon_1 = 16C_{UV}^2\|M^*\|_*^2. \tag{223}$$

Defining convex set $\mathcal{C}$: Consider a convex set $\mathcal{C} = \mathbb{B}(g) = \{X : \|\mathsf{vec}(X)\|_2 \leq g\}$.

First and foremost we need to estimate $\delta_{\mathcal{C}}$ for the following inequality to hold:

$$P(\cap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_{\mathcal{C}}. \tag{224}$$

The probability of $X \in \mathcal{C} = \mathbb{B}(g)$ is equivalent to saying the probability of the event when $\|\mathsf{vec}(X)\|_2 \leq g$. Since, $X_{ij} \sim \mathcal{N}(0, 1/(m \times n))$ as a consequence of Bernstein's Inequality (Vershynin, 2018, Corollary 2.8.3) we have that for any $t \geq 0$,

$$P(|\|\mathsf{vec}(X)\|_2 - 1| \leq t) \geq 1 - 2\exp\left(-cn_X t^2\right) \tag{225}$$

for some constant $c \geq 0$. Now we have

$$P(\|\mathsf{vec}(X)\|_2 \leq g) \begin{cases} \geq 1 - 2\exp\left(-cn_X(g-1)^2\right) & \text{if } g \geq 1 \\ \leq 2\exp\left(-cn_X(g-1)^2\right) & \text{otherwise.} \end{cases} \tag{226}$$

We consider the case where $g \geq 1$, then we have that

$$P(\cap_{i=1}^N X_i \in \mathcal{C}) = P(\cap_{i=1}^N \|\mathsf{vec}(X)\|_2 \leq g) \geq 1 - \underbrace{2N\exp\left(-cn_X(g-1)^2\right)}_{=\delta_{\mathcal{C}}}. \tag{227}$$

We have that $\delta_{\mathcal{C}} = 2N\exp\left(-cn_X(g-1)^2\right)$.

Now we evaluate $B_\ell, B_\Phi, \tilde{L}_\Phi, \tilde{L}_\phi$.

Estimating B_Φ: Recall that $r_\theta = \frac{1}{\sqrt{2}}$. Then we have

$$\begin{aligned}
B_\Phi &= \sup_{Z\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\langle UV^T, Z\rangle\| && (228)\\
&= \sup_{Z\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\langle \mathsf{vec}(UV^T), \mathsf{vec}(Z)\rangle\| && (229)\\
&= g \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\mathsf{vec}(UV^T)\|_2 && (230)\\
&= g \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\sum_{j=1}^R \mathsf{vec}(\mathbf{u}_j\mathbf{v}_j^T)\|_2 && (231)\\
&= gR \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\mathsf{vec}(\mathbf{u}_j\mathbf{v}_j^T)\|_2 && (232)\\
&= gR \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\mathbf{u}_j\mathbf{v}_j^T\|_F && (233)\\
&= gR \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|\mathbf{u}_j\|_2\|\mathbf{v}_j\|_2 && (234)\\
&= gB_uB_vR. && (235)
\end{aligned}$$

Estimating B_ℓ: Similarly, we have

$$\begin{aligned} B_\ell &= \sup_{Z\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_\mathcal{W}} \|\langle UV^T - M^*, Z\rangle\| && (236)\\ &= g \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_\mathcal{W}} \|\mathsf{vec}(UV^T - M^*)\|_2 && (237)\\ &\le g[\|M^*\|_F + B_u B_v R]. && (238)\end{aligned}$$

Estimating $\tilde{L}_\Phi$: Now, we compute the Lipschitz constant with respect to U, V. We have that

$$\begin{aligned} \tilde{L}_\Phi &= \sup_{Z\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_\mathcal{W}} \frac{\|\langle UV^T - U'V'^T, Z\rangle\|}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (239)\\ &= g \sup_{(U,V),(U',V')\in\mathcal{F}_\mathcal{W}} \frac{\|UV^T - U'V'^T\|_F}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (240)\\ &= gR \sup_{(U,V),(U',V')\in\mathcal{F}_\mathcal{W}} \frac{\|\mathbf{u}_j\mathbf{v}_j^T - \mathbf{u}'_j\mathbf{v}'^T_j\|_F}{\sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (241)\\ &= gR \sup_{(U,V),(U',V')\in\mathcal{F}_\mathcal{W}} \frac{\|(\mathbf{u}_j - \mathbf{u}'_j)\mathbf{v}_j^T - \mathbf{u}'_j(\mathbf{v}'_j - \mathbf{v}_j)^T\|_F}{\sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (242)\\ &\le gR \sup_{(U,V),(U',V')\in\mathcal{F}_\mathcal{W}} \frac{\|(\mathbf{u}_j - \mathbf{u}'_j)\|_2\|\mathbf{v}_j\|_2 + \|\mathbf{u}'_j\|_2\|(\mathbf{v}'_j - \mathbf{v}_j)\|_2}{\sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (243)\\ &\le gR \sup_{(U,V),(U',V')\in\mathcal{F}_\mathcal{W}} \sqrt{\|\mathbf{v}_j\|_2^2 + \|\mathbf{u}'_j\|_2^2} && (244)\\ &= g\sqrt{B_u^2 + B_v^2}R. && (245)\end{aligned}$$

Estimating $\tilde{L}_\phi$: Similarly we get $\tilde{L}_\phi = g\sqrt{B_u^2 + B_v^2}$.

Estimating ϵ_2: Recall that

$$\epsilon_2 = \max\{8B_\ell\tilde{L}_\Phi, 8\tilde{L}_\Phi[B_\ell + B_\Phi L], 32\Omega(f_\mu^*)\tilde{L}_\phi \max\{B_\ell, LB_\Phi\}, 4\tilde{L}_\Phi B_\Phi\}. \quad (246)$$

From all the constants computed earlier, we have that

$$\epsilon_2 = k_1 g^2 R^2 (\|M^*\|_F + B_u B_v)^2 \quad (247)$$

for some constant $k_1 \ge 0$.

Next we move on estimating $B(\mathcal{C})$ we need to analyze three terms:
The First Term: We define the first term via

$$T_1 := \sup_{\{W_j\}\in\mathcal{F}_\mathcal{W}} \left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right|.$$

For fixed (U, V), we have

$$\begin{aligned} &\left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| \\ &= \left| \mathbb{E}\left[\langle M^* - UV^T, \mathcal{P}_\mathcal{C}(X)\rangle^2 - \langle M^* - UV^T, X\rangle^2 \right] \right| \\ &= \Bigg| \mathbb{E}\Bigg[\langle \mathsf{vec}(M^* - UV^T)\mathsf{vec}(M^* - UV^T)^T, \mathsf{vec}(\mathcal{P}_\mathcal{C}(X))\mathsf{vec}(\mathcal{P}_\mathcal{C}(X))^T \rangle \\ &- \langle \mathsf{vec}(M^* - UV^T)\mathsf{vec}(M^* - UV^T)^T, vec(X)vec(X)^T \rangle \Bigg] \Bigg| \\ &= \left| \langle \mathsf{vec}(M^* - UV^T)\mathsf{vec}(M^* - UV^T)^T, \mathbb{E}\left[\mathsf{vec}(\mathcal{P}_\mathcal{C}(X))\mathsf{vec}(\mathcal{P}_\mathcal{C}(X))^T - vec(X)vec(X)^T \right] \rangle \right| \end{aligned}$$

From Lemma 4, taking $g \geq 1$,

$$\begin{aligned}\Big|\|f^*_\mu \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|^2_\mu - \|f^*_\mu - \Phi_r(\{W_j\})\|^2_\mu\Big| \\ \leq g e^{-g^2/2} \|\mathsf{vec}(M^* - UV^T)\mathsf{vec}(M^* - UV^T)^T\|_2,\end{aligned} \tag{248}$$

whereupon further simplifying, we obtain

$$\big|\|f^*_\mu \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|^2_\mu - \|f^*_\mu - \Phi_r(\{W_j\})\|^2_\mu\big| \leq g e^{-g^2/2} \|M^* - UV^T\|^2_F. \tag{249}$$

Now, applying triangular inequality and taking the supremum, we obtain

$$T_1 \leq g e^{-g^2/2} (\|M^*\|_F + RB_uB_v)^2. \tag{250}$$

The Second Term: We define the second term via

$$\begin{aligned}T_2 := \sup_{\{W_j\} \in \mathcal{F}_\mathcal{W}, W' \in \mathcal{F}_\theta} \Big| &\langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu \\ &- \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \Big|.\end{aligned} \tag{251}$$

We have

$$\begin{aligned}&\Big|\langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \Big| \\ &\quad = \left|\mathbb{E}\left[\langle UV^T - M^*, \mathcal{P}_\mathcal{C}(X)\rangle \langle \mathbf{u}\mathbf{v}^T, \mathcal{P}_\mathcal{C}(X)\rangle - \langle UV^T - M^*, X\rangle \langle \mathbf{u}\mathbf{v}^T, X\rangle\right]\right| \\ &\quad = \left|\langle \mathsf{vec}(M^* - UV^T)\mathsf{vec}(\mathbf{u}\mathbf{v}^T)^T, \mathbb{E}\left[\mathsf{vec}(X)\mathsf{vec}(X)^T - \mathsf{vec}(\mathcal{P}_\mathcal{C}(X))\mathsf{vec}(\mathcal{P}_\mathcal{C}(X))^T\right]\rangle\right|.\end{aligned}$$

As a consequence of Lemma 4 we have

$$\begin{aligned}&\left|\langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \right| \\ &\leq g e^{-g^2/2} \|\mathsf{vec}(M^* - UV^T)\mathsf{vec}(\mathbf{u}\mathbf{v}^T)^T\|_2 = g e^{-g^2/2} \|M^* - UV^T\|_F \|\mathbf{u}\mathbf{v}^T\|_F.\end{aligned} \tag{252}$$

Now we apply supremum over $(\mathbf{u}, \mathbf{v}) \in \mathcal{F}_\theta$ and then (U, V) obtaining

$$T_2 \leq g e^{-g^2/2} \left[\|M^*\|_F + RB_uB_v\right]. \tag{253}$$

The Third Term: We define

$$\begin{aligned}T_3 := \sup_{\{W_j\} \in \mathcal{F}_\mathcal{W}} \Big| &\langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C} \rangle_\mu \\ &- \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\}) \rangle_\mu \Big|.\end{aligned} \tag{254}$$

Similarly to the earlier item, we rewrite the above as

$$\left|\langle \mathsf{vec}(M^* - UV^T)\mathsf{vec}(UV^T)^T, \mathbb{E}\left[\mathsf{vec}(X)\mathsf{vec}(X)^T - \mathsf{vec}(\mathcal{P}_\mathcal{C}(X))\mathsf{vec}(\mathcal{P}_\mathcal{C}(X))^T\right]\rangle\right| \tag{255}$$

As a consequence of Lemma 4 we have

$$\begin{aligned}&\left|\langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\}) \rangle_\mu \right| \\ &\leq g e^{-g^2/2} \|\mathsf{vec}(M^* - UV^T)\mathsf{vec}(UV^T)^T\|_2 = g e^{-g^2/2} \|M^* - UV^T\|_F \|UV^T\|_F.\end{aligned} \tag{256}$$

Finally, we apply supremum over $(U, V) \in \mathcal{F}_\mathcal{W}$, obtaining

$$T_3 \leq g e^{-g^2/2} B_uB_vR \left[\|M^*\|_F + RB_uB_v\right]. \tag{257}$$

Now combining equations (250), (253), (257) we obtain that

$$B(\mathcal{C}) \leq ge^{-g^2/2}\left[\alpha(\|M^*\|_F + RB_uB_v)^2 + \|M^*\|_F + RB_uB_v + B_uB_vR\left[\|M^*\|_F + RB_uB_v\right]\right] \tag{258}$$

We further upper bound for simplicity as

$$B(\mathcal{C}) \leq 4ge^{-g^2/2}(\|M^*\|_F + RB_uB_v)^2. \tag{259}$$

From Theorem 4 we have that

$$\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \lesssim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \tag{260}$$

$$+ \frac{4}{n_Y}ge^{-g^2/2}\{\|M^*\|_F + RB_uB_v\}^2 + 16C_{UV}^2\|M^*\|_*^2 \times \Bigg(\tag{261}$$

$$\sqrt{\frac{R\mathsf{dim}(\mathcal{W})log\left(C_{UV}\|M^*\|_*k_1g^2R^2(\|M^*\|_F + B_uB_v)^2\frac{1}{\sqrt{2}}\right)\log(N) + \log(1/\delta)}{N}}\Bigg) \tag{262}$$

holds true w.p at least $1 - \delta - 2N\exp\left(-cn_X(g-1)^2\right)$.

Now choose

$$g = 1 + \mathcal{O}\left(\sqrt{\log(\sqrt{NR^{100}}) + \log(1/\delta)}\right). \tag{263}$$

Then we get that

$$\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \tag{264}$$

$$\lesssim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \tag{265}$$

$$+ \frac{4}{n_Y}\{\|M^*\|_F + RB_uB_v\}^2\frac{\delta\sqrt{\log(NR^{100}) + \log(1/\delta)}}{NR^{100}} \tag{266}$$

$$+ 16C_{UV}^2\|M^*\|_*^2\sqrt{\frac{R(m+n)\log\left(C_{UV}\|M^*\|_*k_1[\log(NR)+\log(1/\delta)]R^2(\|M^*\|_F+B_uB_v)^2\frac{1}{\sqrt{2}}\right)\log(N)+\log(1/\delta)}{N}} \tag{267}$$

holds true w.p at least $1-\delta$. Now ignoring *loglog* terms and keeping the right most term because of the dominance, we obtain

$$\frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \lesssim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right]$$

$$+C_{UV}^2\|M^*\|_*^2\sqrt{\frac{Rlog\left(R(C_{UV} + B_uB_v)\right)(m+n)\log(N) + \log(1/\delta)}{N}} \tag{268}$$

holds true w.p at least $1 - \delta$. □

C.2 Structured Matrix Sensing

Next, we move on to a slightly more generalized matrix sensing problem through which we impose certain structure in the factor U. Consider an atomic set $\mathcal{U}$ that represents the set of structured columns, and suppose that U consists of columns that are affine combinations of the atoms in $\mathcal{U}$. We consider a gauge function $\gamma_{\mathcal{U}}(\cdot)$ which is defined via

$$\gamma_{\mathcal{U}}(\mathbf{u}) := \inf\{t, t \geq 0 \text{ such that } \mathbf{u} \in tconv(\mathcal{U})\} \tag{269}$$

For instance, $\mathcal{U}$ can be the intersection of L_2 unit ball and L_1 unit ball, which induces UV^T to be low-rank and U to be sparse. Imposing such structures has been well studied for convex problems by Chandrasekaran et al. (2012). Bach (2013) analyzed such structures for non-convex matrix factorization problems. However, their work was focused primarily on the optimization guarantees whereas our result below provides generalization/recovery guarantees for structured matrix sensing problems. We have the following corollary.

Corollary 5 (Structured matrix sensing)**.** *Consider the true model for (X, y), where $X \in \mathbb{R}^{m\times n}$ is a random matrix with i.i.d. entries $X_{lk} \sim \mathcal{N}(0, \frac{1}{mn})$ and $y = \langle U^* V^{*T}, X\rangle + \epsilon$, where $U^* \in \mathbb{R}^{m\times R^*}$, $V^* \in \mathbb{R}^{n\times R^*}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent from X. For all $i \in [N]$, let (X_i, y_i) be i.i.d. samples from this true model. Consider the estimator $\hat{y} = \langle UV^T, X\rangle$, where $U \in \mathbb{R}^{m\times R}$ and $V \in \mathbb{R}^{n\times R}$. Let $\delta \in (0, 1]$ be fixed. Define the non-convex problem with the atomic set, $\mathcal{U}$*

$$\mathsf{NC}^{\mathsf{SMS}}_{\mu_N}((U, V)) := \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \langle UV^T, X_i\rangle\right)^2 + \lambda\sum_{j=1}^{R}\gamma_{\mathcal{U}}(\mathbf{u}_j)\|\mathbf{v}_j\|_2, \tag{270}$$

and define $\mathsf{NC}^{SMS}_{\mu}((U, V))$ similarly with the sum over i replaced by expectation taken over (X, y). Here $\gamma_{\mathcal{U}}(\mathbf{u}) := \inf\{t; t \geq 0, \mathbf{u} \in tconv(\mathcal{U})\}$ for some specified atomic set, $\mathcal{U}$. Define

$$K_1 := \sum_{j=1}^{r^*}\gamma_{\mathcal{U}}(\mathbf{u}_j^*)\|\mathbf{v}_j^*\|_2; K_2 := \sup_{\|\mathbf{u}\|\leq 1}\gamma_{\mathcal{U}}(\mathbf{u}). \tag{271}$$

Let (U, V) be a stationary point of $\mathsf{NC}^{\mathsf{SMS}}_{\mu_N}((U, V))$. Suppose there exists $C_{UV}, B_u, B_v > 0$ such that $\|UV^T\|_2 \leq C_{UV}K_1$, and for all $j \in [R]$, $\|\mathbf{u}_j\|_2 \leq B_u$, $\|\mathbf{v}_j\|_2 \leq B_v$. Then, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}\left|\mathsf{NC}^{\mathsf{SMS}}_{\mu}((U, V))) - \mathsf{NC}^{\mathsf{SMS}}_{\mu_N}((U, V))\right| &\lesssim K_1\left[K_2\|\frac{1}{N}\sum_{i=1}^{N}(y_i - \langle UV^T, X_i\rangle)X_i\|_2 - \lambda\right] \\ &+ C_{UV}^2K_1^2\sqrt{\frac{R\log\left(R(C_{UV}+B_uB_v)\right)(m+n)\log(N)+\log(1/\delta)}{N}}.\end{aligned} \tag{272}$$

Remarks: Similar to matrix sensing, the sample complexity required for consistency is only that $N \gtrsim R(m+n)$ up to logarithmic terms, assuming a global minimum is found. The sample complexity is similar to low-rank matrix sensing (ignoring the scale and logarithmic dependency). To the best of our knowledge, this problem has not been studied from a statistical perspective, and our sample complexities match the corresponding convex slightly structured matrix sensing of Kakade et al. (2008). Unlike low-rank matrix sensing, the main technical challenge is to compute the polar/supremum term in the optimization error. In general, such a computation is NP-hard when the atomic set $\mathcal{U}$ has non-negative atoms (Hendrickx and Olshevsky, 2010).

Proof. The proof is similar to that of Corollary 1, except for the computation of the polar. Therefore, we only compute the polar.

Estimating $\Omega(f_\mu^*)$: Since M^* is the true matrix the globally optimal solution would be M^*; therefore, from Proposition 2 we have

$$\Omega(f_\mu^*) \leq \left(\sum_{j=1}^{r^*}\gamma_{\mathcal{U}}(\mathbf{u}_j^*)\|\mathbf{v}_j^*\|_2\right). \tag{273}$$

Estimating $\Omega^\circ_{\mu_N}(\cdot)$: Now we move on to compute the polar.

$$\begin{aligned}
\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))\right) &= \Omega^\circ_{\mu_N}\left(\frac{1}{\lambda}(g-\Phi_r(\{W_j\}))\right)\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))\right)\\
&= \Omega^\circ_{\mu_N}\left(\frac{1}{\lambda}(g-\Phi_r(\{W_j\}))\right)\\
&= \sup_{\gamma_{\mathcal{U}}(\mathbf{u})\le 1;\|\mathbf{v}\|\le 1}\frac{1}{N\lambda}\sum_{i=1}^N\langle Y_i-\langle UV^T,X_i\rangle,\mathbf{u}^TX_i\mathbf{v}\rangle\\
&= \sup_{\gamma_{\mathcal{U}}(\mathbf{u})\le 1;\|\mathbf{v}\|\le 1}\frac{1}{N\lambda}\langle\mathbf{v},\sum_{i=1}^N(Y_i-\langle UV^T,X_i\rangle)^T\mathbf{u}^TX_i\rangle\\
&= \sup_{\gamma_{\mathcal{U}}(\mathbf{u})\le 1}\frac{1}{N\lambda}\|\sum_{i=1}^N(Y_i-\langle UV^T,X_i\rangle)\mathbf{u}^TX_i\|\\
&= \sup_{\gamma_{\mathcal{U}}(\mathbf{u})\le 1}\frac{1}{N\lambda}\|\sum_{i=1}^N(Y_i-\langle UV^T,X_i\rangle)X_i^T\mathbf{u}\|.
\end{aligned}$$

This yields

$$\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))\right)\le\left[\sup_{\|\mathbf{u}\|\le 1}\gamma_{\mathcal{U}}(\mathbf{u})\right]\frac{1}{N\lambda}\|\sum_{i=1}^N(Y_i-\langle UV^T,X_i\rangle)X_i^T\|_2. \tag{274}$$

The rest of the proof is the same as that of low-rank matrix sensing (see section C.1). □

C.3 Two-Layer Linear NN

Next, we consider the closely related problem of 2-Layer Linear Neural Networks, which is essentially a multi-dimensional matrix sensing problem; this is also referred to as non-convex linear regression. In practice, this approach has seemed to have better linear convergence (Arora et al., 2019) and generalization capabilities (Allen-Zhu et al., 2019) than vanilla linear regression. Corollary 6 provides generalization error upper bounds.

Corollary 6 (2-Layer Linear Neural Network)**.** *Consider the true model for* $(\mathbf{x},\mathbf{y})$*, where* $\mathbf{x}\sim\mathcal{N}(0,(1/n)I_n)\in\mathbb{R}^n$*,* $\mathbf{y}=U^*V^{*T}\mathbf{x}+\epsilon$*, where* $U^*\in\mathbb{R}^{m\times R^*}$*,* $V^*\in\mathbb{R}^{n\times R^*}$*, and* $\epsilon\sim\mathcal{N}(0,(\sigma^2/m)I_m)\in\mathbb{R}^m$ *independent from* $\mathbf{x}$*. For all* $i\in[N]$*, let* $(\mathbf{x}_i,\mathbf{y}_i)$ *be i.i.d. samples from this true model. Consider the estimator* $\hat{\mathbf{y}}=UV^T\mathbf{x}$*, where* $U\in\mathbb{R}^{m\times R},V\in\mathbb{R}^{n\times R}$*. Let* $\delta\in(0,1]$ *be fixed. Define the non-convex problem*

$$\mathsf{NC}^{\mathsf{2LNN}}_{\mu_N}((U,V)):=\frac{1}{2N}\sum_{i=1}^N\|\mathbf{y}_i-U[V^T\mathbf{x}_i]_+\|_2^2+\frac{\lambda}{2}\left(\|U\|_F^2+\|V\|_F^2\right), \tag{275}$$

and define $\mathsf{NC}^{\mathsf{2LNN}}_{\mu}((U,V))$ *similarly with the sum over* i *replaced by expectation taken over* $(\mathbf{x},\mathbf{y})$*.*

Let (U,V) *be a stationary point of* $\mathsf{NC}^{\mathsf{2LNN}}_{\mu_N}((U,V))$*. Suppose there exists* $C_{UV},B_u,B_v>0$ *such that* $\|UV^T\|_2\le C_{UV}\left[\|U^*\|_F^2+\|V^*\|_F^2\right]$*, and for all* $j\in[R]$*,* $\|\mathbf{u}_j\|_2\le B_u$*,* $\|\mathbf{v}_j\|_2\le B_v$*. Then, with probability at least* $1-\delta$*, it holds that*

$$\begin{aligned}
\frac{1}{m}\left|\mathsf{NC}^{\mathsf{2LNN}}_{\mu}((U,V))-\mathsf{NC}^{\mathsf{2LNN}}_{\mu_N}((U,V))\right| &\lesssim \frac{1}{2m}\left[\|U^*\|_F^2+\|V^*\|_F^2\right]\left[\frac{1}{N}\sum_{i=1}^N\|\mathbf{y}_i-\hat{\mathbf{y}}_i\|_2\|\mathbf{x}_i\|_2-\lambda\right]\\
&C_{UV}^2\left[\|U^*\|_F^2+\|V^*\|_F^2\right]^2\sqrt{\frac{R\log\left(R\left(C_{UV}+B_u^2+B_v^2\right)\right)(m+n)\log(N)+\log(1/\delta)}{N}}.
\end{aligned}$$

Similar to matrix sensing, we require that $N\gtrsim R(m+n)$, with $\frac{R(m+n)}{N}\to 0$ for consistency at a global minimum. This matches classical results for (convex) linear regression.

Proof. To obtain a generalization bound from Theorem 4 for this setting, we set the following problem parameters:

$$\ell(Y,\hat{Y})=\frac{1}{2}\|Y-\hat{Y}\|\implies(\alpha,L)=(0,1); \tag{276}$$

$$\phi(W) = \langle \mathbf{v}, \mathbf{x} \rangle \mathbf{u}; \tag{277}$$

$$\theta(W) = \frac{1}{2}\left[\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2\right]. \tag{278}$$

Estimating $\Omega(f_\mu^*)$: From Proposition 2 we have that

$$\Omega(f_\mu^*) \leq \frac{\|U^*\|_F^2 + \|V^*\|_F^2}{2} \tag{279}$$

Choosing $\mathcal{F}_\theta$:

$$\mathcal{F}_\theta := \{(\mathbf{u}, \mathbf{v}) : \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 \leq 2, \|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1\}. \tag{280}$$

Estimating L_ϕ:

$$L_\phi = \sup_{(\mathbf{u},\mathbf{v})\in\mathcal{F}_\theta} \|\mathbf{u}\mathbf{v}^T(\cdot)\|_{\text{Lip}} = \sup_{(\mathbf{u},\mathbf{v})\in\mathcal{F}_\theta} \|\mathbf{u}\mathbf{v}^T\|_2 = 1. \tag{281}$$

Estimating r_θ: For any $(\mathbf{u}, \mathbf{v}) \in \mathcal{F}_\theta$, we have that,

$$\frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2}{2} \leq \sqrt{\frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2}{2}} \implies \mathcal{F}_\theta \subseteq \mathbb{B}(1/\sqrt{2}). \tag{282}$$

Then we have $r_\theta = 1/\sqrt{2}$.

Choosing $\mathcal{F}_\mathcal{W}$:

$$\mathcal{F}_\mathcal{W} := \{(U, V) : \|UV^T\|_2 \leq \gamma, \|\mathbf{u}_j\| \leq B_u, \|\mathbf{v}_j\| \leq B_v\}. \tag{283}$$

As $\gamma \geq \Omega(f_\mu^*)L_\phi = \frac{\|U^*\|_F^2 + \|V^*\|_F^2}{2}$, we may take $\gamma = C_{UV}\left[\frac{\|U^*\|_F^2 + \|V^*\|_F^2}{2}\right]$ for some C_{UV}. We have that

$$\mathcal{F}_\mathcal{W} = \left\{\{(\mathbf{u}_j, \mathbf{v}_j)\} : \|\langle UV^T, \cdot\rangle\|_{\text{Lip}} = \|UV^T\|_2 \leq C_{UV}\left[\frac{\|U^*\|_F^2 + \|V^*\|_F^2}{2}\right], \|\mathbf{u}_j\| \leq B_u, \|\mathbf{v}_j\| \leq B_v\right\}. \tag{284}$$

Estimating ϵ_0: From the data generating mechanism we have $\|g\|_{\text{Lip}} = \|M^*\|_2$, $\sigma_X = 1$, $\sigma_{Y|X} = \sigma$, which yields the following constants from Theorem 4:

$$\epsilon_0 = 16\gamma^2\sigma_X^2 \min\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}, \tag{285}$$

which evaluates to when $C_{UV} \leq 0.5\sqrt{(1+\sigma^2)}$

$$\epsilon_0 = 8C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\min\left\{1, \frac{1+\sigma^2}{4C_{UV}^2}\right\} = 2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]. \tag{286}$$

Estimating ϵ_1: Similarly, we evaluate

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}, \tag{287}$$

obtaining

$$\epsilon_1 = 8C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]. \tag{288}$$

Choosing the convex set $\mathcal{C}$: Consider a convex set $\mathcal{C} = \mathbb{B}(g) = \{X : \|\mathsf{vec}(X)\|_2 \leq g\}$.

First and foremost we need to estimate $\delta_\mathcal{C}$ for the following inequality to hold:

$$P(\cap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_\mathcal{C}. \tag{289}$$

The probability of $\mathbf{x} \in \mathcal{C} = \mathbb{B}(g)$ is equivalent to saying the probability of the event when $\|\mathbf{x}\|_2 \leq g$. Since, $x_i \sim \mathcal{N}(0, 1/n)$ as a consequence of Bernstein's Inequality (Vershynin, 2018, Corollary 2.8.3) we have that for any $t \geq 0$,

$$P(|\|\mathbf{x}\|_2 - 1| \leq t) \geq 1 - 2\exp\left(-cn_X t^2\right) \tag{290}$$

for some constant $c \geq 0$. Now we have

$$P(\|\mathbf{x}\|_2 \leq g) \begin{cases} \geq 1 - 2\exp\left(-cn_X(g-1)^2\right) & \text{if } g \geq 1 \\ \leq 2\exp\left(-cn_X(g-1)^2\right) & \text{otherwise} \end{cases} \tag{291}$$

We consider the case where $g \geq 1$, then we have that

$$P(\cap_{i=1}^N X_i \in \mathcal{C}) = P(\cap_{i=1}^N \|\mathbf{x}\|_2 \leq g) \geq 1 - \underbrace{2N\exp\left(-cn_X(g-1)^2\right)}_{=\delta_{\mathcal{C}}}. \tag{292}$$

We have that $\delta_{\mathcal{C}} = 2N\exp\left(-cn(g-1)^2\right)$.

Now we evaluate $B_\ell, B_\Phi, \tilde{L}_\Phi, \tilde{L}_\phi$.

Estimating B_Φ: We have

$$\begin{aligned} B_\Phi &= \sup_{\mathbf{z}\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|UV^T\mathbf{z}\| && (293)\\ &= g \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|UV^T\|_2 && (294)\\ &= g\gamma && (295) \end{aligned}$$

Estimating B_ℓ: Similarly, we have

$$\begin{aligned} B_\ell &= \sup_{\mathbf{z}\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|(UV^T - U^*V^{*T})\mathbf{z}\| && (296)\\ &= g \sup_{\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|UV^T - U^*V^{*T}\|_2 && (297)\\ &= g(\|U^*V^{*T}\|_2 + \gamma). && (298) \end{aligned}$$

Estimating $\tilde{L}_\Phi$: Now, we compute the Lipschitz constant with respect to U, V. We have

$$\begin{aligned} \tilde{L}_\Phi &= \sup_{\mathbf{z}\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|(UV^T - U'V'^T)\mathbf{z}\|}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (299)\\ &= g \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|UV^T - U'V'^T\|_2}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (300)\\ &\leq g \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|UV^T - U'V'^T\|_F}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (301)\\ &= g \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|\sum_{j=1}^R \mathbf{u}_j\mathbf{v}_j^T - \mathbf{u}'_j\mathbf{v}'^T_j\|_F}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (302)\\ &= gR \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|\mathbf{u}_j\mathbf{v}_j^T - \mathbf{u}'_j\mathbf{v}'^T_j\|_F}{\sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (303)\\ &= gR \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|(\mathbf{u}_j - \mathbf{u}'_j)\mathbf{v}_j^T - \mathbf{u}'_j(\mathbf{v}'_j - \mathbf{v}_j)^T\|_2}{\sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (304)\\ &\leq gR \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|(\mathbf{u}_j - \mathbf{u}'_j)\|_2\|\mathbf{v}_j\|_2 + \|\mathbf{u}'_j\|_2\|(\mathbf{v}'_j - \mathbf{v}_j)\|_2}{\sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (305)\\ &= gR \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \sqrt{\|\mathbf{v}_j\|_2^2 + \|\mathbf{u}'_j\|_2^2} && (306)\\ &= g\sqrt{B_u^2 + B_v^2}R. && (307) \end{aligned}$$

Estimating $\tilde{L}_\phi$: Similarly we get $\tilde{L}_\phi = g\sqrt{B_u^2 + B_v^2}$.

Estimating ϵ_2: Recall that

$$\epsilon_2 = \max\{8B_\ell \tilde{L}_\Phi, 8\tilde{L}_\Phi[B_\ell + B_\Phi L], 32\Omega(f_\mu^*)\tilde{L}_\phi \max\{B_\ell, LB_\Phi\}, 4\tilde{L}_\Phi B_\Phi\}. \tag{308}$$

From all the constants computed earlier, we have that

$$\epsilon_2 = k_1 g^2 R^2 C_{UV}^2 (\|U^*\|_F^2 + \|V^*\|_F^2)\sqrt{B_u^2 + B_u^2} \tag{309}$$

for some constant $k_1 \geq 0$.

Next, we move on to estimating $B(\mathcal{C})$. We need to analyze three terms:
The First Term: Define

$$T_1 := \sup_{\{W_j\} \in \mathcal{F}_\mathcal{W}} \left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right|$$

We have

$$\left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| = \tag{310}$$

$$= \left| \mathbb{E}\left[\|(U^*V^{*T} - UV^T)\mathcal{P}_\mathcal{C}(\mathbf{x})\|^2 - \|(U^*V^{*T} - UV^T)\mathbf{x}\|^2 \right] \right| \tag{311}$$

$$\begin{aligned} = \Big| \mathbb{E}\Big[&\langle (U^*V^{*T} - UV^T)(U^*V^{*T} - UV^T)^T, (\mathcal{P}_\mathcal{C}(\mathbf{x}))(\mathcal{P}_\mathcal{C}(\mathbf{x}))^T \rangle \\ &- \langle (U^*V^{*T} - UV^T)(U^*V^{*T} - UV^T)^T, \mathbf{x}\mathbf{x}^T \rangle \Big] \Big| \end{aligned} \tag{312}$$

$$= \left| \langle (U^*V^{*T} - UV^T)(U^*V^{*T} - UV^T)^T, \mathbb{E}\left[(\mathcal{P}_\mathcal{C}(\mathbf{x}))(\mathcal{P}_\mathcal{C}(\mathbf{x}))^T - \mathbf{x}\mathbf{x}^T \right] \rangle \right| \tag{313}$$

From Lemma 4 we obtain that (taking $g \geq 1$)

$$\begin{aligned} &\left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| \\ &\leq g e^{-g^2/2} \|(U^*V^{*T} - UV^T)(U^*V^{*T} - UV^T)^T\|_2, \end{aligned} \tag{314}$$

on further simplifying, we get

$$\left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| \leq g e^{-g^2/2} \|U^*V^{*T} - UV^T\|_2^2. \tag{315}$$

Now applying triangular inequality and taking the supremum, we obtain

$$T_1 \leq g e^{-g^2/2} (\|U^*V^{*T}\|_2 + \gamma)^2, \tag{316}$$

The Second Term: Define

$$\begin{aligned} T_2 := \sup_{\{W_j\} \in \mathcal{F}_\mathcal{W}, W' \in \mathcal{F}_\theta} \Big| &\langle \nabla_{\hat{Y}} \ell\left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu \\ &- \langle \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \Big|. \end{aligned} \tag{317}$$

We have

$$\left| \langle \nabla_{\hat{Y}} \ell\left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \right| =$$

$$\left| \mathbb{E}\left[\langle (UV^T - U^*V^{*T})\mathcal{P}_\mathcal{C}(\mathbf{x}), \mathbf{u}\mathbf{v}^T \mathcal{P}_\mathcal{C}(\mathbf{x}) \rangle - \langle (UV^T - U^*V^{*T})\mathbf{x}, \mathbf{u}\mathbf{v}^T \mathbf{x} \rangle \right] \right| \tag{318}$$

is the same as

$$= \left| \langle (U^*V^{*T} - UV^T)(\mathbf{u}\mathbf{v}^T)^T, \mathbb{E}\left[\mathbf{x}\mathbf{x}^T - (\mathcal{P}_\mathcal{C}(\mathbf{x}))(\mathcal{P}_\mathcal{C}(\mathbf{x}))^T \right] \rangle \right| \tag{319}$$

As a consequence of Lemma 4 we have

$$\left| \langle \nabla_{\hat{Y}} \ell\left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell\left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \right|$$

$$\leq ge^{-g^2/2}\|(U^*V^{*T} - UV^T)(\mathbf{u}\mathbf{v}^T)^T\|_2 = ge^{-g^2/2}\|U^*V^{*T} - UV^T\|_F\|\mathbf{u}\mathbf{v}^T\|_F. \tag{320}$$

Now we apply supremum over $(\mathbf{u}, \mathbf{v}) \in \mathcal{F}_\theta$, and then over $(U, V) \in \mathcal{F}_\mathcal{W}$, yielding

$$T_2 \leq ge^{-g^2/2}\left[\|U^*V^{*T}\|_2 + \gamma\right]. \tag{321}$$

The Third Term: Define

$$\begin{aligned} T_3 := \sup_{\{W_j\}\in\mathcal{F}_\mathcal{W}} \Big| &\langle \nabla_{\hat{Y}}\ell\left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\rangle_\mu \\ &- \langle \nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu \Big|. \end{aligned} \tag{322}$$

Similarly to the earlier item, we rewrite the above as

$$= \left|\langle (U^*V^{*T} - UV^T)(UV^T)^T, \mathbb{E}\left[\mathbf{x}\mathbf{x}^T - (\mathcal{P}_\mathcal{C}(\mathbf{x}))(\mathcal{P}_\mathcal{C}(\mathbf{x}))^T\right]\rangle\right| \tag{323}$$

As a consequence of Lemma 4 we have

$$\begin{aligned} &\left|\langle \nabla_{\hat{Y}}\ell\left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\rangle_\mu - \langle \nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu\right| \\ &\leq ge^{-g^2/2}\|(U^*V^{*T} - UV^T)(UV^T)^T\|_2 \leq ge^{-g^2/2}\|U^*V^{*T} - UV^T\|_2\|UV^T\|_2. \end{aligned} \tag{324}$$

Finally, we apply supremum over $(U, V) \in \mathcal{F}_\mathcal{W}$, obtaining

$$T_3 \leq ge^{-g^2/2}\gamma\left[\|U^*V^{*T}\|_2 + \gamma\right]. \tag{325}$$

Now combining equations (316), (321), (325) we obtain that

$$B(\mathcal{C}) \leq ge^{-g^2/2}\left[\alpha(\|U^*V^{*T}\|_2 + \gamma)^2 + \|U^*V^{*T}\|_2 + \gamma + \gamma\left[\|U^*V^{*T}\|_2 + \gamma\right]\right]. \tag{326}$$

We further upper bound for simplicity as

$$B(\mathcal{C}) \leq ge^{-g^2/2}(1+\gamma)(\|U^*V^{*T}\|_2 + \gamma). \tag{327}$$

From Theorem 4 we have that

$$\begin{aligned} \frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| &\lesssim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \\ &+ \frac{4}{n_Y}ge^{-g^2/2}(1+\gamma)(\|U^*V^{*T}\|_2 + \gamma) + 8C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]^2 \times \Bigg(\\ &\sqrt{\frac{R(m+n)\log\left(k_1g^2R^2C_{UV}^2(\|U^*\|_F^2 + \|V^*\|_F^2)\sqrt{B_u^2 + B_u^2}C_{UV}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\right)\log(N) + \log(1/\delta)}{N}}\Bigg) \end{aligned} \tag{328}$$

holds true w.p at least $1 - \delta - 2N\exp\left(-cn_X(g-1)^2\right)$.

Now choose

$$g = 1 + \mathcal{O}\left(\sqrt{\log(N) + \log(1/\delta)}\right). \tag{329}$$

Now ignoring *loglog* terms and keep the right most term because of the dominance,

$$\begin{aligned} \frac{1}{n_Y}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| &\lesssim \frac{\lambda}{n_Y}\Omega(f_\mu^*)\left[\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right)\right) - 1\right] \\ &+ C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]^2\sqrt{\frac{Rlog\left(R\left(C_{UV} + B_u^2 + B_v^2\right)\right)(m+n)\log(N) + \log(1/\delta)}{N}} \end{aligned} \tag{330}$$

holds true w.p at least $1 - \delta$.

□

C.4 Two-Layer ReLU NN

Next, we present and prove the generalization bound for the two-layer ReLU neural network. This is one step ahead of all the linear models that were discussed earlier. Similarly to the Gaussian projections discussed in matrix sensing, we discuss ReLU projection results that will be used in the main proof.

Lemma 5 (ReLU projection 1). *Consider $U_1, U_2 \in \mathbb{R}^{m\times r}$, $V_1, V_2 \in \mathbb{R}^{n\times r}$. Denote, convex set $\mathcal{C} = \mathbb{B}(g)$ that is g-radius hyper sphere, then we have that*

$$\begin{aligned}\Big|\mathbb{E}\Big[\|U_1[V_1^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+ - U_2[V_2^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+\|^2 - \|U_1[V_1^T\mathbf{x}]_+ - U_2[V_2^T\mathbf{x}]_+\|^2\Big]\Big| \\ \leq 2ge^{-g^2/2}[\|U_1\|_F^2\|V_1\|_F^2 + \|U_2\|_F^2\|V_2\|_F^2].\end{aligned} \tag{331}$$

Proof. First, we re-write

$$\|U_1[V_1^T\mathbf{x}]_+ - U_2[V_2^T\mathbf{x}]_+\|^2 = \|\sum_{j=1}^r \mathbf{u}_{j1}[\mathbf{v}_{j1}^T\mathbf{x}]_+ - \mathbf{u}_{j2}[\mathbf{v}_{j2}^T\mathbf{x}]_+\|^2 \tag{332}$$

$$= \|\sum_{j=1}^r \mathbf{u}_{j1}\mathbf{v}_{j1}^T\mathbf{x}\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0} - \mathbf{u}_{j2}\mathbf{v}_{j2}^T\mathbf{x}\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\|^2 \tag{333}$$

$$\begin{aligned}= \sum_{j=1}^r\sum_{l=1}^r \Big[\langle(\mathbf{u}_{j1}\mathbf{v}_{j1}^T)^T(\mathbf{u}_{j1}\mathbf{v}_{j1}^T), \mathbf{x}\mathbf{x}^T\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\rangle + \langle(\mathbf{u}_{j2}\mathbf{v}_{j2}^T)^T(\mathbf{u}_{j2}\mathbf{v}_{j2}^T), \mathbf{x}\mathbf{x}^T\mathbf{1}_{\mathbf{v}_{j2}^T\mathbf{x}\geq 0}\rangle \\ - 2\langle(\mathbf{u}_{j2}\mathbf{v}_{j2}^T)^T(\mathbf{u}_{j1}\mathbf{v}_{j1}^T), \mathbf{x}\mathbf{x}^T\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\mathbf{1}_{\mathbf{v}_{j2}^T\mathbf{x}\geq 0}\rangle\Big].\end{aligned} \tag{334}$$

Note that $\mathbf{1}_{\mathbf{v}^T\mathbf{x}>0} = \mathbf{1}_{\mathbf{v}^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})>0}$. Similarly, we have

$$\begin{aligned}\|U_1[V_1^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+ - U_2[V_2^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+\|^2 = \sum_{j=1}^r\sum_{l=1}^r \Big[\langle(\mathbf{u}_{j1}\mathbf{v}_{j1}^T)^T(\mathbf{u}_{j1}\mathbf{v}_{j1}^T), \mathcal{P}_{\mathcal{C}}(\mathbf{x})\mathcal{P}_{\mathcal{C}}(\mathbf{x})^T\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\rangle \\ + \langle(\mathbf{u}_{j2}\mathbf{v}_{j2}^T)^T(\mathbf{u}_{j2}\mathbf{v}_{j2}^T), \mathcal{P}_{\mathcal{C}}(\mathbf{x})\mathcal{P}_{\mathcal{C}}(\mathbf{x})^T\mathbf{1}_{\mathbf{v}_{j2}^T\mathbf{x}\geq 0}\rangle \\ - 2\langle(\mathbf{u}_{j2}\mathbf{v}_{j2}^T)^T(\mathbf{u}_{j1}\mathbf{v}_{j1}^T), \mathcal{P}_{\mathcal{C}}(\mathbf{x})\mathcal{P}_{\mathcal{C}}(\mathbf{x})^T\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\mathbf{1}_{\mathbf{v}_{j2}^T\mathbf{x}\geq 0}\rangle\Big].\end{aligned} \tag{335}$$

Now computing the difference between equations (334), and (335) we get

$$\begin{aligned}\big|\mathbb{E}\left[\|U_1[V_1^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+ - U_2[V_2^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+\|^2 - \|U_1[V_1^T\mathbf{x}]_+ - U_2[V_2^T\mathbf{x}]_+\|^2\right]\big| = \\ \Big|\sum_{j=1}^r\sum_{l=1}^r \mathbb{E}\Big[\langle(\mathbf{u}_{j1}\mathbf{v}_{j1}^T)^T(\mathbf{u}_{j1}\mathbf{v}_{j1}^T), (\mathcal{P}_{\mathcal{C}}(\mathbf{x})\mathcal{P}_{\mathcal{C}}(\mathbf{x})^T - \mathbf{x}\mathbf{x}^T)\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\rangle \\ + \langle(\mathbf{u}_{j2}\mathbf{v}_{j2}^T)^T(\mathbf{u}_{j2}\mathbf{v}_{j2}^T), (\mathcal{P}_{\mathcal{C}}(\mathbf{x})\mathcal{P}_{\mathcal{C}}(\mathbf{x})^T - \mathbf{x}\mathbf{x}^T)\mathbf{1}_{\mathbf{v}_{j2}^T\mathbf{x}\geq 0}\rangle \\ - 2\langle(\mathbf{u}_{j2}\mathbf{v}_{j2}^T)^T(\mathbf{u}_{j1}\mathbf{v}_{j1}^T), (\mathcal{P}_{\mathcal{C}}(\mathbf{x})\mathcal{P}_{\mathcal{C}}(\mathbf{x})^T - \mathbf{x}\mathbf{x}^T)\mathbf{1}_{\mathbf{v}_{j1}^T\mathbf{x}\geq 0}\mathbf{1}_{\mathbf{v}_{j2}^T\mathbf{x}\geq 0}\rangle\Big]\Big|.\end{aligned} \tag{336}$$

After applying Lemma 4 and the triangular inequality we obtain, for $g \geq 1$,

$$\begin{aligned}\Big|\mathbb{E}\Big[\|U_1[V_1^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+ - U_2[V_2^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+\|^2 - \|U_1[V_1^T\mathbf{x}]_+ - U_2[V_2^T\mathbf{x}]_+\|^2\Big]\Big| \\ \overset{(a)}{\leq} ge^{-g^2/2}\sum_{j=1}^r\Big[\sum_{l=1}^r \|\mathbf{u}_{j1}\|^2\|\mathbf{v}_{j1}\|^2 \\ + 2\|\mathbf{u}_{j1}\|\|\mathbf{u}_{j2}\|\|\mathbf{v}_{j1}\|\|\mathbf{v}_{j2}\| + \|\mathbf{u}_{j2}\|^2\|\mathbf{v}_{j2}\|^2\Big] \\ \overset{(b)}{\leq} 2ge^{-g^2/2}\sum_{j=1}^r\sum_{l=1}^r\Big[\|\mathbf{u}_{j1}\|^2\|\mathbf{v}_{j1}\|^2 + \|\mathbf{u}_{j2}\|^2\|\mathbf{v}_{j2}\|^2\Big] \\ \overset{(c)}{\leq} 2ge^{-g^2/2}[\|U_1\|_F^2\|V_1\|_F^2 + \|U_2\|_F^2\|V_2\|_F^2].\end{aligned} \tag{337}$$

In (a) we apply triangular inequality and apply Lemma 4, (b) we use the identity that $a^2+2ab+b^2=(a+b)^2$, and (c) we use the identity that $(a^2c^2+b^2d^2)\leq(a^2+b^2)(c^2+d^2)$. This completes the proof.

□

Lemma 6 (ReLU projection 2). *Consider $U_1,U_2\in\mathbb{R}^{m\times r}$, $V_1,V_2\in\mathbb{R}^{n\times r}$. Denote the convex set $\mathcal{C}=\mathbb{B}(g)$; that is, the g-radius hyper sphere. Then we have that*

$$\begin{aligned}&\Big|\mathbb{E}\Big[\langle U_1[V_1^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+-U_2[V_2^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+,U_1'[V_1'^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+-U_2'[V_2'^T\mathcal{P}_{\mathcal{C}}(\mathbf{x})]_+\rangle\\&-\langle U_1[V_1^T\mathbf{x}]_+-U_2[V_2^T\mathbf{x}]_+,U_1'[V_1'^T\mathbf{x}]_+-U_2'[V_2'^T\mathbf{x}]_+\rangle\Big]\Big|\\&\leq 2ge^{-g^2/2}[\|U_1\|_F\|U_1'\|_F\|V_1\|_F\|V_1'\|_F+\|U_2\|_F\|U_2'\|_F\|V_2\|_F\|V_2'\|_F].\end{aligned}\tag{338}$$

Proof. The proof is similar to the proof of Lemma 5. □

Corollary 7 (Two-Layer ReLU Neural Network). *Consider the true model for $(\mathbf{x},\mathbf{y})$, where $\mathbf{x}\sim\mathcal{N}(0,(1/n)I_n)\in\mathbb{R}^n$, $\mathbf{y}=U^*[V^{*T}\mathbf{x}]_+\epsilon$, where $U^*\in\mathbb{R}^{m\times R^*}$, $V^*\in\mathbb{R}^{n\times R^*}$, and $\epsilon\sim\mathcal{N}(0,(\sigma^2/m)I_m)\in\mathbb{R}^m$ independent from $\mathbf{x}$. For all $i\in[N]$, let $(\mathbf{x}_i,\mathbf{y}_i)$ be i.i.d. samples from this true model. Consider the estimator $\hat{\mathbf{y}}=U[V^T\mathbf{x}]_+$, where $U\in\mathbb{R}^{m\times R},V\in\mathbb{R}^{n\times R}$. Let $\delta\in(0,1]$ be fixed. Define the non-convex problem*

$$\mathsf{NC}_{\mu_N}^{\mathsf{ReLU}}((U,V)):=\frac{1}{2N}\sum_{i=1}^N\|\mathbf{y}_i-U[V^T\mathbf{x}_i]_+\|_2^2+\frac{\lambda}{2}\left(\|U\|_F^2+\|V\|_F^2\right),\tag{339}$$

and define $\mathsf{NC}_{\mu}^{\mathsf{ReLU}}((U,V))$ similarly with the sum over i replaced by expectation taken over $(\mathbf{x},\mathbf{y})$.

Let (U,V) be a stationary point of $\mathsf{NC}_{\mu_N}^{\mathsf{ReLU}}((U,V))$. Suppose there exists $C_{UV},B_u,B_v>0$ such that $\|UV^T\|_2\leq C_{UV}\left[\|U^*\|_F^2+\|V^*\|_F^2\right]$, and for all $j\in[R]$, $\|\mathbf{u}_j\|_2\leq B_u$, $\|\mathbf{v}_j\|_2\leq B_v$. Then with probability at least $1-\delta$, it holds that

$$\begin{aligned}\frac{1}{m}\left|\mathsf{NC}_{\mu}^{\mathsf{ReLU}}((U,V))-\mathsf{NC}_{\mu_N}^{\mathsf{ReLU}}((U,V))\right|&\lesssim\frac{1}{2m}\left[\|U^*\|_F^2+\|V^*\|_F^2\right]\left[\frac{1}{N}\sum_{i=1}^N\|\mathbf{y}_i-\hat{\mathbf{y}}_i\|_2\|\mathbf{x}_i\|_2-\lambda\right]\\&+C_{UV}^2\left[\|U^*\|_F^2+\|V^*\|_F^2\right]\left[\frac{R(m+n)\log\left(R(m+n)(C_{UV}+B_u^2+B_v^2)\right)\log(N)+\log(1/\delta)}{N}\right]^{1/2}.\end{aligned}$$

Proof. To obtain a generalization bound from Theorem 4 for this setting, we set the following problem parameters:

$$\ell(Y,\hat{Y})=\frac{1}{2}\|Y-\hat{Y}\|\implies(\alpha,L)=(0,1)\tag{340}$$

$$\phi(W)=[\langle\mathbf{v},\mathbf{x}\rangle]_+\mathbf{u};\tag{341}$$

$$\theta(W)=\frac{1}{2}\left[\|\mathbf{u}\|_2^2+\|\mathbf{v}\|_2^2\right].\tag{342}$$

Estimating $\Omega(f_\mu^*)$: From Proposition 2 we have that

$$\Omega(f_\mu^*)\leq\frac{\|U^*\|_F^2+\|V^*\|_F^2}{2}\tag{343}$$

Choosing $\mathcal{F}_\theta$:

$$\mathcal{F}_\theta:=\{(\mathbf{u},\mathbf{v}):\|\mathbf{u}\|^2+\|\mathbf{v}\|^2\leq 2,\|\mathbf{u}\|_2\leq 1,\|\mathbf{v}\|_2\leq 1\}.\tag{344}$$

Estimating L_ϕ: The Lipschtiz constant L_ϕ in the function $\mathcal{F}_\theta$ is $L_\phi:=\sup_{\|\mathbf{u}\|\leq1,\|\mathbf{v}\|\leq1}\|\mathbf{u}[\mathbf{v}^T.]_+\|_{\text{Lip}}\leq\sup_{\|\mathbf{u}\|\leq1,\|\mathbf{v}\|\leq1}\|\mathbf{u}\|\|\mathbf{v}\|=1$.

Estimating r_θ: For any $(\mathbf{u},\mathbf{v})\in\mathcal{F}_\theta$, we have that,

$$\frac{\|\mathbf{u}\|^2+\|\mathbf{v}\|^2}{2}\leq\sqrt{\frac{\|\mathbf{u}\|^2+\|\mathbf{v}\|^2}{2}}\implies\mathcal{F}_\theta\subseteq\mathbb{B}(1/\sqrt{2}).\tag{345}$$

Then we have $r_\theta = 1/\sqrt{2}$.

Choosing $\mathcal{F}_\mathcal{W}$: From the corollary's assumptions we have that $\mathcal{B}_R := \{(\mathbf{u}, \mathbf{v}) : \|\mathbf{u}\|_2 \leq B_u, \|\mathbf{v}\|_2 \leq B_v\}$; our hypothesis class is defined as

$$\mathcal{F}_\mathcal{W} := \{(U, V) : \|U[V^T.]_+\|_{\text{Lip}} \leq \|U\|_2\|V\|_2 \leq \gamma, \|\mathbf{u}_j\| \leq B_u, \|\mathbf{v}_j\| \leq B_v\}. \tag{346}$$

From Proposition 2, we have that, $\Omega(f^*_\mu) \leq \frac{1}{2}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]$. As we require $\gamma \geq \Omega(f^*_\mu)L_\phi = \frac{1}{2}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]$, we set $\gamma = C_{UV}\frac{\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]}{2}$.

$$\begin{aligned}\mathcal{F}_\mathcal{W} = \Big\{\{(\mathbf{u}_j, \mathbf{v}_j)\} : \|U[V^T]_+\cdot\|_{\text{Lip}} = \|UV^T\|_2 \leq \frac{C_{UV}}{2}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right], \\ \|\mathbf{u}_j\| \leq B_u, \|\mathbf{v}_j\| \leq B_v\Big\}.\end{aligned} \tag{347}$$

Estimating $\Omega^\circ_{\mu_N}(\cdot)$: We have

$$\begin{aligned}\Omega^\circ_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g, \Phi_r(\{W_j\}))\right) &= \Omega^\circ_{\mu_N}\left(\frac{1}{\lambda}(g - \Phi_r(\{W_j\}))\right) && (348)\\ &= \sup_{\|\mathbf{u}\|\leq 1;\|\mathbf{v}\|\leq 1} \frac{1}{N\lambda}\sum_{i=1}^N \langle Y_i - U[V^T\mathbf{x}_i]_+, \mathbf{u}[\mathbf{v}^T\mathbf{x}_i]_+\rangle && (349)\\ &= \sup_{\|\mathbf{v}\|\leq 1} \frac{1}{N\lambda}\sum_{i=1}^N [\mathbf{v}^T\mathbf{x}_i]_+\|Y_i - \hat{Y}_i\|_2 && (350)\\ &\leq \frac{1}{N\lambda}\sum_{i=1}^N \sup_{\|\mathbf{v}\|\leq 1} [\mathbf{v}^T\mathbf{x}_i]_+\|Y_i - \hat{Y}_i\|_2 && (351)\\ &= \frac{1}{N\lambda}\sum_{i=1}^N \|\mathbf{x}_i\|_2\|Y_i - \hat{Y}_i\|_2. && (352)\end{aligned}$$

Estimating ϵ_0: From the data generating mechanism we have $\|g\|_{\text{Lip}} \leq \|U^*\|_2\|V^*\|_2 \leq \frac{1}{2}\left[\|U^*\|_F + \|V^*\|_F\right]$, $\sigma_X = 1$, $\sigma_{Y|X} = \sigma$. Then we have the following constants from Theorem 4:

$$\epsilon_0 = 16\gamma^2\sigma_X^2 \min\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|^2_{\text{Lip}}}{\gamma^2}\left(1 + \frac{\sigma^2_{Y|X}}{\sigma^2_X}\right)\right]\right\}, \tag{353}$$

which evaluates to

$$\epsilon_0 = 8C^2_{UV}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\min\left\{1, \frac{(1+\sigma^2)}{4C^2_{UV}}\right\}. \tag{354}$$

Let $C_{UV} \leq 0.5\sqrt{1+\sigma^2}$ then we have

$$\epsilon_0 = 2(1+\sigma^2)\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]. \tag{355}$$

Estimating ϵ_1: Similarly,

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|^2_{\text{Lip}}}{\gamma^2}\left(1 + \frac{\sigma^2_{Y|X}}{\sigma^2_X}\right)\right]\right\}, \tag{356}$$

obtaining

$$\epsilon_1 = 8C^2_{UV}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]. \tag{357}$$

Defining a convex set $\mathcal{C}$: Consider a convex set $\mathcal{C} = \mathbb{B}(g) = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq g\}$.

First and foremost we need to estimate, $\delta_{\mathcal{C}}$ for the following inequality to hold:

$$P(\cap_{i=1}^{N} \mathbf{x}_i \in \mathcal{C}) \geq 1 - \delta_{\mathcal{C}}. \tag{358}$$

The probability of $\mathbf{x} \in \mathcal{C} = \mathbb{B}(g)$ is equivalent to saying the probability of the event when $\|\mathbf{x}\|_2 \leq g$. Since, $x_i \sim \mathcal{N}(0, 1/n)$ as a consequence of Bernstein's Inequality (Vershynin, 2018, Corollary 2.8.3) we have that, for any $t \geq 0$,

$$P(|\|\mathbf{x}\|_2 - 1| \leq t) \geq 1 - 2\exp\left(-cn_X t^2\right), \tag{359}$$

for some constant $c \geq 0$. Now we have

$$P(\|\mathbf{x}\|_2 \leq g) \begin{cases} \geq 1 - 2\exp\left(-cn_X(g-1)^2\right) & \text{if } g \geq 1 \\ \leq 2\exp\left(-cn_X(g-1)^2\right) & \text{otherwise.} \end{cases} \tag{360}$$

We consider the case where $g \geq 1$ yielding

$$P(\cap_{i=1}^{N} \mathbf{x}_i \in \mathcal{C}) = P(\cap_{i=1}^{N} \|\mathbf{x}\|_2 \leq g) \geq 1 - \underbrace{2N\exp\left(-cn_X(g-1)^2\right)}_{=\delta_{\mathcal{C}}}. \tag{361}$$

We have that $\delta_{\mathcal{C}} = 2N\exp\left(-cn(g-1)^2\right)$.

Now we evaluate $B_\ell, B_\Phi, \tilde{L}_\Phi, \tilde{L}_\phi$.

Estimating B_Φ: We have

$$\begin{aligned} B_\Phi &= \sup_{\mathbf{z}\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|U[V^T\mathbf{z}]_+\|_2 && (362)\\ &\leq \sup_{\mathbf{z}\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|U\|_2\|[V^T\mathbf{z}]_+\|_2 && (363)\\ &\leq \sup_{\mathbf{z}\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|U\|_2\|V\|_2\|\mathbf{z}\|_2 && (364)\\ &= g\gamma. && (365) \end{aligned}$$

Estimating B_ℓ: Similarly, we have

$$\begin{aligned} B_\ell &= \sup_{\mathbf{z}\in\mathcal{C},\{(\mathbf{u}_j,\mathbf{v}_j)\}\in\mathcal{F}_{\mathcal{W}}} \|U[V^T\mathbf{z}]_+ - U^*[V^{*T}\mathbf{z}]\|_2 && (366)\\ &= 2g\gamma && (367) \end{aligned}$$

Estimating $\tilde{L}_\Phi$: Now, we compute the Lipschitz constant with respect to U, V. We have

$$\begin{aligned} \tilde{L}_\Phi &= \sup_{\mathbf{z}\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|U[V^T\mathbf{z}]_+ - U'[V'^T\mathbf{z}]_+\|}{\max_j \sqrt{\|\mathbf{u}_j - \mathbf{u}'_j\|^2 + \|\mathbf{v}_j - \mathbf{v}'_j\|^2}} && (368)\\ &= R \sup_{\mathbf{z}\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|\mathbf{u}[\mathbf{v}^T\mathbf{z}]_+ - \mathbf{u}'[\mathbf{v}'^T\mathbf{z}]_+\|}{\sqrt{\|\mathbf{u} - \mathbf{u}'\|^2 + \|\mathbf{v} - \mathbf{v}'\|^2}} && (369)\\ &= R \sup_{\mathbf{z}\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|(\mathbf{u} - \mathbf{u}')[\mathbf{v}^T\mathbf{z}]_+ - \mathbf{u}'[[\mathbf{v}'^T\mathbf{z}]_+ - [\mathbf{v}^T\mathbf{z}]_+]\|}{\sqrt{\|\mathbf{u} - \mathbf{u}'\|^2 + \|\mathbf{v} - \mathbf{v}'\|^2}} && (370)\\ &\leq R \sup_{\mathbf{z}\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|(\mathbf{u} - \mathbf{u}')[\mathbf{v}^T\mathbf{z}]_+\| + \|\mathbf{u}'[[\mathbf{v}'^T\mathbf{z}]_+ - [\mathbf{v}^T\mathbf{z}]_+]\|}{\sqrt{\|\mathbf{u} - \mathbf{u}'\|^2 + \|\mathbf{v} - \mathbf{v}'\|^2}} && (371)\\ &\leq R \sup_{\mathbf{z}\in\mathcal{C},(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{\|(\mathbf{u} - \mathbf{u}')[\mathbf{v}^T\mathbf{z}]_+\| + \|\mathbf{u}'[[\mathbf{v}'^T\mathbf{z}]_+ - [\mathbf{v}^T\mathbf{z}]_+]\|}{\sqrt{\|\mathbf{u} - \mathbf{u}'\|^2 + \|\mathbf{v} - \mathbf{v}'\|^2}} && (372)\\ &\leq gR \sup_{(U,V),(U',V')\in\mathcal{F}_{\mathcal{W}}} \frac{B_v\|(\mathbf{u} - \mathbf{u}')\| + B_u\|\mathbf{v} - \mathbf{v}'\|}{\sqrt{\|\mathbf{u} - \mathbf{u}'\|^2 + \|\mathbf{v} - \mathbf{v}'\|^2}} && (373)\\ &= g\sqrt{B_u^2 + B_v^2}R. && (374) \end{aligned}$$

Estimating $\tilde{L}_\phi$: Similarly we get $\tilde{L}_\phi = g\sqrt{B_u^2 + B_v^2}$.

Estimating ϵ_2: Recall that

$$\epsilon_2 = \max\{8B_\ell \tilde{L}_\Phi, 8\tilde{L}_\Phi[B_\ell + B_\Phi L], 32\Omega(f_\mu^*)\tilde{L}_\phi \max\{B_\ell, LB_\Phi\}, 4\tilde{L}_\Phi B_\Phi\}. \tag{375}$$

From all the constants computed earlier, we have that

$$\epsilon_2 = k_1 g^2 R^2 C_{UV} \sqrt{B_u^2 + B_v^2} \left[\|U^*\|_F^2 + \|V^*\|_F^2\right], \tag{376}$$

for some constant $k_1 \geq 0$.

Next, we move on to estimating $B(\mathcal{C})$. We need to analyze three terms:
The First Term: Define

$$T_1 := \sup_{\{W_j\} \in \mathcal{F}_\mathcal{W}} \left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right|. \tag{377}$$

For a fixed (U, V), we have

$$\begin{aligned} &\left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| \\ &= \left| \mathbb{E}\Big[\|U^*[V^{*T}\mathcal{P}_\mathcal{C}(\mathbf{x})]_+ - U[V^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+\|^2 - \|U^*[V^{*T}\mathbf{x})_+ - U[V^T\mathbf{x}]_+\|^2 \Big] \right|. \end{aligned} \tag{378}$$

From Lemma 5 taking $g \geq 1$ we have

$$\begin{aligned} &\left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| \\ &\qquad \leq 2ge^{-g^2/2} \left[\|U^*\|_F^2 \|V^*\|_F^2 + \|U\|_F^2 \|V\|_F^2\right], \end{aligned} \tag{379}$$

whereupon further simplifying, we obtain

$$\begin{aligned} &\sup_{(U,V) \in \mathcal{F}_\mathcal{W}} \left| \|f_\mu^* \circ \mathcal{P}_\mathcal{C} - \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\|_\mu^2 - \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \right| \\ &\qquad \leq 2ge^{-g^2/2}(\|U^*\|_F^2 \|V^*\|_F^2 + R^2\gamma^2). \end{aligned} \tag{380}$$

Now, applying triangular inequality and taking the supremum we obtain

$$T_2 \leq 2ge^{-g^2/2}(\|U^*\|_F^2 \|V^*\|_F^2 + R^2\gamma^2). \tag{381}$$

The Second Term: Define

$$\begin{aligned} T_2 := \sup_{\{W_j\} \in \mathcal{F}_\mathcal{W}, W' \in \mathcal{F}_\theta} &\Big| \langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu \\ &- \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \Big|. \end{aligned} \tag{382}$$

For a fixed $(U, V), (\mathbf{u}, \mathbf{v})$ we have

$$\begin{aligned} &\left| \langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \right| = \\ &\qquad \Big| \mathbb{E}\Big[\langle U[V^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+ - U^*[V^{*T}\mathcal{P}_\mathcal{C}(\mathbf{x})]_+, \mathbf{u}[\mathbf{v}^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+ \rangle \\ &\qquad\qquad - \langle U[V^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+ - U^*[V^{*T}\mathbf{x}]_+, \mathbf{u}[\mathbf{v}^T\mathbf{x}]_+ \rangle \Big] \Big|. \end{aligned} \tag{383}$$

As a consequence of Lemma 6 we have

$$\left| \langle \nabla_{\hat{Y}} \ell \left(g \circ \mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\}) \circ \mathcal{P}_\mathcal{C}\right), \phi(W') \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell \left(g, \Phi_r(\{W_j\})\right), \phi(W') \rangle_\mu \right|$$

$$\leq 2ge^{-g^2/2}\left[\|U\|_F\|V\|_F\|\mathbf{u}\|\|\mathbf{v}\| + \|U^*\|_F\|V^*\|_F\|\mathbf{u}\|\|\mathbf{v}\|\right]. \tag{384}$$

Now we apply supremum over $(\mathbf{u},\mathbf{v}) \in \mathcal{F}_\theta$, obtaining

$$T_2 \leq 2ge^{-g^2/2}\left[\|U\|_F\|V\|_F + \|U^*\|_F\|V^*\|_F\right]. \tag{385}$$

Finally, we apply supremum over $(U,V) \in \mathcal{F}_\mathcal{W}$, obtaining

$$T_2 \leq 2ge^{-g^2/2}\left[\|U^*\|_F\|V^*\|_F + R\gamma\right]. \tag{386}$$

The Third Term: Define

$$\begin{aligned} T_3 := \sup_{\{W_j\}\in\mathcal{F}_\mathcal{W}} \Big| &\langle \nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\})\circ\mathcal{P}_\mathcal{C}\right), \Phi_r(\{W_j\})\circ\mathcal{P}_\mathcal{C}\rangle_\mu \\ &- \langle \nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu \Big|. \end{aligned} \tag{387}$$

For a fixed (U,V) we can rewrite the above to

$$\begin{aligned} \Big|\mathbb{E}\Big[&\langle U[V^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+ - U^*[V^{*T}\mathcal{P}_\mathcal{C}(\mathbf{x})]_+, U[V^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+\rangle \\ &-\langle U[V^T\mathcal{P}_\mathcal{C}(\mathbf{x})]_+ - U^*[V^{*T}\mathbf{x}]_+, U[V^T\mathbf{x}]_+\rangle\Big]\Big|. \end{aligned} \tag{388}$$

As a consequence of Lemma 4 we have

$$\begin{aligned} &\left|\langle \nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_\mathcal{C}, \Phi_r(\{W_j\})\circ\mathcal{P}_\mathcal{C}\right), \Phi_r(\{W_j\})\circ\mathcal{P}_\mathcal{C}\rangle_\mu - \langle \nabla_{\hat{Y}}\ell\left(g, \Phi_r(\{W_j\})\right), \Phi_r(\{W_j\})\rangle_\mu\right| \\ &\leq 2ge^{-g^2/2}\|U\|_F\|V\|_F\left[\|U\|_F\|V\|_F + \|U^*\|_F\|V^*\|_F\right]. \end{aligned} \tag{389}$$

Finally, we apply supremum over $(U,V) \in \mathcal{F}_\mathcal{W}$, obtaining

$$T_3 \leq 2ge^{-g^2/2}R\gamma\left[\|U^*\|_F\|V^*\|_F + R\gamma\right]. \tag{390}$$

Now combining T_1, T_2 and T_3 from equations (381), (386), (390) we have

$$B(\mathcal{C}) \leq 2ge^{-g^2/2}\left[\alpha(\|U^*\|_F^2\|V^*\|_F + R^2\gamma^2) + \|U\|_F\|V\|_F + R\gamma + R\gamma\left[\|U^*\|_F\|V^*\|_F + R\gamma\right]\right]. \tag{391}$$

We further upper bound for simplicity via

$$B(\mathcal{C}) \leq 4Rge^{-g^2/2}\gamma\left[\|U\|_F\|V\|_F + \gamma\right]. \tag{392}$$

From Theorem 4 we have that

$$\begin{aligned} \frac{1}{m}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| &\lesssim \frac{\lambda}{2m}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\left[\sup_{\|\mathbf{v}\|\leq 1}\frac{1}{N}\sum_{i=1}^N [\mathbf{v}^T\mathbf{x}_i]_+\|Y_i - U[V^T\mathbf{x}_i]_+\| - \lambda\right] \\ &+\frac{2}{m}Rge^{-g^2/2}C_{UV}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\left[\|U\|_F\|V\|_F + C_{UV}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\right] + C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]^2 \\ &\sqrt{\frac{R(m+n)\log\left(k_1g^2R^2C_{UV}^2\sqrt{B_u^2+B_v^2}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]^2\right)\log(N) + \log(1/\delta)}{N}}, \end{aligned} \tag{393}$$

holds true w.p at least $1 - \delta - 2N\exp\left(-cn_X(g-1)^2\right)$.

Now choose

$$g = 1 + \mathcal{O}\left(\sqrt{\log(NR) + \log(1/\delta)}\right). \tag{394}$$

After ignoring all the log-log terms and using only dominant terms, we have that

$$\frac{1}{m}\left|\mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \lesssim$$

$$\frac{\lambda}{2m}\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]\left[\sup_{\|\mathbf{v}\|\leq 1}\frac{1}{N}\sum_{i=1}^{N}[\mathbf{v}^T\mathbf{x}_i]_+\|Y_i - U[V^T\mathbf{x}_i]_+\| - \lambda\right]$$

$$+C_{UV}^2\left[\|U^*\|_F^2 + \|V^*\|_F^2\right]^2\sqrt{\frac{R(m+n)\log\left(R(m+n)(C_{UV}+B_u^2+B_v^2)\right)\log(N)+\log(1/\delta)}{N}}, \tag{395}$$

holds true w.p at least $1-\delta$. □

C.5 Multi-head Attention

Next, we move to applying Theorem 4 to the single-layer multi-head attention problem. We require similar Gaussian projections arguments onto convex sets are needed to be established. For this application, we require Gaussian projections onto softmax, which is analyzed through Lemma 7. First, we define the soft max operation, $\sigma_t(\cdot):\mathbb{R}^n\to\mathbb{R}^n$.

$$[\sigma_t(\mathbf{u})]_i := \exp(tu_i)/(\sum_{j=1}^{n}\exp(tu_j)), \tag{396}$$

where t is called the temperature.

A discrete version of soft-max is known as hard-max that is defined as

$$[\sigma(\mathbf{u})]_i := \mathbf{1}_{u_i=\max_i u_i}. \tag{397}$$

Note that when $t\to\infty$, $\sigma_t(\mathbf{u})\to\sigma(\mathbf{u})$.

Lemma 7 (Gaussian Softmax Projection)**.** *Let $X\in\mathbb{R}^{m\times n}$ and $X_{ij}\sim\mathcal{N}(0,1/(mn))$ be independent random variables. Suppose $\sigma_t(\cdot)$ is a softmax with temperature, t, and M is fixed matrix in $\mathbb{R}^{m\times n}$. Consider a convex set $\mathcal{C}=\{X=(\mathbf{x}_1,\ldots,\mathbf{x}_n):\forall j\in[n];\|\mathbf{x}_j\|\leq g\}$ for $g\geq 1$. Then*

$$\sup_{\mathbf{z}\in\mathcal{F}_{\mathbf{z}}}\left|\langle M,\mathbb{E}\left[X\sigma_t(X^T\mathbf{z})(X\sigma_t(X^T\mathbf{z}))^T - \mathcal{P}_{\mathcal{C}}(X)\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z})(\mathcal{P}_{\mathcal{C}}(X)\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z}))^T\right]\rangle\right|$$

$$\leq c_1m^2\|M\|_F g\exp\left(-c_2mg^2\right). \tag{398}$$

for some positive constant, c_1,c_2.

Proof. Denote

$$T := \sup_{\mathbf{z}\in\mathcal{F}_{\mathbf{z}}}\left|\langle M,\mathbb{E}\left[X\sigma_t(X^T\mathbf{z})(X\sigma_t(X^T\mathbf{z}))^T - \mathcal{P}_{\mathcal{C}}(X)\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z})(\mathcal{P}_{\mathcal{C}}(X)\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z}))^T\right]\rangle\right|. \tag{399}$$

Firstly, we upper bound the earlier term by Cauchy-Schwartz inequality:

$$T\leq\|M\|_F\sup_{\mathbf{z}\in\mathcal{F}_{\mathbf{z}}}\|\mathbb{E}\left[X\sigma_t(X^T\mathbf{z})(X\sigma_t(X^T\mathbf{z}))^T - \mathcal{P}_{\mathcal{C}}(X)\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z})(\mathcal{P}_{\mathcal{C}}(X)\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z}))^T\right]\|_F. \tag{400}$$

Denote $a_t(\mathbf{z},i,j)=\sigma_t(X^T\mathbf{z})_i\sigma_t(X^T\mathbf{z})_j$ and $\tilde{a}_t(\mathbf{z},i,j)=\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z})_i\sigma_t(\mathcal{P}_{\mathcal{C}}(X)^T\mathbf{z})_j$,

$$T\leq\|M\|_F\sup_{\mathbf{z}\in\mathcal{F}_{\mathbf{z}}}\|\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j^Ta_t(\mathbf{z},i,j) - \mathcal{P}_{\mathcal{C}}(\mathbf{x}_i)\mathcal{P}_{\mathcal{C}}(\mathbf{x}_j)^T\tilde{a}_t(\mathbf{z},i,j)\right]\|_F. \tag{401}$$

Now we apply triangular inequality,

$$T\leq\|M\|_F\sup_{\mathbf{z}\in\mathcal{F}_{\mathbf{z}}}\sum_{i=1}^{T}\sum_{j=1}^{T}\|\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j^Ta_t(\mathbf{z},i,j) - \mathcal{P}_{\mathcal{C}}(\mathbf{x}_i)\mathcal{P}_{\mathcal{C}}(\mathbf{x}_j)^T\tilde{a}_t(\mathbf{z},i,j)\right]\|_F. \tag{402}$$

Now we apply Cauchy-Schwartz,

$$T \leq \|M\|_F \sum_{i=1}^{m}\sum_{j=1}^{m} \sup_{\mathbf{z}\in\mathcal{F}_\mathbf{z}} \|\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j^T a_t(\mathbf{z},i,j) - \mathcal{P}_\mathcal{C}(\mathbf{x}_i)\mathcal{P}_\mathcal{C}(\mathbf{x}_j)^T\tilde{a}_t(\mathbf{z},i,j)\right]\|_F. \tag{403}$$

We can upper bound the earlier term via taking a supremum over indices $i,j\in[T]$ then we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\sup_{\mathbf{z}\in\mathcal{F}_\mathbf{z}} \|\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j^T a_t(\mathbf{z},i,j) - \mathcal{P}_\mathcal{C}(\mathbf{x}_i)\mathcal{P}_\mathcal{C}(\mathbf{x}_j)^T\tilde{a}_t(\mathbf{z},i,j)\right]\|_F. \tag{404}$$

Observe that argument inside the expectation is 0 on the event $X\in\mathcal{C}$, thereby we only have the case where $X\notin\mathcal{C}$ then we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\sup_{\mathbf{z}\in\mathcal{F}_\mathbf{z}} \|\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j^T a_t(\mathbf{z},i,j) - \mathcal{P}_\mathcal{C}(\mathbf{x}_i)\mathcal{P}_\mathcal{C}(\mathbf{x}_j)^T\tilde{a}_t(\mathbf{z},i,j)\mathbf{1}_{\mathcal{E}^c}\right]\|_F. \tag{405}$$

On application Cauchy-Schwartz identity again we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\sup_{\mathbf{z}\in\mathcal{F}_\mathbf{z}} \mathbb{E}\left[\|\mathbf{x}_i\mathbf{x}_j^T a_t(\mathbf{z},i,j) - \mathcal{P}_\mathcal{C}(\mathbf{x}_i)\mathcal{P}_\mathcal{C}(\mathbf{x}_j)^T\tilde{a}_t(\mathbf{z},i,j)\|_F\mathbf{1}_{\mathcal{E}^c}\right]. \tag{406}$$

We have that $\mathbf{x}\notin\mathcal{C}, \mathcal{P}_\mathcal{C}(\mathbf{x}) = g\mathbf{x}/\|\mathbf{x}\|$. By using this fact we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\sup_{\mathbf{z}\in\mathcal{F}_\mathbf{z}} \mathbb{E}\left[\|\left(a_t(\mathbf{z},i,j) - \frac{g^2}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}\tilde{a}_t(\mathbf{z},i,j)\right)\mathbf{x}_i\mathbf{x}_j^T\|_F\mathbf{1}_{\mathcal{E}^c}\right]. \tag{407}$$

Now we recall Reverse Fatou's Lemma, for any function sequence, $f_n\in L^2(\mu)$, we have

$$\limsup_{n\to\infty}\int f_n d\mu \leq \int \limsup_{n\to\infty} f_n d\mu. \tag{408}$$

on applying this identity we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\mathbb{E}\left[\sup_{\mathbf{z}\in\mathcal{F}_\mathbf{z}}\left|a_t(\mathbf{z},i,j) - \frac{g^2}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}\tilde{a}_t(\mathbf{z},i,j)\right|\|\mathbf{x}_i\mathbf{x}_j^T\|_F\mathbf{1}_{\mathcal{E}^c}\right]. \tag{409}$$

Observe that $a_t(\mathbf{z},i,j)\leq 1$ and $\frac{g^2}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}\tilde{a}_t(\mathbf{z},i,j)\leq 1$ when $X\notin\mathcal{C}$ therefore we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\mathbb{E}\left[\|\mathbf{x}_i\mathbf{x}_j^T\|_F\mathbf{1}_{\mathcal{E}^c}\right]. \tag{410}$$

Now since $\mathbf{x}_i$ and $\mathbf{x}_j$ are iid we have

$$T \leq \|M\|_F m^2 \sup_{i,j}\mathbb{E}\left[\|\mathbf{x}_i\|_2\mathbf{1}_{\mathcal{E}^c}\right]\mathbb{E}\left[\|\mathbf{x}_j\|_2\mathbf{1}_{\mathcal{E}^c}\right]. \tag{411}$$

By Gaussian integral over norm for $g\geq 1$ we have

$$T \lesssim m^2\|M\|_F g\exp\left(-mg^2\right) \tag{412}$$

□

With the above result on Gaussian softmax projection, we now state the corollary for the single-layer multi-head attention problem and its proof.

Corollary 8 (Transformers)**.** *Consider the true model for $(X,\mathbf{y})$, where $X\in\mathbb{R}^{n\times T}$ is a random matrix with i.i.d. entries $X_{lk}\sim\mathcal{N}(0,1/(nT))$ and $\mathbf{y} = A^*X\mathbf{b}^*+\epsilon$, where $A^*\in\mathbb{R}^{m\times n}$, $\mathbf{b}^*\in\mathbb{S}^{T-1}$ and $\epsilon\sim\mathcal{N}(0,(\sigma^2/m)I_m)$ is independent from X. For all $i\in[N]$, let $(X_i,\mathbf{y}_i)$ be i.i.d. samples from this true model. Consider the estimator $\hat{\mathbf{y}} = \sum_{j=1}^R V_jX\sigma(X^T\mathbf{z}_j)$, $V_j\in\mathbb{R}^n, \mathbf{z}_j\in\mathbb{R}^n$. Let $\delta\in(0,1]$ be fixed. Define the non-convex problem*

$$\mathsf{NC}^{\mathsf{TF}}_{\mu_N}(\{(V_j,\mathbf{z}_j)\}):=\frac{1}{2N}\sum_{i=1}^{N}\|\mathbf{y}_i-\sum_{j=1}^{R}V_jX_i\sigma_t(X_i^T\mathbf{z}_j)\|_2^2+\lambda\sum_{j=1}^{R}\left[\|V_j\|_F+\delta_{\{\mathbf{z}:\|\mathbf{z}\|_2\le 1\}}(\mathbf{z}_j)\right],$$

where, $\sigma_t(\cdot)$ is softmax function with temperature t, for $k\in[T]$ defined $\sigma_t(\mathbf{u})_k:=\exp(tu_k)/\sum_{l=1}^{T}\exp(tu_l)$ and define $\mathsf{NC}^{\mathsf{TF}}_{\mu}(\{(V_j,\mathbf{z}_j)\})$ similarly with the sum over i replaced by expectation taken over $(X,\mathbf{y})$.

Let $\{(V_j,\mathbf{z}_j)\}$ be a stationary point of $\mathsf{NC}^{\mathsf{TF}}_{\mu_N}(\{(V_j,\mathbf{z}_j)\})$. Suppose there exists $C_V,B_V>0$ such that $\sum_{j=1}^{R}\|V_j\|_F\le C_V\|A^*\|_F$, and for all $j\in[R]$, $\|V_j\|_F\le B_V$.

Then with probability at least $1-\delta$, it holds that

$$\frac{1}{m}\left|\mathsf{NC}^{\mathsf{TF}}_{\mu}(\{(V_j,\mathbf{z}_j)\})-\mathsf{NC}^{\mathsf{TF}}_{\mu_N}(\{(V_j,\mathbf{z}_j)\})\right|\lesssim\frac{1}{2m}\|A^*\|_F\left[\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i-\hat{\mathbf{y}}_i\|_2\|X_i\|_2-\lambda\right]$$
$$+C_V^2\|A^*\|_F^2\sqrt{\frac{R(m+n)\log(R(m+n)(C_V+B_V))\log(N)+\log(1/\delta)}{N}}.$$

Proof. To obtain a generalization bound from Theorem 4 for the case of matrix sensing, we set the following problem parameters.

$$\ell(Y,\hat{Y})=\frac{1}{2}\|Y-\hat{Y}\|\implies(\alpha,L)=(0,1);\tag{413}$$
$$\phi(W)=VX\sigma_t(X^T\mathbf{z});\tag{414}$$
$$\theta(W)=\|V\|_F+\delta_{\mathbf{z}\in\mathbb{B}(1)}.\tag{415}$$

Estimating $\Omega_{\mu_N}(\cdot)$: Now we move on to compute the polar:

$$\begin{aligned}\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))\right)&=\Omega^{\circ}_{\mu_N}\left(\frac{1}{\lambda}(g-\Phi_r(\{W_j\}))\right),\\&=\sup_{\|V\|_F\le1,\|\mathbf{z}\|\le1}\frac{1}{N\lambda}\sum_{i=1}^{N}\langle Y_i-\hat{Y}_i,VX_i\sigma_t(X^T\mathbf{z})\rangle,\\&=\sup_{\|V\|_F\le1,\|z\|\le1}\frac{1}{N\lambda}\|\sum_{i=1}^{N}\langle(Y_i-\hat{Y}_i)(X_i\sigma_t(X^T\mathbf{z}))^T,V\rangle\|,\\&=\sup_{\|V\|_F\le1,\|z\|\le1}\frac{1}{N\lambda}\|\sum_{i=1}^{N}\langle(Y_i-\hat{Y}_i)(X_i\sigma_t(X^T\mathbf{z}))^T,V\rangle\|,\end{aligned}$$

$$\begin{aligned}\Omega^{\circ}_{\mu_N}\left(-\frac{1}{\lambda}\nabla_{\hat{Y}}\ell(g,\Phi_r(\{W_j\}))\right)&\le\sup_{\|z\|\le1}\frac{1}{N\lambda}\|\sum_{i=1}^{N}(Y_i-\hat{Y}_i)(X_i\sigma_t(X^T\mathbf{z}))^T\|_F,\\&\le\frac{1}{N\lambda}\sum_{i=1}^{N}\sup_{\|z\|\le1}\|(Y_i-\hat{Y}_i)(X_i\sigma_t(X^T\mathbf{z}))^T\|_F,\\&\le\frac{1}{N\lambda}\sum_{i=1}^{N}\sup_{\|z\|\le1}\|Y_i-\hat{Y}_i\|_F\|(X_i\sigma_t(X^T\mathbf{z}))^T\|_F,\\&\le\frac{1}{N\lambda}\sum_{i=1}^{N}\sup_{\|z\|\le1}\|Y_i-\hat{Y}_i\|_F\|(X_i\sigma_t(X^T\mathbf{z}))^T\|_F,\\&\le\frac{1}{N\lambda}\sum_{i=1}^{N}\|Y_i-\hat{Y}_i\|_2\|X_i\|_2.\end{aligned}$$

Choose $\mathcal{F}_\theta$: From Assumptions 4 suppose that

$$\mathcal{F}_\theta := \{(V, \mathbf{z}) : \|V\|_2 \leq 1, \|\mathbf{z}\|_2 \leq 1\}; \tag{416}$$

Computing L_ϕ: The Lipschtiz constant L_ϕ in the function $\mathcal{F}_\theta$ is $L_\phi := \sup_{\|V\|\leq 1, \|\mathbf{z}\|\leq 1} \|V(\cdot)\sigma((\cdot)^T\mathbf{z})\|_{\text{Lip}} \leq \sup_{\|V\|\leq 1} \|V\| = 1$.

Computing r_ϕ: Clearly, when $r_\theta = \sqrt{2}$ we have that $\mathcal{F}_\theta \subseteq \mathbb{B}(\sqrt{2})$.

Choose $\mathcal{F}_\mathcal{W}$: From the corollary's assumptions we have that, $\mathcal{B}_R := \{(V, \mathbf{z}) : \|V\|_2 \leq B_V, \|\mathbf{z}\|_2 \leq 1\}$; our hypothesis class is defined as

$$\mathcal{F}_\mathcal{W} := \{\{(V_j, \mathbf{z}_j)\} : \|\sum_{j=1}^{r} V_j \cdot \sigma((\cdot)^T\mathbf{z}_j)\|_{\text{Lip}} \leq \sum_{j=1}^{r} \|V_j\|_F \leq \gamma, \|V_j\|_F \leq B_V, \|\mathbf{z}\| \leq 1\}. \tag{417}$$

From Proposition 2 we have $\Omega(f_\mu^*) \leq \|A^*\|_F$. We have $\gamma \geq \Omega(f_\mu^*)L_\phi = \|A^*\|_F$, then we set $\gamma = C_V\|A^*\|_F$. We have that

$$\begin{aligned}\mathcal{F}_\mathcal{W} := \Big\{\{(V_j, \mathbf{z}_j)\} : \|\sum_{j=1}^{r} V_j \cdot \sigma((\cdot)^T\mathbf{z}_j)\|_{\text{Lip}} \leq \sum_{j=1}^{r} \|V_j\|_F \leq C_V\|A^*\|_F, \\ \|V_j\|_F \leq B_V, \|\mathbf{z}\| \leq 1\Big\}.\end{aligned} \tag{418}$$

Estimating ϵ_0: From the data generating mechanism we have $\|g\|_{\text{Lip}} \leq \|A^*\|_F$, $\sigma_X = 1$, $\sigma_{Y|X} = \sigma$, then we have the following constants from Theorem 4:

$$\epsilon_0 = 16\gamma^2\sigma_X^2 \min\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}, \tag{419}$$

this evaluates to,

$$\epsilon_0 = 16(\sigma^2 + 1)\|A^*\|_F^2, \tag{420}$$

when $C_V \leq \sqrt{1 + \sigma^2}$.

Estimating ϵ_1: Similarly, we evaluate

$$\epsilon_1 = 16\gamma^2\sigma_X^2 \max\left\{1, \frac{L}{4}\left[1 + \frac{\|g\|_{\text{Lip}}^2}{\gamma^2}\left(1 + \frac{\sigma_{Y|X}^2}{\sigma_X^2}\right)\right]\right\}, \tag{421}$$

obtaining,

$$\epsilon_1 = 16C_V^2\|A^*\|_F^2, \tag{422}$$

when $C_V \leq \sqrt{1 + \sigma^2}$.

Choosing the convex set $\mathcal{C}$: Consider a convex set $\mathcal{C} = \{X = (\mathbf{x}_1, \ldots, \mathbf{x}_T) : \|\mathbf{x}_j\|_2 \leq g/\sqrt{T}\}$.

First and foremost we need to estimate $\delta_\mathcal{C}$ for the inequality to hold:

$$P(\cap_{i=1}^{N}\mathbf{x}_i \in \mathcal{C}) \geq 1 - \delta_\mathcal{C}. \tag{423}$$

The probability of $\mathbf{x} \in \mathcal{C} = \mathbb{B}(g)$ is equivalent to saying the probability of the event when $\|\mathbf{x}\|_2 \leq g$. Since, $x_i \sim \mathcal{N}(0, 1/n)$ as a consequence of Bernstein's Inequality (Vershynin, 2018, Corollary 2.8.3) we have that, for any $t \geq 0$.

$$P(|\|\mathbf{x}\|_2 - 1| \leq t) \geq 1 - 2\exp\left(-cnTt^2\right), \tag{424}$$

for some constant $c \geq 0$. Now we have

$$P(\|\mathbf{x}\|_2 \leq g/\sqrt{T}) \begin{cases} \geq 1 - 2\exp\left(-cn(g-1)^2\right) & \text{if } g \geq 1 \\ \leq 2\exp\left(-cn(g-1)^2\right) & \text{otherwise} \end{cases}. \tag{425}$$

We consider the case where $g \geq 1$. Then we have that

$$P(\cap_{i=1}^{N} \mathbf{x}_i \in \mathcal{C}) = P(\cap_{i=1}^{N} \|\mathbf{x}\|_2 \leq g/\sqrt{T}) \geq 1 - \underbrace{2N \exp\left(-cn(g-1)^2\right)}_{=\delta_{\mathcal{C}}}. \tag{426}$$

We have that $\delta_{\mathcal{C}} = 2N \exp\left(-cn(g-1)^2\right)$.

Estimating B_Φ:

$$\begin{aligned} B_\Phi &= \sup_{X \in \mathcal{C}, \{(V_j, \mathbf{z}_j)\} \in \mathcal{F}_{\mathcal{W}}} \| \sum_{j=1}^{r} V_j X \sigma_t(X^T \mathbf{z}_j) \|_2, && (427) \\ &\leq R \sup_{X \in \mathcal{C}, \{(V_j, \mathbf{z}_j)\} \in \mathcal{F}_{\mathcal{W}}} \|V_j\|_F \|X \sigma_t(X^T \mathbf{z})\|_2, && (428) \\ &\leq R \sup_{X \in \mathcal{C}, \{(V_j, \mathbf{z}_j)\} \in \mathcal{F}_{\mathcal{W}}} \|V_j\|_F \|X \sigma_t(X^T \mathbf{z})\|_2, && (429) \\ &\leq gRB_V/\sqrt{T}. && (430) \end{aligned}$$

Estimating B_ℓ: we have

$$\begin{aligned} B_\ell &= \sup_{X \in \mathcal{C}, \{(V_j, \mathbf{z}_j)\} \in \mathcal{F}_{\mathcal{W}}} \| \sum_{j=1}^{r} V_j X \sigma_t(X^T \mathbf{z}_j) - A^* X B^* \|, && (431) \\ &= g\left[RB_V + \|A^*\|_F\right]/\sqrt{T}. && (432) \end{aligned}$$

Estimating $\tilde{L}_\Phi$: We have

$$\tilde{L}_\Phi = gR\sqrt{B_V^2 + 1}/\sqrt{T}. \tag{433}$$

Estimating $\tilde{L}_\phi$: Similarly we get $\tilde{L}_\phi = g\sqrt{B_V^2 + 1}/\sqrt{T}$ as we have only one slice of factor.

Estimating ϵ_2: Recall that,

$$\epsilon_2 = \max\{8B_\ell \tilde{L}_\Phi, 8\tilde{L}_\Phi[B_\ell + B_\Phi L], 32\Omega(f_\mu^*)\tilde{L}_\phi \max\{B_\ell, LB_\Phi\}, 4\tilde{L}_\Phi B_\Phi\}. \tag{434}$$

From all the constants computed earlier, we have that,

$$\epsilon_2 = k_1 g^2 R^2 B_V^2/T, \tag{435}$$

for some constant $k_1 \geq 0$.

Next, we move on estimating $B(\mathcal{C})$ we need to analyze three terms

The First Term is defined via

$$\begin{aligned} T_1 := \sup_{\{W_j\} \in \mathcal{F}_{\mathcal{W}}} \Big| &\|f_\mu^* \circ \mathcal{P}_{\mathcal{C}} - \Phi_r(\{W_j\}) \circ \mathcal{P}_{\mathcal{C}}\|_\mu^2 \\ &- \|f_\mu^* - \Phi_r(\{W_j\})\|_\mu^2 \Big|. \end{aligned} \tag{436}$$

From Lemma 7 we obtain that, taking $g \geq 1$ and further simplifying, we get,

$$T_1 \leq c_1 T^2 g e^{-c_2 T g^2} \left[R^2\gamma^2 + \|A^*\|_2\right], \tag{437}$$

The Second Term is defined via

$$T_2 := \sup_{\{W_j\}\in\mathcal{F}_{\mathcal{W}},W'\in\mathcal{F}_\theta} \Big|\langle \nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_{\mathcal{C}},\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\right),\phi(W')\circ\mathcal{P}_{\mathcal{C}}\rangle_\mu \\ - \langle \nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right),\phi(W')\rangle_\mu\Big|. \tag{438}$$

As a consequence of Lemma 7 we have

$$T_2 \le c_1 T^2 g e^{-c_2 T g^2}\left[\|A^*\|_2 + R\gamma\right]. \tag{439}$$

The Third Term is defined via

$$T_3 := \sup_{\{W_j\}\in\mathcal{F}_{\mathcal{W}}} \Big|\langle \nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_{\mathcal{C}},\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\right),\Phi_r(\{W_j\})\circ\mathcal{P}_{\mathcal{C}}\rangle_\mu \\ - \langle \nabla_{\hat{Y}}\ell\left(g,\Phi_r(\{W_j\})\right),\Phi_r(\{W_j\})\rangle_\mu\Big|. \tag{440}$$

As a consequence of Lemma 7 we have

$$T_3 \le c_1 T^2 g e^{-c_2 T g^2} R\gamma\left[\|A^*\|_2 + R\gamma\right], \tag{441}$$

for some positive constant, c_1, c_2.

Now combining T_1, T_2, and T_3 from equations (437), (439), (441) we obtain that

$$B(\mathcal{C}) \le c_1 T^2 g e^{-c_2 T g^2}\left[\alpha(\|A^*\|_2 + R^2\gamma^2) + \|A^*\|_2 + R\gamma + R\gamma\left[\|A^*\|_2 + R\gamma\right]\right]. \tag{442}$$

We further upper bound for simplicity as

$$B(\mathcal{C}) \le 4Rc_1\sqrt{\frac{\log(T)}{T^5}} g e^{-c_2 g^2}\gamma\left[\|A^*\|_2 + \gamma\right]. \tag{443}$$

From Theorem 4 we have that

$$\frac{1}{m}\left|\mathsf{NC}_\mu(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \lesssim \frac{1}{2m}\|A^*\|_F\left[\sup_{\|z\|\le 1}\frac{1}{N}\|\sum_{i=1}^N (Y_i - \hat{Y}_i)^T(X_i\sigma_t(X^T\mathbf{z}))\|_F - \lambda\right] \\ + \frac{2}{m}Rc_1T^2ge^{-c_2Tg^2}C_V\left[\|A^*\|_F\right]\left[\|A^*\|_2+\gamma\right] + C_V^2\left[\|A^*\|_F\right] \\ \times\sqrt{\frac{R(m+n)\log\left(C_V^2\left[\|U^*\|_F^2+\|V^*\|_F^2\right]^2 k_1 g^2 RB_V/T\right)\log(N)+\log(1/\delta)}{N}}. \tag{444}$$

holds true w.p at least $1-\delta-2N\exp\left(-cn_X(g-1)^2\right)$.

Now choose

$$g = 1 + \tilde{\mathcal{O}}\left(\frac{1}{T}\log\left(\frac{N}{R(m+n)+log(1/\delta)}\right)\right). \tag{445}$$

Then we have

$$\frac{2}{m}Rc_1T^2ge^{-c_2Tg^2}C_V\left[\|A^*\|_F\right]\left[\|A^*\|_2+\gamma\right] \lesssim \sqrt{\frac{R(m+n)\log\left(C_V^2\left[\|U^*\|_F^2+\|V^*\|_F^2\right]^2 k_1 g^2 RB_V/T\right)\log(N)+\log(1/\delta)}{N}}. \tag{446}$$

Therefore we can upper bound the middle term to the right most leaving us behind

$$\frac{1}{m}\left|\mathsf{NC}_{\mu}(\{W_j\}) - \mathsf{NC}_{\mu_N}(\{W_j\})\right| \lesssim \frac{1}{2m}\|A^*\|_F \left[\frac{1}{N}\sum_{i=1}^{N}\|Y_i - \hat{Y}_i\|_2\|X_i\|_2 - \lambda\right]$$

$$+C_V^2\left[\|A^*\|_F\right]\sqrt{\frac{R(m+n)\log\left(R(m+n)(C_V+B_V)\right)\log(N)+\log(1/\delta)}{N}}, \tag{447}$$

holds true w.p at least $1-\delta$. □

D GOODS EVENTS

In this section, we provide compute the probabilities of events defined in the proof of Theorem 4. Recall the definition of our function classes:

$$\mathcal{F}_\theta := \{\{W_j\} : \|\Phi_R(\{W_j\})\|_{\text{Lip}} \leq \gamma, \Theta_R(\{W_j\}) \leq \gamma/L_\phi\}; \tag{448}$$
$$\mathcal{F}_\mathcal{W} := \{\{W_j\} : \|\Phi_R(\{W_j\})\|_{\text{Lip}} \leq \gamma, \Theta_R(\{W_j\}) \leq \gamma/L_\phi\}; \tag{449}$$
$$\mathcal{F}_\Phi := \{\Phi_R(\zeta) : \forall \zeta \in \mathcal{F}_\mathcal{W}\} \tag{450}$$

We define the below events:

$$\mathcal{E}_{cvx}(\epsilon) := \{\forall \zeta \in \mathcal{F}_\mathcal{W} : |\mathsf{C}_{\mu_N}(f_\zeta) - \mathsf{C}_\mu(f_\zeta)| \leq \epsilon + B_{nrm}(\mathcal{C})\}; \tag{451}$$
$$\mathcal{E}_{eql}(\epsilon) := \{\forall \zeta \in \mathcal{F}_\mathcal{W} : \left|\langle \nabla_{\hat{Y}} \ell\left(g, f_\zeta\right), f_\zeta\rangle_{\mu_N} - \langle \nabla_{\hat{Y}} \ell\left(g, f_\zeta\right), f_\zeta\rangle_{\mu}\right| \leq \epsilon + B_{eql}(\mathcal{C})\}; \tag{452}$$
$$\mathcal{E}_{plr}(\epsilon) := \{\forall \zeta \in \mathcal{F}_\mathcal{W} : \left|\Omega^\circ_{\mu_N}\left(\nabla_{\hat{Y}} \ell\left(g, f_\zeta\right)\right) - \Omega^\circ_{\mu}\left(\nabla_{\hat{Y}} \ell\left(g, f_\zeta\right)\right)\right| \leq \epsilon + B_{plr}(\mathcal{C})\}; \tag{453}$$
$$\mathcal{E}_{nrm}(\epsilon) := \{\forall \zeta \in \mathcal{F}_\mathcal{W} : \left|\|f^*_\mu - f_\zeta\|^2_{\mu_N} - \|f^*_\mu - f_\zeta\|^2_{\mu}\right| \leq \epsilon + B_{nrm}(\mathcal{C})\}. \tag{454}$$

In each of the sections below, we discuss the technical analysis to estimate the probability of the events, $\mathcal{E}_{cvx}(\epsilon)$, $\mathcal{E}_{eql}(\epsilon)$, $\mathcal{E}_{plr}(\epsilon)$ and $\mathcal{E}_{nrm}(\epsilon)$.

D.1 Concentration of Norms

In this section, we upper bound the probability of the event, $\mathcal{E}_{nrm}(\epsilon)$ through Lemma 8.

Lemma 8 (Concentration of Norms)**.** *Consider an n_X-dimensional sub-Gaussian vector $X \sim SG(0, (\sigma_X^2/n_X)I)$, and set of functions $f_\zeta : \mathbb{R}^{n_X} \to \mathbb{R}$ as parameterized by $\zeta \in \mathcal{F}_\mathcal{W}$. Let $\mathcal{C}$ be some convex obeying $P(\bigcap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_\mathcal{C}$ for i.i.d samples $\{X_i\}_{i=1}^N$. Assume that for any fixed, $\zeta, \zeta' \in \mathcal{F}_\mathcal{W}$, and fixed $Z \in \mathcal{C}$, we have*

$$\|f_\zeta(Z) - f_{\zeta'}(Z)\| \leq \tilde{L}_\Phi d(\zeta_1, \zeta_2) \text{ and } \|f_\zeta(Z)\| \leq B_\Phi. \tag{455}$$

Denote,

$$B_{nrm}(\mathcal{C}) := \sup_{\zeta \in \mathcal{F}_\mathcal{W}} \left|\|f^*_\mu \circ \mathcal{P}_\mathcal{C} - f_\zeta \circ \mathcal{P}_\mathcal{C}\|^2_\mu - \|f^*_\mu - f_\zeta\|^2_\mu\right|, \tag{456}$$

where $\mathcal{P}_\mathcal{C}(\cdot)$ denotes the Euclidean projection onto the set $\mathcal{C}$. Define,

$$K := 64 n_Y \gamma^2 \sigma_X^2. \tag{457}$$

Then for any $\epsilon \in [0, K]$,

$$\mathbb{P}\left(\sup_{\zeta \in \mathcal{F}_\mathcal{W}} \left|\|f^*_\mu - f_\zeta\|^2_{\mu_N} - \|f^*_\mu - f_\zeta\|^2_{\mu}\right| \geq \epsilon + B_{nrm}(\mathcal{C})\right)$$
$$\leq \delta_\mathcal{C} + c \exp\left(\log(C_{\mathcal{F}_\mathcal{W}}\left(\frac{\epsilon}{4\tilde{L}_\Phi B_\Phi}\right)) - N\frac{\epsilon^2}{K^2}\right). \tag{458}$$

for some positive constant, c and $C_{\mathcal{F}_\mathcal{W}}(\nu)$ is the ν-net covering number of the set $\mathcal{F}_\mathcal{W}$.

Proof. If $X \in \mathbb{R}^{n_X} \sim SG\left(\frac{\sigma_X^2}{n_X} I_{n_X \times n_X}\right)$, The function map, $\|f^*_\mu - f_\zeta\|$ has Lipschitz constant of $\|f^*_\mu\|_{\text{Lip}} + \|f_\zeta\|_{\text{Lip}} \leq 2\gamma$; as $f_\zeta, f^*_\mu \in \mathcal{F}_\Phi$. Therefore from Theorem 5.1.4 in Vershynin (2018) we have that, $f^*_\mu(X) - f_\zeta(X) \sim SG\left(4\gamma^2\sigma_X^2 I_{n_Y \times n_Y}\right)$. Thus, $\|f^*_\mu(X) - f_\zeta(X)\|^2 \sim SE\left(4 n_Y \gamma^2 \sigma_X^2\right)$

Now, applying the concentration inequality for sub-exponential from Theorem 2.8.1 Vershynin (2018) for a fixed $\zeta \in \mathcal{F}_\mathcal{W}$, we have that

$$\mathbb{P}\left(\left|\|f^*_\mu - f_\zeta\|^2_{\mu_N} - \|f^*_\mu - f_\zeta\|^2_{\mu}\right| \geq \epsilon\right) \leq C \exp\left(-N \min\left\{\frac{\epsilon^2}{16 n_Y^2 \gamma^4 \sigma_X^4}, \frac{\epsilon}{4 n_Y \gamma^2 \sigma_X^2}\right\}\right), \tag{459}$$

for some positive constant, $C \geq 0$. We use Lemma 14 for applying the concentration bounds. Now set

$$g_\theta = \|f_\mu^* - f_\zeta\|^2. \tag{460}$$

We need to check if the function, g, is Lipschitz on some metric and convex set $\mathcal{C} \subseteq \mathbb{R}^{n_X}$, choose for any $Z \in \mathcal{C}$. We have $P(\bigcap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_{\mathcal{C}}$. Recall that

1. $\forall \zeta_1, \zeta_2 \in \mathcal{F}_{\mathcal{W}} : \|f_{\zeta_1}(Z) - f_{\zeta_2}(Z)\| \leq \tilde{L}_\Phi d(\zeta_1, \zeta_2)$, for all $Z \in \mathcal{C}$.
2. $\forall \zeta \in \mathcal{F}_{\mathcal{W}} : \|f_\zeta(Z)\| \leq B_\Phi$, for all $Z \in \mathcal{C}$.
3. For a fixed $\zeta \in \mathcal{F}_{\mathcal{W}}$,

$$\left|\mathbb{E}\left[\|f_\mu^*(\mathcal{P}_{\mathcal{C}}(X)) - f_\zeta(\mathcal{P}_{\mathcal{C}}(X))\|^2 - \|f_\mu^*(X) - f_\zeta(X)\|^2\right]\right| \leq B_{nrm}(\mathcal{C}). \tag{461}$$

By exploiting the above items we have

$$\begin{aligned} |g_{\theta_1} - g_{\theta_2}| &= \left|\|f_\mu^* - f_{\zeta_1}\|^2 - \|f_\mu^* - f_{\zeta_2}\|^2\right|, = \left|\langle 2f_\mu^* - (f_{\zeta_1} + f_{\zeta_2}), f_{\zeta_1} - f_{\zeta_2}\rangle\right|, \\ &\leq \|2f_\mu^* - (f_{\zeta_1} + f_{\zeta_2})\|^\circ \|f_{\zeta_1} - f_{\zeta_2}\|, \\ &\leq 4\tilde{L}_\Phi B_\Phi d(\zeta_1, \zeta_2). \end{aligned}$$

Then we have that for covering number $C_{\mathcal{F}_{\mathcal{W}}}(\nu) = \mathcal{N}(\mathcal{F}_{\mathcal{W}}, d(.,.), \nu)$, and $K = 4\tilde{L}_\Phi B_\Phi$

$$\mathbb{P}\left(\sup_{\zeta \in \mathcal{F}_{\mathcal{W}}} \left|\|f_\mu^* - f_\zeta\|_{\mu_N}^2 - \|f_\mu^* - f_\zeta\|_\mu^2\right| \geq \epsilon + B_{nrm}\right)$$

$$\leq \delta_{\mathcal{C}} + C\exp\left(\log(C_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{4\tilde{L}_\Phi B_\Phi}\right)) - N\min\left\{\frac{\epsilon^2}{256 n_Y^2 \gamma^4 \sigma_X^4}, \frac{\epsilon}{64 n_Y \gamma^2 \sigma_X^2}\right\}\right). \tag{462}$$

We conclude the result by choosing $\epsilon \in [0, 64 n_Y \gamma^2 \sigma_X^2]$. □

D.2 Concentration of Convex functions

In this section, we upper bound the probability of the event, $\mathcal{E}_{cvx}(\epsilon)$ through Lemma 9. In this, we consider strongly and smooth convex function (see assumption 3) through Taylor expansion of the function is always bounded quadratically. Lemma 8 plays an important role in establishing Lemma 9.

Lemma 9 (Concentration of Convex functions)**.** *Consider an n_X-dimensional sub-Gaussian vector $X \sim SG(0, (\sigma_X^2/n_X)I)$, and set of functions $f_\zeta : \mathbb{R}^{n_X} \to \mathbb{R}$ as parameterized by $\zeta \in \mathcal{F}_{\mathcal{W}}$. Let $\mathcal{C}$ be some convex obeying $P(\bigcap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_{\mathcal{C}}$ for i.i.d samples $\{X_i\}_{i=1}^N$. Assume that for any fixed, $\zeta, \zeta' \in \mathcal{F}_{\mathcal{W}}$, and fixed $Z \in \mathcal{C}$, we have*

$$\|f_\zeta(Z) - f_{\zeta'}(Z)\| \leq \tilde{L}_\Phi d(\zeta_1, \zeta_2) \text{ and } \|f_\zeta(Z)\| \leq B_\Phi. \tag{463}$$

Denote

$$B_{nrm}(\mathcal{C}) := \sup_{\zeta \in \mathcal{F}_{\mathcal{W}}} \left|\|f_\mu^* \circ \mathcal{P}_{\mathcal{C}} - f_\zeta \circ \mathcal{P}_{\mathcal{C}}\|_\mu^2 - \|f_\mu^* - f_\zeta\|_\mu^2\right|. \tag{464}$$

where $\mathcal{P}_{\mathcal{C}}(\cdot)$ denotes the Euclidean projection onto the set $\mathcal{C}$. Define

$$K := n_Y L\left[(\gamma^2 + \|g\|_{Lip}^2)\sigma_X^2 + \|g\|_{Lip}^2 \sigma_{Y|X}^2\right]. \tag{465}$$

Then for any $\epsilon \in [0, K]$,

$$\mathbb{P}\left(\sup_{\zeta \in \mathcal{F}_{\mathcal{W}}} |C_{\mu_N}(f_\zeta) - C_\mu(f_\zeta)| \geq \epsilon + B_{nrm}(\mathcal{C})\right) \tag{466}$$

$$\leq \delta_{\mathcal{C}} + 2\exp\left(\log\left(C_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{2B_\ell \tilde{L}_\Phi}\right)\right) - cN\left(\frac{\epsilon}{K}\right)^2\right),$$

for some positive constant, c and $C_{\mathcal{F}_{\mathcal{W}}}(\nu)$ is the ν-net covering number of the set $\mathcal{F}_{\mathcal{W}}$.

Proof. Recall the definitions of the convex functions:

$$C_{\mu_N}(f) := \ell(g,f)_{\mu_N} + \lambda\Omega(f), \text{ and } C_{\mu}(f) := \ell(g,f)_{\mu} + \lambda\Omega(f). \tag{467}$$

The difference between these two terms is

$$|C_{\mu_N}(f) - C_{\mu}(f)| = |\ell(g,f)_{\mu_N} - \ell(g,f)_{\mu}|\,. \tag{468}$$

From assumption 3, $\ell(.,.)$ is second-order differentiable in the second argument. By 2nd-order Taylor's theorem, we have

$$\ell(Y,\hat{Y}) = \ell(Y,\hat{Y}_0) + \langle \nabla_{\hat{Y}}\ell(Y,\hat{Y}_0), \hat{Y}-\hat{Y}_0\rangle + \langle \int_0^1 t\nabla^2_{\hat{Y}}\ell(Y,\hat{Y}_0 + t(\hat{Y}-\hat{Y}_0))dt, (\hat{Y}-\hat{Y}_0)(\hat{Y}-\hat{Y}_0)^T\rangle. \tag{469}$$

Now choose $Y = \hat{Y}_0 = g(X(\omega), E(\omega))$, and $\hat{Y} = f_\zeta(X(\omega))$. As $\ell(Y,Y) = 0$, and $\nabla_{\hat{Y}}\ell(Y,Y) = \mathbf{0}$. Plugging these parameters in the Taylor expansion we have that (ignoring the inputs, $(X(\omega), E(\omega))$ for simplicity),

$$\ell(g,f) = \langle \int_0^1 t\nabla^2_{\hat{Y}}\ell(g, g + t(f_\zeta - g))dt, (f_\zeta - g)(f_\zeta - g)^T\rangle. \tag{470}$$

Now we apply expectation over the measure μ_N, and μ respectively on the above equality. Then we have that

$$\ell(g,f_\zeta)_{\mu} = \langle \int_0^1 t\nabla^2_{\hat{Y}}\ell(g, g + t(f_\zeta - g))dt, (f_\zeta - g)(f_\zeta - g)^T\rangle_{\mu}; \tag{471}$$

$$\ell(g,f_\zeta)_{\mu_N} = \langle \int_0^1 t\nabla^2_{\hat{Y}}\ell(g, g + t(f_\zeta - g))dt, (f_\zeta - g)(f_\zeta - g)^T\rangle_{\mu_N}. \tag{472}$$

Since f_ζ and g are Lipschitz functions and the inputs are sub-Gaussian, we have that $f_\zeta - g$ is a sub-Gaussian vector. As a consequence of Lemma 2.7.6 from Vershynin (2018) we obtain that $(f_\zeta - g)(f_\zeta - g)^T$ follows a sub-exponential distribution, whose concentration is well-studied.

As a consequence of assumption 3 the hessian is bounded, i.e, $\alpha I \preceq \nabla^2_{\hat{Y}}\ell(.,.) \preceq LI$. We can argue that the product of a bounded RV and sub-exponential RV is sub-exponential. Recall Item (iii) from Proposition 2.7.1 of Vershynin (2018). The random variable Z is sub-exponential iff

$$\mathbb{E}_Z\left[e^{\lambda|Z|}\right] \leq e^{\lambda K}; \forall\lambda \in [0, 1/K], \tag{473}$$

for some positive constant, $K \geq 0$.

We now verify if $\langle H(\mathbf{x},\mathbf{z}), \mathbf{x}\mathbf{x}^T\rangle$ is sub-exponential. Given that $\mathbf{x} \sim SG(\sigma_X^2/n_x I_{n_x\times n_x})$, $\mathbf{z}$ is a R.V. Suppose $A \preceq H(\mathbf{x},\mathbf{z}) \preceq B$ *a.s.* Then we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathbf{z}}\left[e^{\lambda\left|\langle H(\mathbf{x},\mathbf{z}), \mathbf{x}\mathbf{x}^T\rangle\right|}\right] &\leq \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[e^{\lambda\|H(\mathbf{x},\mathbf{z})\|_2\|\mathbf{x}\mathbf{x}^T\|_2}\right], &(474)\\
&\leq \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[e^{\lambda\max\{\rho(A),\rho(B)\}\|\mathbf{x}\|_2^2}\right], &(475)\\
& &(476)
\end{aligned}$$

where, $\rho(A)$ is the spectral radius of the matrix, A. Since, $\mathbf{x} \sim SG(\sigma_X^2/n_x I_{n_x\times n_x})$, we have $\|\mathbf{x}\|^2 \sim SE(\sigma_X^2)$. Then,

$$\mathbb{E}_{\mathbf{x},\mathbf{z}}\left[e^{\lambda\left|\langle H(\mathbf{x},\mathbf{z}), \mathbf{x}\mathbf{x}^T\rangle\right|}\right] \leq e^{\lambda\max\{\rho(A),\rho(B)\}\sigma_X^2},$$

implies that $\langle H(\mathbf{x},\mathbf{z}), \mathbf{x}\mathbf{x}^T\rangle \sim SE(\max\{\rho(A),\rho(B)\}\sigma_X^2)$. From this analysis we have

$$\langle \int_0^1 t\nabla^2_{\hat{Y}}\ell(g, g + t(f_\zeta - g))dt, \underbrace{(f_\zeta - g)(f_\zeta - g)^T}_{\sim SE\left(\left[(\|f_\zeta\|^2_{\text{Lip}}+\|g\|^2_{\text{Lip}})\sigma_X^2+\|g\|^2_{\text{Lip}}\sigma^2_{Y|X}\right]I_{n_y\times n_y}\right)} \rangle$$

$$\sim SE\left(n_Y \frac{L}{2}\left[(\|f_\zeta\|_{\text{Lip}}^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2\sigma_{Y|X}^2\right]\right). \tag{477}$$

For convex functions, we know that $0 \leq \alpha \leq L$. As a consequence, we have $\frac{\alpha}{2}I \preceq \int_0^1 t\nabla_{\hat{Y}}^2 \ell(g, g+t(f_\zeta - g))dt \preceq \frac{L}{2}I$. Now we apply sub-exponential concentration for a fixed $\zeta \in \mathcal{F}_\mathcal{W}$, yielding

$$\begin{aligned}
&\mathbb{P}\left(|C_{\mu_N}(f_\zeta) - C_\mu(f_\zeta)| \geq \epsilon\right) \\
&\leq 2\exp\Big(-cN\min\Big\{\left(\frac{2\epsilon}{n_Y L\left[(\gamma^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2\sigma_{Y|X}^2\right]}\right)^2, \\
&\qquad \frac{2\epsilon}{n_Y L\left[(\gamma^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2\sigma_{Y|X}^2\right]}\Big\}\Big),
\end{aligned} \tag{478}$$

for some positive constant, $c \geq 0$.

Next, we move on to obtain a uniform concentration for all $\zeta \in \mathcal{F}_\mathcal{W}$. Now we apply covering argument from Lemma 14, and set

$$g_\theta = \ell(g, f_\zeta). \tag{479}$$

We need to check if the function, g, is Lipschitz on some metric and convex set $\mathcal{C} \subseteq \mathbb{R}^{n_X}$, choose for any $Z \in \mathcal{C}$. We have $P(\bigcap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_\mathcal{C}$. Recall that

1. $\forall \zeta_1, \zeta_2 \in \mathcal{F}_\mathcal{W} : \|f_{\zeta_1}(Z) - f_{\zeta_2}(Z)\| \leq \tilde{L}_\Phi d(\zeta_1, \zeta_2)$, for all $Z \in \mathcal{C}$.
2. $\forall \zeta \in \mathcal{F}_\mathcal{W} : \|\nabla_{\hat{Y}}\ell(g(Z), f_\zeta(Z))\| \leq B_\ell$, for all $Z \in \mathcal{C}$.
3. For a fixed $\zeta \in \mathcal{F}_\mathcal{W}$,

$$\left|\mathbb{E}\left[\|f_\mu^*(\mathcal{P}_\mathcal{C}(X)) - f_\zeta(\mathcal{P}_\mathcal{C}(X))\|^2 - \|f_\mu^*(X) - f_\zeta(X)\|^2\right]\right| \leq B_{nrm}(\mathcal{C}). \tag{480}$$

From Taylor expansion we have that,

$$\begin{aligned}
|g_{\theta_1} - g_{\theta_2}| &= |\langle \int_t \nabla_{\hat{Y}}\ell(g, f_{\zeta_1} + t(f_{\zeta_2} - f_{\zeta_1}))dt, f_{\zeta_1} - f_{\zeta_2}\rangle|, \\
&\leq \|\int_t \nabla_{\hat{Y}}\ell(g, f_{\zeta_1} + t(f_{\zeta_2} - f_{\zeta_1}))dt\| \|f_{\zeta_2} - f_{\zeta_1}\|, \\
&\leq B_\ell \|f_{\zeta_2} - f_{\zeta_1}\|, \\
&\leq B_\ell \tilde{L}_\Phi d(\zeta_1, \zeta_2).
\end{aligned}$$

From Lemma 14 we have

$$\mathbb{P}\left(\sup_{\zeta \in \mathcal{F}_\mathcal{W}} |C_{\mu_N}(f_\zeta) - C_\mu(f_\zeta)| \geq \epsilon\right) \tag{481}$$

$$\begin{aligned}
&\leq \delta_\mathcal{C} + 2\exp\left(\log\left(C_{\mathcal{F}_\mathcal{W}}\left(\frac{\epsilon}{2B_\ell\tilde{L}_\Phi}\right)\right) - cN\min\left\{\left(\frac{\epsilon}{n_Y L\left[(\gamma^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2\sigma_{Y|X}^2\right]}\right)^2,\right.\right. \\
&\qquad \left.\left.\frac{\epsilon}{n_Y L\left[(\gamma^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2\sigma_{Y|X}^2\right]}\right\}\right),
\end{aligned} \tag{482}$$

for some positive constant, c. Now restrict $\epsilon \in \left[0, n_Y L\left[(\gamma^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2\sigma_{Y|X}^2\right]\right]$. This completes our proof. □

D.3 Concentration of Equilibria

In this section, we upper bound the probability of the event, $\mathcal{E}_{eql}(\epsilon)$ through Lemma 10. We first present the concentration of bi-Lipschtiz functions for sub-Gaussian inputs.

Proposition 4. *Let $X \sim SG(\frac{\sigma_X^2}{n_x} I_{n_x \times n_x})$ and $Y|X \sim SG(\frac{\sigma_{Y|X}^2}{n_y} I_{n_y \times n_y})$, where $\sigma_{Y|X}$ is independent of $X = x$, then for any Lipschitz function, $\phi : \mathcal{Z} \to \mathbb{R}$ and a function, $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, satisfies the following. Then $\phi(f(X,Y)) \sim SG\left(4\|\phi\|_{Lip}^2 \|f\|_{Lip}^2 \left[\sigma_X^2 + \sigma_{Y|X}^2\right]\right)$.*

$$\|f(X_2, Y_2) - f(X_1, Y_1)\| \leq \|f\|_{Lip} \left[\|X_2 - X_1\| + \|Y_2 - Y_1\|\right]. \tag{483}$$

Proof. Let us compute the moments of the random variable $\phi(f(X,Y))$ for any Lipschitz function, $\phi : \mathcal{Z} \to \mathbb{R}$.

By symmetrization, we have

$$\mathbb{E}_{X,Y}\left[\exp(\lambda[\phi(f(X,Y)) - \mathbb{E}_{X,Y}[\phi(f(X,Y))]])\right]$$

$$= \mathbb{E}_{X,Y}\left[\exp(\lambda[\phi(f(X,Y)) - \mathbb{E}_{X',Y'}[\phi(f(X',Y'))])\right] \tag{484}$$

$$= \mathbb{E}_{X,Y}\left[\exp(\lambda \mathbb{E}_{X',Y'}[\phi(f(X,Y)) - \phi(f(X',Y'))])\right]. \tag{485}$$

By Jensen's inequality we have

$$\leq \mathbb{E}_{X,Y,X',Y'}\left[\exp(\lambda[\phi(f(X,Y)) - \phi(f(X',Y'))])\right], \tag{486}$$

By Lipschitz continuity we have

$$\leq \mathbb{E}_{X,Y,X',Y'}\left[\exp(\|\phi\|_{\text{Lip}} \lambda[\|f(X,Y) - f(X',Y')\|]\right]. \tag{487}$$

By construction, we have.

$$\leq \mathbb{E}_{X,Y,X',Y'}\left[\exp(\|\phi\|_{\text{Lip}} \|f\|_{\text{Lip}} \lambda[\|X - X'\| + \|Y - Y'\|]\right]. \tag{488}$$

By Cauchy-Schwartz's inequality we have

$$\leq \mathbb{E}_{X,Y,X',Y'}\left[\exp(\|\phi\|_{\text{Lip}} \|f\|_{\text{Lip}} \lambda[\|X\| + \|X'\| + \|Y\| + \|Y'\|]\right]. \tag{489}$$

As the symmetrized random variables are independent,

$$\leq \mathbb{E}_{X,Y}\left[\exp(2\|\phi\|_{\text{Lip}} \|f\|_{\text{Lip}} \lambda[\|X\| + \|Y'\|]\right]. \tag{490}$$

Now perform conditional expectation,

$$\leq \mathbb{E}_X \mathbb{E}_{Y|X}\left[\exp(2\|\phi\|_{\text{Lip}} \|f\|_{\text{Lip}} \lambda[\|X\| + \|Y\|]\right], \tag{491}$$

$$\leq \mathbb{E}_X \exp(2\|\phi\|_{\text{Lip}} \|f\|_{\text{Lip}} \lambda[\|X\|]) \mathbb{E}_{Y|X}\left[\exp(2\|\phi\|_{\text{Lip}} \|f\|_{\text{Lip}} \lambda[\|Y\|]\right], \tag{492}$$

$$\leq \exp\left(\frac{\lambda^2}{2} 4\|\phi\|_{\text{Lip}}^2 \|f\|_{\text{Lip}}^2 \left[\sigma_X^2 + \sigma_{Y|X}^2\right] + \lambda[\mathbb{E}_X[\|X\|] + \mathbb{E}_{Y|X}[\|Y\|]]\right), \tag{493}$$

$$\leq K \exp\left(\frac{\lambda^2}{2} 4\|\phi\|_{\text{Lip}}^2 \|f\|_{\text{Lip}}^2 \left[\sigma_X^2 + \sigma_{Y|X}^2\right]\right). \tag{494}$$

for some constant, $K \geq 0$.

This implies that, $\phi(f(X,Y)) \sim SG\left(4\|\phi\|_{\text{Lip}}^2 \|f\|_{\text{Lip}}^2 \left[\sigma_X^2 + \sigma_{Y|X}^2\right]\right)$. □

With the above result, we now state and prove the probability of the event, $\mathcal{E}_{eql}(\epsilon)$.

Lemma 10 (Concentration of Equilibria). *Consider an n_X-dimensional sub-Gaussian vector $X \sim SG(0, (\sigma_X^2/n_X)I)$, and set of functions $f_\zeta : \mathbb{R}^{n_X} \to \mathbb{R}$ as parameterized by $\zeta \in \mathcal{F}_\mathcal{W}$. Let $\mathcal{C}$ be some convex obeying $P(\bigcap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_\mathcal{C}$ for i.i.d samples $\{X_i\}_{i=1}^N$. Assume that for any fixed, $\zeta_1, \zeta_2 \in \mathcal{F}_\mathcal{W}$, and fixed $Z \in \mathcal{C}$, we have*

$$\|\nabla_{\hat{Y}} \ell(g(Z), f_\zeta(Z))\| \leq B_\ell, \ \|f_\zeta(Z)\| \leq B_\Phi, \text{ and} \tag{495}$$

$$\|f_\zeta(Z) - f_{\zeta'}(Z)\| \leq \tilde{L}_\Phi d(\zeta, \zeta'). \tag{496}$$

In addition, we have that,

$$\sup_{\zeta \in \mathcal{F}_\mathcal{W}} \left| \mathbb{E}\left[\langle \nabla_{\hat{Y}} \ell\left(g \circ \mathcal{P}_\mathcal{C}, f_\zeta \circ \mathcal{P}_\mathcal{C}\right), f_{\zeta'} \circ \mathcal{P}_\mathcal{C} \rangle_\mu - \langle \nabla_{\hat{Y}} \ell\left(g, f_\zeta\right), f_{\zeta'} \rangle_\mu \right] \right| = B_{eql}(\mathcal{C}). \tag{497}$$

Define

$$K := 4 n_y \gamma \|\nabla_{\hat{Y}} \ell\|_{Lip} \sigma_X \sqrt{(\gamma^2 + \|g\|_{Lip}^2)\sigma_X^2 + \|g\|_{Lip}^2 \sigma_{E|X}^2}. \tag{498}$$

Then for any $\epsilon \in [0, K]$,

$$\mathbb{P}\left(\sup_{\zeta \in \mathcal{F}_\mathcal{W}} \left| \langle \nabla_{\hat{Y}} \ell(g, f_\zeta), f_\zeta \rangle_{\mu_N} - \langle \nabla_{\hat{Y}} \ell(g, f_\zeta), f_\zeta \rangle_\mu \right| \geq \epsilon + B_{eql}(\mathcal{C}) \right) \leq$$

$$\delta_\mathcal{C} + c \exp\left(\log\left(\mathcal{C}_{\mathcal{F}_\mathcal{W}}\left(\frac{\epsilon}{2\tilde{L}_\Phi \left[B_\ell + B_\Phi L\right]} \right) \right) - N \frac{\epsilon^2}{K^2} \right), \tag{499}$$

for some positive constant, c and $C_{\mathcal{F}_\mathcal{W}}(\nu)$ is the ν-net covering number of the set $\mathcal{F}_\mathcal{W}$..

Proof. From Assumptions 1-5, we have that for any $g_1, f_1, g_2, f_2 \in L^2(\mu)$

$$\|\nabla_{\hat{Y}} \ell(g_2(X(\omega),E(\omega)), f_2(X(\omega))) - \nabla_{\hat{Y}} \ell(g_1(X(\omega), E(\omega)), f_1(X(\omega)))\| \tag{500}$$

$$\leq \|\nabla_{\hat{Y}} \ell\|_{\text{Lip}} \left[\|g_2(X(\omega), E(\omega)) - g_1(X(\omega), E(\omega))\| + \|f_2(X(\omega)) - f_1(X(\omega))\| \right]. \tag{501}$$

Since $X(\omega)$ and $E(\omega)$ are Lipschitz concentrated R.Vs, it holds that

$$g(X,E)|E \sim SG(\|g\|_{\text{Lip}}^2 \sigma_X^2 I_{n_y \times n_y}), g(X,E)|X \sim SG(\|g\|_{\text{Lip}}^2 \sigma_{E|X}^2 I_{n_y \times n_y}), \tag{502}$$

$$\text{and } f_\zeta(X) \sim SG(\|f_\zeta\|_{\text{Lip}}^2 \sigma_X^2 I_{n_y \times n_y}). \tag{503}$$

From Proposition 4 we have

$$\nabla_{\hat{Y}} \ell(g(X(\omega), E(\omega)), f_\zeta(X(\omega))) \sim$$

$$SG\left(4\|\nabla_{\hat{Y}} \ell\|_{\text{Lip}}^2 \left[(\|f_\zeta\|_{\text{Lip}}^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2 \sigma_{E|X}^2 \right] I_{n_y} \right). \tag{504}$$

Now we have the inner product between two sub-Gaussian random variables from Proposition 7 we have that the result is sub-exponential, i.e.,

$$\Big\langle \underbrace{\nabla_{\hat{Y}} \ell(g(X(\omega), E(\omega)), f_\zeta(X(\omega)))}_{\sim SG\left(4\|\nabla_{\hat{Y}} \ell\|_{\text{Lip}}^2 \left[(\|f_\zeta\|_{\text{Lip}}^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2 \sigma_{E|X}^2 \right] I_{n_y \times n_y}\right)}, \underbrace{f_\zeta(X(\omega))}_{\sim SG(\|f_\zeta\|_{\text{Lip}}^2 \sigma_X^2 I_{n_y})} \Big\rangle$$

$$\sim SE\left(2n_y \|\nabla_{\hat{Y}} \ell\|_{\text{Lip}} \|f_\zeta\|_{\text{Lip}} \sigma_X \sqrt{(\|f_\zeta\|_{\text{Lip}}^2 + \|g\|_{\text{Lip}}^2)\sigma_X^2 + \|g\|_{\text{Lip}}^2 \sigma_{E|X}^2} \right). \tag{505}$$

The class of functions, f_ζ for $\zeta \in \mathcal{F}_\mathcal{W}$, has bounded Lipschitz constant γ. As a consequence of the sub-exponential concentration bound from Theorem 2.8.1 in Vershynin (2018), we have that for a fixed $\zeta \in \mathcal{F}_\mathcal{W}$,

$$\mathbb{P}\left(\left| \langle \nabla_{\hat{Y}} \ell(g, f_\zeta), f_\zeta \rangle_{\mu_N} - \langle \nabla_{\hat{Y}} \ell(g, f_\zeta, f_\zeta \rangle_\mu \right| \geq \epsilon \right) \leq C \exp\left(-N \min\left\{ \frac{\epsilon^2}{K^2}, \frac{\epsilon}{K} \right\} \right). \tag{506}$$

where $K := 2n_y\gamma\|\nabla_{\hat{Y}}\ell\|_{\text{Lip}}\sigma_X\sqrt{(\gamma^2+\|g\|^2_{\text{Lip}})\sigma_X^2+\|g\|^2_{\text{Lip}}\sigma^2_{E|X}}$ and some positive constant, C.

Now we move on to providing a uniform concentration in the inequality (506). We will apply uniform concentration result from Lemma 14, for this set:

$$g_\theta = \langle\nabla_{\hat{Y}}\ell(g, f_\zeta), f_\zeta\rangle. \tag{507}$$

Recall the below items:

1. For a fixed $Z \in \mathcal{C}$ we have $\forall\zeta \in \mathcal{F}_\mathcal{W} : \|\nabla_{\hat{Y}}\ell(g(Z), f_\zeta(Z))\| \le B_\ell$.
2. For a fixed $Z \in \mathcal{C}$ we have $\forall\zeta \in \mathcal{F}_\mathcal{W} : \|f_\zeta(Z)\| \le B_\Phi$.
3. For a fixed $Z \in \mathcal{C}$ we have $\forall\zeta,\zeta' \in \mathcal{F}_\mathcal{W} : \|f_\zeta(Z) - f_{\zeta'}(Z)\| \le \tilde{L}_\Phi d(\zeta,\zeta')$.
4. For a any $\hat{Y}_1, \hat{Y}_2 \in \mathbb{R}^{n_Y}$ we have $\|\nabla_{\hat{Y}}\ell(Y,\hat{Y}_1) - \nabla_{\hat{Y}}\ell(Y,\hat{Y}_2)\| \le L\|\hat{Y}_1 - \hat{Y}_2\|$.
5. For a fixed $\zeta \in \mathcal{F}_\mathcal{W}$,

$$\sup_{\zeta\in\mathcal{F}_\mathcal{W}} \left|\mathbb{E}\left[\langle\nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_\mathcal{C}, f_\zeta\circ\mathcal{P}_\mathcal{C}\right), f_\zeta\circ\mathcal{P}_\mathcal{C}\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g, f_\zeta\right), f_\zeta\rangle_\mu\right]\right| = B_{eql}(\mathcal{C}). \tag{508}$$

Now we check the Lipschitz continuity of the function g_θ:

$$\begin{aligned}
|g_{\theta_1} - g_{\theta_2}| &= \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_1}\rangle - \langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_2}),f_{\zeta_2}\rangle\right|,\\
&= \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_1}-f_{\zeta_2}\rangle - \langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_2})-\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_2}\rangle\right|,\\
&\le \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_1}-f_{\zeta_2}\rangle\right| + \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_2})-\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_2}\rangle\right|,\\
&\le \|\nabla_{\hat{Y}}\ell(g,f_{\zeta_1})\|\|f_{\zeta_1}-f_{\zeta_2}\| + \|\nabla_{\hat{Y}}\ell(g,f_{\zeta_2})-\nabla_{\hat{Y}}\ell(g,f_{\zeta_1})\|\|f_{\zeta_2}\|,\\
&\le B_\ell\|f_{\zeta_1}-f_{\zeta_2})\| + B_\Phi L\|f_{\zeta_1}-f_{\zeta_2}\|,\\
&\le \tilde{L}_\Phi\left[B_\ell + B_\Phi L\right]d(\zeta_1,\zeta_2).
\end{aligned}$$

Then from Lemma 14 we have that

$$\mathbb{P}\left(\sup_{\zeta\in\mathcal{F}_\mathcal{W}}\left|\langle\nabla_{\hat{Y}}\ell(g,f_\zeta),f_\zeta\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell(g,f_\zeta),f_\zeta\rangle_\mu\right| \ge \epsilon + B_{eql}(\mathcal{C})\right) \tag{509}$$

$$\le \delta_\mathcal{C} + C\exp\left(\log\left(\mathcal{C}_{\mathcal{F}_\mathcal{W}}\left(\frac{\epsilon}{2\tilde{L}_\Phi\left[B_\ell + B_\Phi L\right]}\right)\right) - N\min\left\{\frac{\epsilon^2}{4K^2}, \frac{\epsilon}{2K}\right\}\right). \tag{510}$$

□

D.4 Concentration of Polar

In this section, we compute the probability of the occurrence of the event, $\mathcal{E}_{plr}(\epsilon)$ through Lemma 11. The analysis of $\mathcal{E}_{plr}(\epsilon)$ resembles to that of $\mathcal{E}_{eql}(\epsilon)$ following similar arguments.

Lemma 11 (Concentration of Polar). *Consider an n_X-dimensional sub-Gaussian vector $X \sim SG(0,(\sigma_X^2/n_X)I)$, and set of functions $f_\zeta : \mathbb{R}^{\aleph\times\aleph} \to \mathbb{R}$ as parameterized by $\zeta \in \mathcal{F}_\mathcal{W}$. Let $\mathcal{C}$ be some convex obeying $P(\bigcap_{i=1}^N X_i \in \mathcal{C}) \ge 1-\delta_\mathcal{C}$ for i.i.d samples $\{X_i\}_{i=1}^N$. Assume that for any fixed, $\zeta_1,\zeta_2 \in \mathcal{F}_\mathcal{W}$, $\zeta_1',\zeta_2' \in \mathcal{F}_\theta$, and fixed $Z \in \mathcal{C}$, we have*

$$\|\nabla_{\hat{Y}}\ell(g(Z),f_\zeta(Z))\| \le B_\ell,\ \|f_\zeta(Z)\| \le B_\Phi, \tag{511}$$

$$\|f_{\zeta_1}(Z) - f_{\zeta_2}(Z)\| \le \tilde{L}_\Phi d(\zeta_1,\zeta_2),\ \text{and}\ \|f_{\zeta_1'}(Z) - f_{\zeta_2'}(Z)\| \le \tilde{L}_\phi d(\zeta_1',\zeta_2'). \tag{512}$$

In addition, we have that,

$$\sup_{\zeta\in\mathcal{F}_\mathcal{W},\zeta'\in\mathcal{F}_\theta}\left|\mathbb{E}\left[\langle\nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_\mathcal{C}, f_\zeta\circ\mathcal{P}_\mathcal{C}\right), f_{\zeta'}\circ\mathcal{P}_\mathcal{C}\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g, f_\zeta\right), f_{\zeta'}\rangle_\mu\right]\right| = B_{plr}(\mathcal{C}). \tag{513}$$

Define

$$K := 4n_Y\|\nabla_{\hat{Y}}\ell\|_{Lip}L_\phi\sigma_X\sqrt{(\gamma^2+\|g\|^2_{Lip})\sigma_X^2+\|g\|^2_{Lip}\sigma^2_{E|X}}. \tag{514}$$

Then for any $\epsilon \in [0, K]$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\zeta\in\mathcal{F}_{\mathcal{W}}} : \left|\Omega^{\circ}_{\mu_N}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) - \Omega^{\circ}_{\mu}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right)\right| \geq \epsilon + B_{plr}(\mathcal{C})\right) \\
&\leq \delta_{\mathcal{C}} + c\exp\left(\log\left(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\left(\frac{\epsilon}{8\max\{\tilde{L}_{\phi}B_{\ell}, L\tilde{L}_{\Phi}B_{\Phi}\}}\right)\right)\right. \\
&\qquad\left. + \log\left(\mathcal{C}_{\mathcal{F}_{\theta}}\left(\frac{\epsilon}{8\max\{\tilde{L}_{\phi}B_{\ell}, L\tilde{L}_{\Phi}B_{\Phi}\}}\right)\right) - N\frac{\epsilon^2}{K^2}\right),
\end{aligned} \tag{515}$$

for some positive constant, c *and* $C_{\mathcal{F}_{\mathcal{W}}}(\nu)$ *(and* $C_{\mathcal{F}_{\theta}}(\nu)$*) is the* ν*-net covering number of the set* $\mathcal{F}_{\mathcal{W}}$ *(and* $\mathcal{F}_{\theta}$*).*

Proof. Recall the definition of the polar in Equation 6:

$$\Omega^{\circ}_{\mu_N}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) := \sup_{\zeta'\in\mathcal{F}_{\theta}} \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu_N}, \tag{516}$$

$$\Omega^{\circ}_{\mu}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) := \sup_{\zeta'\in\mathcal{F}_{\theta}} \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu}. \tag{517}$$

Now, by taking the difference between the above two polars, we have

$$\begin{aligned}
&\left|\Omega^{\circ}_{\mu_N}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) - \Omega^{\circ}_{\mu}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right)\right| \\
&\qquad = \left|\sup_{\zeta'\in\mathcal{F}_{\theta}} \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu_N} - \sup_{\zeta'\in\mathcal{F}_{\theta}} \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu}\right|.
\end{aligned} \tag{518}$$

Denote, $\zeta'^{*}_{\mu} = arg\sup_{\zeta'\in\mathcal{F}_{\theta}}\langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu}$ and $\zeta'^{*}_{\mu_N} = arg\sup_{\zeta'\in\mathcal{F}_{\theta}}\langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu_N}$, then, by definition we have that

$$\begin{aligned}
&-\langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'^{*}_{\mu}}\rangle_{\mu_N} + \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'^{*}_{\mu}}\rangle_{\mu} \leq \Omega^{\circ}_{\mu_N}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) - \Omega^{\circ}_{\mu}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) \leq \\
&\qquad \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'^{*}_{\mu}}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'^{*}_{\mu}}\rangle_{\mu}.
\end{aligned} \tag{519}$$

Applying modulus on both sides we obtain

$$\begin{aligned}
&\left|\Omega^{\circ}_{\mu_N}\Big(\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big)\Big) - \Omega^{\circ}_{\mu}\Big(\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big)\Big)\right| \\
&\qquad \leq \max\left\{\left|\langle\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big), f_{\zeta'^{*}_{\mu}}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big), f_{\zeta'^{*}_{\mu}}\rangle_{\mu}\right|,\right. \\
&\qquad \left.\left|\langle\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big), f_{\zeta'^{*}_{\mu_N}}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big), f_{\zeta'^{*}_{\mu_N}}\rangle_{\mu}\right|\right\} \\
&\qquad \leq \sup_{\zeta'\in\mathcal{F}_{\theta}}\left|\langle\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big), f_{\zeta'}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\Big(g, f_{\zeta}\Big), f_{\zeta'}\rangle_{\mu}\right|.
\end{aligned} \tag{520}$$

Now, we have to compute a lower bound on

$$\mathbb{P}\left(\sup_{\zeta'\in\mathcal{F}_{\theta}}\left|\langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right), f_{\zeta'}\rangle_{\mu}\right| \leq \epsilon\right). \tag{521}$$

The computation of Equation (521) is similar to that of Lemma 10. We can re-write the concentration of polars and apply the monotonicty of probability in inequality (520) by doing this we have

$$\mathbb{P}\left(\sup_{\zeta\in\mathcal{F}_{\mathcal{W}}} : \left|\Omega^{\circ}_{\mu_N}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right) - \Omega^{\circ}_{\mu}\left(\nabla_{\hat{Y}}\ell\left(g, f_{\zeta}\right)\right)\right| \geq \epsilon\right)$$

$$\leq \mathbb{P}\left(\sup_{\zeta\in\mathcal{F}_\Phi,\zeta'\in\mathcal{F}_\theta} : \left|\langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_{\mu}\right| \leq \epsilon\right). \tag{522}$$

As the data follow the sub-Gaussian distribution from Proposition **??** and Proposition 7, we have

$$\Big\langle \underbrace{\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right)}_{\sim SG\left(4\|\nabla_{\hat{Y}}\ell\|^2_{\text{Lip}}\left[(\|f_\zeta\|^2_{\text{Lip}}+\|g\|^2_{\text{Lip}})\sigma^2_X+\|g\|^2_{\text{Lip}}\sigma^2_{E|X}\right]I_{n_Y\times n_Y}\right)}, \underbrace{f_{\zeta'}}_{\sim SG(\|f_{\zeta'}\|^2_{\text{Lip}}\sigma^2_X I_{n_Y\times n_Y})} \Big\rangle_{\mu_N}$$

$$\sim SE\left(2n_Y\|\nabla_{\hat{Y}}\ell\|_{\text{Lip}}\|f_{\zeta'}\|_{\text{Lip}}\sigma_X\sqrt{(\|f_\zeta\|^2_{\text{Lip}}+\|g\|^2_{\text{Lip}})\sigma^2_X+\|g\|^2_{\text{Lip}}\sigma^2_{E|X}}\right). \tag{523}$$

From Assumption 6 the class, $\mathcal{F}_\Phi$ has Lipschitz constant at most, γ. From the assumption 4 $\mathcal{F}_\theta$ has a Lipschitz constant at most γ_θ. Therefore, the inner product described above is concentrated as a consequence of Theorem 2.8.1 from Vershynin (2018). Now for a fixed $\zeta\in\mathcal{F}_\mathcal{W},\zeta'\in\mathcal{F}_\theta$, we have that

$$\mathbb{P}\left(\left|\langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_{\mu}\right| \leq \epsilon\right) \leq C\exp\left(-N\min\left\{\frac{\epsilon^2}{K^2},\frac{\epsilon}{K}\right\}\right). \tag{524}$$

where, $K = 2n_Y\|\nabla_{\hat{Y}}\ell\|_{\text{Lip}}L_\phi\sigma_X\sqrt{(\gamma^2+\|g\|^2_{\text{Lip}})\sigma^2_X+\|g\|^2_{\text{Lip}}\sigma^2_{E|X}}$.

Now we utilize Lemma 14 to have this concentration uniformly for all, $\zeta\in\mathcal{F}_\mathcal{W},\zeta'\in\mathcal{F}_\theta$. Set

$$g_\theta = \langle\nabla_{\hat{Y}}\ell(g,f_\zeta),f_{\zeta'}\rangle. \tag{525}$$

Recall the below items:

1. For a fixed $Z\in\mathcal{C}$ we have $\forall\zeta\in\mathcal{F}_\mathcal{W}:\|\nabla_{\hat{Y}}\ell(g(Z),f_\zeta(Z))\|\leq B_\ell$.
2. For a fixed $Z\in\mathcal{C}$ we have $\forall\zeta\in\mathcal{F}_\mathcal{W}:\|f_\zeta(Z)\|\leq B_\Phi$.
3. For a fixed $Z\in\mathcal{C}$ we have $\forall\zeta,\zeta'\in\mathcal{F}_\mathcal{W}:\|f_\zeta(Z)-f_{\zeta'}(Z)\|\leq\tilde{L}_\Phi d(\zeta,\zeta')$.
4. For a fixed $Z\in\mathcal{C}$ we have $\forall\zeta,\zeta'\in\mathcal{F}_\theta:\|f_\zeta(Z)-f_{\zeta'}(Z)\|\leq\tilde{L}_\phi d(\zeta,\zeta')$.
5. For a any $\hat{Y}_1,\hat{Y}_2\in\mathbb{R}^{n_Y}$ we have $\|\nabla_{\hat{Y}}\ell(Y,\hat{Y}_1)-\nabla_{\hat{Y}}\ell(Y,\hat{Y}_2)\|\leq L\|\hat{Y}_1-\hat{Y}_2\|$.
6. For a fixed $\zeta\in\mathcal{F}_\mathcal{W},\zeta'\in\mathcal{F}_\theta$,

$$\sup_{\zeta\in\mathcal{F}_\mathcal{W},\zeta'\in\mathcal{F}_\theta}\left|\mathbb{E}\left[\langle\nabla_{\hat{Y}}\ell\left(g\circ\mathcal{P}_\mathcal{C},f_\zeta\circ\mathcal{P}_\mathcal{C}\right),f_{\zeta'}\circ\mathcal{P}_\mathcal{C}\rangle_\mu - \langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_\mu\right]\right| = B_{plr}(\mathcal{C}). \tag{526}$$

Now we check the Lipschitzness of g:

$$\begin{aligned}
|g_{\theta_1}-g_{\theta_2}| &= \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_1'}\rangle - \langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_2}),f_{\zeta_2'}\rangle\right| \\
&= \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_1'}-f_{\zeta_2'}\rangle - \langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_2})-\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_2'}\rangle\right|, \\
&\leq \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_1'}-f_{\zeta_2'}\rangle\right| + \left|\langle\nabla_{\hat{Y}}\ell(g,f_{\zeta_2})-\nabla_{\hat{Y}}\ell(g,f_{\zeta_1}),f_{\zeta_2'}\rangle\right|, \\
&\leq \|\nabla_{\hat{Y}}\ell(g,f_{\zeta_1})\|\|f_{\zeta_1'}-f_{\zeta_2'}\|,+\|\nabla_{\hat{Y}}\ell(g,f_{\zeta_2})-\nabla_{\hat{Y}}\ell(g,f_{\zeta_1})\|\|f_{\zeta_2'}\|, \\
&\leq B_\ell\tilde{L}_\phi d(\zeta_1',\zeta_2') + B_\Phi L\tilde{L}_\Phi d(\zeta_1,\zeta_2), \\
&\leq 2\max\{\tilde{L}_\phi B_\ell, L\tilde{L}_\Phi B_\Phi\}\max\{d(\zeta_1',\zeta_2'),d(\zeta_1,\zeta_2)\}.
\end{aligned}$$

Now, we have a product of two metric spaces whose metric is a maximum of individual metrics, therefore simply we can upper bound the covering number by product of these two metric spaces.,i.e.,

$$\mathcal{N}(\mathcal{F}_\mathcal{W}\times\mathcal{F}_\theta,\|\cdot\|_{\infty,d(.,.)},\nu)\leq\mathcal{N}(\mathcal{F}_\mathcal{W},d(.,.),\nu)\mathcal{N}(\mathcal{F}_\theta,d(.,.),\nu). \tag{527}$$

From Lemma 14 we have that

$$\mathbb{P}\left(\sup_{\zeta\in\mathcal{F}_\mathcal{W},\zeta'\in\mathcal{F}_\theta} : \left|\langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_{\mu_N} - \langle\nabla_{\hat{Y}}\ell\left(g,f_\zeta\right),f_{\zeta'}\rangle_{\mu}\right| \geq \epsilon + B_{plr}(\mathcal{C})\right) \tag{528}$$

$$\leq \delta_{\mathcal{C}} + C \exp\Big(\log\Big(\mathcal{C}_{\mathcal{F}_{\mathcal{W}}}\Big(\frac{\epsilon}{8\tilde{L}_{\Phi} \max\{B_{\ell}, LB_{\Phi}\}} \Big)\Big) + \log\Big(\mathcal{C}_{\mathcal{F}_{\theta}}\Big(\frac{\epsilon}{8\tilde{L}_{\Phi} \max\{B_{\ell}, LB_{\Phi}\}} \Big)\Big) - N \min\Big\{ \frac{\epsilon^2}{4K^2}, \frac{\epsilon}{2K} \Big\} \Big). \tag{529}$$

This completes our result. □

E NUMERICAL EXPERIMENTS

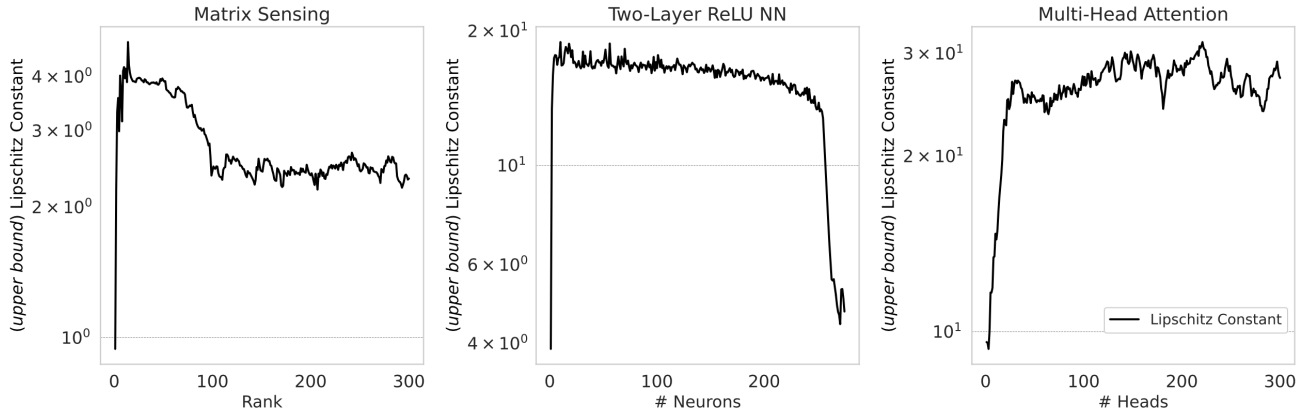

Figure 1: Numerical simulations of the Lipschitz constant (or upper bound thereof) obtained for different model widths (r).

In this section, we present numerical simulations for the problems of low-rank matrix sensing, two-layer ReLU neural networks, and multi-head attention. In each simulation shown in Figure E, we generated data using a teacher model with random initialization of parameters, $Y = \Phi_{r^*}(\{W_j\})(X) + \epsilon$, where $r^* = 64$, $X \sim \mathcal{N}(0, (\sigma_X^2/n_X)I)$, and $\epsilon \in \mathcal{N}(0, (\sigma_E^2/n_Y)I)$. We used gradient descent to reach a stationary point for each R (rank, number of neurons, or number of heads), starting from 1 and increasing up to 300. The first factor was initialized with small-scale random values. For each subsequent factor, we initialized the new factor with the supremum obtained from the polar equation (14), following the algorithm in Haeffele and Vidal (2015, 2020).

In each problem shown in Figure E, we plot the upper bounds on the Lipschitz constant for these problems. For matrix sensing, the Lipschitz constant is trivially upper-bounded by $\|\mathbf{U}\mathbf{V}^T\|_2$; for the ReLU neural network, it is upper-bounded by $\|\mathbf{U}\|_2\|\mathbf{V}\|_2$; and for multi-head attention, it is upper-bounded by $\sum_{j=1}^{r} \|\mathbf{V_j}\|_2$. We can observe from Figure E upper bounds on the Lipschitz constants are uniformly bounded, indicating that our Assumption 6 is realistic and holds empirically.

We conjecture that it is possible to show that the Lipschitz constants are uniformly bounded for any stationary. However, the analysis of this is beyond the scope of this work. Similar analyses based on gradient descent can be found in Oymak and Soltanolkotabi (2019).

F OTHER RELATED WORKS

In this section, we provide a comprehensive study of the related works of the applications that are of the concern in this work.

Statistical Learning Theory (SLT): SLT provides a theoretical framework for analyzing generalization error, often producing results of the form (530). The seminal work by Vapnik (2000) established a systematic approach to deriving bounds of this nature. Over time, various approaches in SLT have attempted to estimate $\epsilon(\mathcal{F}, N, \delta)$, as summarized in Table 2. A recurring challenge in these bounds is the need to quantify the "capacity" of the model's hypothesis class, which is particularly difficult for DNNs.

While there have been attempts to estimate the VC-dimension, such as those based on the norm of the parameters (Neyshabur et al., 2017), the resulting bounds heavily depend on the norm of the parameters. Consequently, it

remains unclear how to accurately estimate the sample complexity of models when varying the depth or width of DNNs. More recent work, such as Imaizumi and Schmidt-Hieber (2023), presents bounds that are tight but still dependent on the norm of the weights, assuming that the SGD iterates converge to a specific class of parameters.

Another line of research by Muthukumar and Sulam (2023) explores bounds that leverage the sparsity of feed-forward neural networks. However, there is still a lack of data-dependent bounds that do not rely on capacity estimates for models trained on random labels.

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{X,Y}\left[\ell(Y,f(X))\right]-\frac{1}{N}\sum_{i=1}^{N}\ell(Y_i,f(X_i))\right|\leq\epsilon(\mathcal{F},\delta,N)\right)\geq 1-\delta \tag{530}$$

Description	$\epsilon(\mathcal{F},N,\delta)$
Vapnik-Chernoviks Dimension, (Vapnik, 2000)	$\sqrt{\frac{\text{VCdim}(\mathcal{F})-\log(\delta)}{N}}$
Rademacher Complexity, (Bartlett and Mendelson, 2001)	$R_N(\mathcal{F})+\sqrt{\frac{-\log(\delta)}{N}}$
PAC-Bayes Bounds, (McAllester, 1999)	$\frac{KL(Q\|\|P)-\log(\delta)}{N}$
Gaussian Complexity, (Bartlett and Mendelson, 2001)	$G_N(\mathcal{F})+\sqrt{\frac{-\log(\delta)}{N}}$
Information-theoretic Bounds, (**?**)	$\frac{1}{N}\sum_{i=1}^{N}\sqrt{I(W;(X_i,Y_i))}$
Algorithmic Stability, (Feldman and Vondrak, 2019)	$\beta+\sqrt{\frac{-\log(\delta)}{N}}$

Table 2: SLT frameworks (in chronological order)

Matrix recovery: This is a fundamental problem in signal processing, where we seek to recover a matrix by indirect measurements, like random measurements, and random entry access. We typically have limited measurements; the problem itself is ill-posed when reconstructing the matrix. However, if the underlying matrix has certain special structures like low-rankedness, or sparsity in entries, the problem becomes tractable so as to reconstruct the true matrix. In practice, the problem tends to have low-rankedness, therefore having immense literature in this area, our work also presents such results, considering the optimization.

Let, $Y_i=\langle M^*,X_i\rangle+\epsilon\in\mathbb{R}$, where, $X_i\in\mathbb{R}^{m\times n}(m\geq n)$ is Gaussian entried matrix, $\epsilon\sim\mathcal{N}(0,\sigma^2)$ and $M^*\in\mathbb{R}^{m\times n}$ is a r^*-rank matrix. Consider the below problem

$$\begin{aligned}\min_{M\in\mathbb{R}^{m\times n}}\quad&\text{rank}(M)\\ \text{s.t.}\quad&\|Y_i-\langle M,X_i\rangle\|\leq\delta\end{aligned}\tag{531}$$

$$\begin{aligned}\min_{M\in\mathbb{R}^{m\times n}}\quad&\|M\|_*\\ \text{s.t.}\quad&\|Y_i-\langle M,X_i\rangle\|\leq\delta\end{aligned}\tag{532}$$

$$\begin{aligned}\min_{r\in\mathbb{N},U\in\mathbb{R}^{m\times r},V\in\mathbb{R}^{n\times r}}\quad&\|UV^T\|_*\\ \text{s.t.}\quad&\|Y_i-\langle UV^T,X_i\rangle\|\leq\delta\end{aligned}\tag{533}$$

$$\begin{aligned}\min_{r\in\mathbb{N},U\in\mathbb{R}^{m\times r},V\in\mathbb{R}^{n\times r}}\quad&\frac{1}{2}\left(\|U\|_F^2+\|V\|_F^2\right)\\ \text{s.t.}\quad&\|Y_i-\langle UV^T,X_i\rangle\|\leq\delta\end{aligned}\tag{534}$$

Table 3: Optimization problems for matrix sensing

$$\min_{r\in\mathbb{N},U\in\mathbb{R}^{m\times r},V\in\mathbb{R}^{n\times r}}\|Y_i-\langle UV^T,X_i\rangle\|^2+\frac{\lambda}{2}\left[\|U\|_F^2+\|V\|_F^2\right]\tag{535}$$

The optimization problem in (531) is non-convex due to its rank-minimization nature, an NP-HARD problem. However, under certain specific conditions on the measurement matrices X_i, the convex relaxation (532) can recover solutions to (531), as demonstrated in Recht et al. (2008). Solving the convex program (532) requires computing the Singular Value Decomposition (SVD), which has a computational complexity of $\mathcal{O}(mn^2)$.

To mitigate this computational burden, the Burer-Monteiro (BM) factorization (Burer and Monteiro, 2003) is employed, yielding the bilinear factorization in the non-convex program (533). This approach is more efficient than (532) because it introduces an implicit rank constraint, $rank(UV^T)\leq\min(n,r)$, which reduces the runtime of SVD to $\mathcal{O}((m+n)r^2)$. Additionally, the equivalence between the nuclear norm and the sum of Frobenius norms,

as shown by Giampouras et al. (2020), further accelerates the optimization process, reducing the complexity to $\mathcal{O}((m+n)r)$.

While the BM factorization program (533) is non-convex, in contrast to the convex program (532), gradient descent (GD) algorithms typically guarantee only local minima for non-convex optimization problems (Reddy and Vidyasagar, 2023). However, Ge et al. (2017) has proven that the program (533) has no spurious local minima, and any local minimum is indeed a global minimum. Numerous studies (Jia et al., 2023) have explored the optimization landscapes and the convergence to global minima.

Our work primarily focuses on the generalization capabilities of the BM factorization program (535), which represents the Lagrangian form of the program (533). Table F summarizes the results from the literature that provide matrix recovery guarantees; from this we can suggest there are no bounds in the literature for low-rank matrix recovery with nuclear norm regularization under noisy settings with generic parameterization. Our work presents results first of its kind.

Measurement Type	Scenario	Reference	Result
Exact	Under-Parameterized ($r < r^*$)	N/A	N/A
	Exactly-Parameterized ($r = r^*$)	N/A	Not directly available.
	Over-Parameterized ($r > r^*$)	(Stöger and Soltanolkotabi, 2021)	$\|UU^T - M^*\|_F \lesssim r^{*1/8}(r-r^*)^{3/8}$ when $r \in (r^*, 2r^*)$.
	Generic Parameterization ($r \geq 1$)	(Jin et al., 2023)	GD learns rank incrementally, $\|M^* - UU^T\|_F \lesssim \alpha^{\frac{1}{\widetilde{\sigma}_2 k^2}}$, but analysis is algorithmic.
	SDP Relaxation (*Full SDP Matrix*)	N/A	Not directly available.
Noisy	Under-Parameterized ($r < r^*$)	N/A	N/A
	Exactly-Parameterized ($r = r^*$)	(Ma et al., 2020)	$\|M^* - UU^T\|_F \lesssim \sqrt{\frac{\log(m)}{N}}$ under RIP assumptions $\delta_{4r^*} \leq 0.1$.
		(Negahban and Wainwright, 2011)	$\|\hat{M} - M\|_F \lesssim \sqrt{r^* \frac{m+n}{N}}$.
	Over-Parameterized ($r > r^*$)	(Ma et al., 2020)	$\|M^* - UU^T\|_F \lesssim \sqrt{\delta_{r+r^*}}\|M^*\|_2$.
	Generic Parameterization ($r \geq 1$)	N/A	N/A
	SDP Relaxation (*Full SDP Matrix*)	(Candès and Plan, 2011)	$\|\hat{M} - M^*\|_F \lesssim \sqrt{nr^*/N}$ under RIP assumptions.
		(Koltchinskii et al., 2011)	$\|\hat{M} - M\|_F \lesssim \frac{mnr^*\log(N)}{N}$ under uniform noisy measurements.

Table 4: Summary of Related Works on Matrix Recovery. N/A is an acronym for "Not Available".

Transformers: The remarkable success of Large Language Models (LLMs) (Team, 2024) can largely be attributed to their foundational architecture—Transformers (Vaswani et al., 2017). The optimization dynamics of Transformers have been a subject of extensive recent research (Bordelon et al., 2024), (Singh, 2023), (Yang et al., 2022), (Tian et al., 2023), (Nichani et al., 2024). Although Transformers exhibit impressive generalization capabilities in practical applications (Zhou et al., 2024), there is still a significant gap in the theoretical analysis of their generalization error.

To apply classical SLT bounds, one must determine the capacities of the function classes induced by Transformers. Previous attempts, such as in (Edelman et al., 2022), have made progress but were limited to scenarios where input data is bounded. In contrast, our work extends these results to settings where the inputs are not necessarily bounded.

Another line of research (Li et al., 2023), (Deora et al., 2024) has provided bounds that depend on step sizes and initialization choices for Gradient Descent (GD). For instance, Li et al. (2023) offered bounds within the context of in-context learning (Zhang et al., 2024), yet without evaluating the capacities of the stable algorithms used to train these Transformers.

In the broader literature, existing studies on generalization bounds often rely on strong assumptions, such as (i) bounded input data, (ii) algorithmic stability in some defined sense, and (iii) Lipschitz continuity of the loss function (which does not hold globally for mean squared error). Our results address these limitations by providing near-tight sample complexity bounds, offering a more comprehensive understanding of generalization in Transformer models.

G PRELIMINARIES

This section provides preliminaries of convex analysis and concentration of measure.

G.1 Convex Functions

Definition 1 (L^2 functions). *A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be square integrable on measure μ, i.e., $L^2(\mu)$ if and only if,*

$$\langle f, f \rangle_\mu = \int_{x \in \mathcal{X}} \langle f(x), f(x) \rangle_{\mathcal{Y}} d\mu(x) < \infty. \tag{536}$$

Definition 2 (Convex Set, (Rockafellar, 1970)). *A set $\mathcal{C}$ is said to be convex if and only if $\forall f, g \in \mathcal{C}$, $\alpha f + (1 - \alpha) g \in \mathcal{C}; \forall \alpha \in [0, 1]$.*

Definition 3 (Convex functions, (Rockafellar, 1970)). *A function, Ω is said to be convex if and only if $dom(\Omega)$ is convex and $\forall f, g \in dom(\Omega)$ and any $\alpha \in [0, 1]$.*

$$\Omega(\alpha f + (1 - \alpha) g) \leq \alpha \Omega(f) + (1 - \alpha) \Omega(g). \tag{537}$$

Definition 4 (Gauge function, (Rockafellar, 1970)). *The gauge function or the Minkowski functional is defined in a set $\mathcal{C} \in L^2(\mu)$ for a point f as follows,*

$$\sigma_{\mathcal{C}}(f) := \inf \left\{ t \geq 0; \text{ such that } f \in t conv(\mathcal{C}) \right\}. \tag{538}$$

Definition 5 (Polar Set, (Rockafellar, 1970)). *The polar set of any set $\mathcal{C} \subseteq L^2(\mu)$ is given be*

$$\mathcal{C}^\circ := \left\{ g \in L^2(\mu) : \text{ such that } \langle g, f \rangle_\mu \leq 1; \forall f \in \mathcal{C} \right\}. \tag{539}$$

Proposition 5 (Polar Properties).

Definition 6 (Polar function, (Rockafellar, 1970)). *The polar function of any gauge function, σ defined in the set $\mathcal{C} \subseteq L^2(\mu)$ is given be*

$$\sigma^\circ_{\mathcal{C}}(g) := \sigma_{\mathcal{C}^\circ}(g). \tag{540}$$

Definition 7 (Fenchal dual, (Rockafellar, 1970)). *The fenchal-dual for any μ-measurable function, Ω evaluated at $g \in L^2(\mu)$ is defined by,*

$$\Omega^*(g) := \sup_{f \in L^2(\mu)} \langle g, f \rangle_\mu - \Omega(f). \tag{541}$$

Lemma 12 (First Convexity, (Rockafellar, 1970)). *Any function Ω that is first-order differentiable, $\Omega \in \mathcal{C}^1$ is convex if and only if for any $f, g \in dom(\Omega)$*

$$\Omega(f) \geq \Omega(g) + \langle \nabla \Omega(g), f - g \rangle_\mu. \tag{542}$$

Lemma 13 (Strongly Convex, (Rockafellar, 1970)). *Any function Ω that is first-order differentiable, $\Omega \in \mathcal{C}^1$ is said to be $\lambda(\geq 0)$-strongly convex if and only if for any $f, g \in dom(\Omega)$*

$$\Omega(f) \geq \Omega(g) + \langle \nabla \Omega(g), f - g \rangle_\mu + \frac{\lambda}{2} \|f - g\|^2_\mu. \tag{543}$$

Definition 8 (Lipschitz Continuous). *A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be Lipschitz continuous with Lipschitz constant $\|f\|_{Lip}$ if for any $x_2, x_2 \in \mathcal{X}$*

$$\|f(x_1) - f(x_2)\|_{\mathcal{Y}} \leq \|f\|_{Lip} \|x_1 - x_2\|_{\mathcal{X}}. \tag{544}$$

Remark: Lipschitz constant, $\|f\|_{\text{Lip}}$ is not a norm but only a semi-norm. Because $\|f\|_{\text{Lip}} = 0$, it implies that f can be any constant function.

Definition 9 (Lipschitz Smooth). *A first-order differentiable function $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{C}^1$ is said to be lipchtiz smooth if ∇f is Lipschitz continous.*

Definition 10 ((L, λ) convex function). *A first-order differentiable function $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{C}^1$ is said to be (L, λ) convex if and only if f is L-Lipschitz smooth and λ-strongly convex, here $L \geq \lambda \geq 0$.*

Proposition 6 (Properties of Lipschitz). *The below are few properties of Lipschitz functions,*

1. *If function $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{C}^1$ then $\sup_{x \in \mathcal{X}} \frac{\|\langle \nabla f(x), x \rangle\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}} = \|f\|_{Lip}$.*
2. *If convex function $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{C}^1$ is L-Lipschitz smooth then,*

$$f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle_{\mathcal{Y}} \leq f(x) \leq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle_{\mathcal{Y}} + \frac{L}{2}\|x - x_0\|^2_{\mathcal{X}}. \tag{545}$$

3. *If convex function $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{C}^1$ is (L, λ) convex then,*

$$\begin{aligned} f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle_{\mathcal{Y}} + \frac{\lambda}{2}\|x - x_0\|^2_{\mathcal{X}} &\leq f(x) \\ &\leq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle_{\mathcal{Y}} + \frac{L}{2}\|x - x_0\|^2_{\mathcal{X}}. \end{aligned} \tag{546}$$

G.2 Concentration of Measure

Definition 11 (Greater than or approximately equal to). *The inequality $f \gtrsim g$ means that $\exists C > 0$ such that $f \geq Cg$.*

Definition 12 (Sub-Gaussianity). *A random variable, X is said to be sub-Gaussian with proxy variance, σ^2 if the following is satisfied,*

$$\mathbb{E}_X\left[e^t[X - \mathbb{E}[X]]\right] \leq \exp\left(-\frac{t^2\sigma^2}{2}\right); \forall t \geq 0. \tag{547}$$

We denote, $X \sim SG(\sigma^2)$.

Definition 13 (Sub-exponential). *A random variable X is said to be subexponential with the proxy parameter λ if the following is satisfied*

$$\mathbb{E}_X\left[e^{t[X - \mathbb{E}[X]]}\right] \leq \exp\left(-\frac{t\lambda}{2}\right); \forall t \geq 0. \tag{548}$$

We denote $X \sim SE(\lambda)$.

Proposition 7 (Properties of Sub-Gaussianity and Sub-exponential). *Let X, Y be two random variables that need not be independent.*

1. *$X \in \mathbb{R}^n \sim SG\left(\frac{\sigma_X^2}{n} I_{n \times n}\right)$ if and only if $\|X\|^2 \sim SE(\sigma_X^2)$.*
2. *If $X \in \mathbb{R}^n \sim SG\left(\frac{\sigma_X^2}{n} I_{n \times n}\right)$, then for any Lipschitz function $\phi : \mathcal{X} \to \mathbb{R}$, $\phi(X) \sim SG(\|\phi\|^2_{Lip}\sigma_X^2/n)$.*
3. *If $X \in \mathbb{R}^n \sim SG\left(\frac{\sigma_X^2}{n} I_{n \times n}\right)$, and $Y \in \mathbb{R}^n \sim SG\left(\frac{\sigma_Y^2}{n} I_{n \times n}\right)$, then $\langle X, Y \rangle \sim SE(\sigma_X \sigma_Y)$.*

Lemma 14 (Uniform concentration of function). *Consider an n_X-dimensional vector X, and a parameterized function, $g_\theta : \mathcal{X} \to \mathbb{R}$, where $\theta \in \mathcal{F}_\theta$. Let $\mathcal{C}$ be some convex set obeying $P(\cap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_{\mathcal{C}}$. Assume that for any fixed $\theta_1, \theta_1 \in \mathcal{F}_\theta$ and any $Z \in \mathcal{C}$ we have*

$$|g_{\theta_1}(Z) - g_{\theta_2}(Z)| \leq K d(\theta_1, \theta_2). \tag{549}$$

In addition, suppose that for any fixed $\theta \in \mathcal{F}_\theta$, we have

$$|\mathbb{E}\left[g_\theta(\mathcal{P}_{\mathcal{C}}(X)) - g_\theta(X)\right]| \leq B, \tag{550}$$

where $\mathcal{P}_{\mathcal{C}}(\cdot)$ denotes the Euclidean projection onto the set $\mathcal{C}$. Finally, suppose that for any fixed θ and $\epsilon \in [-t, t]$ it holds that

$$P\left(\left|\int_{\omega}(g_\theta \circ \mathcal{P}_{\mathcal{C}})d\mu_N(\omega) - \int_{\omega}(g_\theta \circ \mathcal{P}_{\mathcal{C}})d\mu(\omega)\right| \geq \epsilon\right) \leq \delta(\varepsilon), \tag{551}$$

Then for any $\epsilon \in [-t, t]$,

$$P\left(\sup_{\theta\in\mathcal{F}_\theta}\left|\int_{\omega} g_\theta d\mu_N(\omega) - \int_{\omega} g_\theta d\mu(\omega)\right| \geq \epsilon + B\right) \leq \mathcal{N}(\mathcal{F}_\theta, d(.,.), \epsilon/(2K))\delta(\epsilon/4) + \delta_{\mathcal{C}}. \tag{552}$$

Proof. The proof technique is similar to that of (Li and Wei, 2023, Lemma 6) but includes more general parameter sets $\mathcal{F}_\theta$. Let us define

$$h_\theta(X) := g_\theta(\mathcal{P}_{\mathcal{C}}(X)), \tag{553}$$

from the assumptions in the lemma, we have that,

$$P\left(\left|\int_{\omega} h_\theta d\mu_N(\omega) - \int_{\omega} h_\theta d\mu(\omega)\right| \geq \epsilon\right) \leq \delta(\varepsilon), \tag{554}$$

Next, we must establish uniform concentration overall $\theta \in \mathcal{E}_\theta$. Let us construct a ν-net for $\mathcal{F}_\theta$. For any $\theta' \in \mathcal{N}_\nu(\mathcal{F}_\theta, d(.,.))$, $\theta \in \mathcal{F}_\theta$ from the triangular inequality, and as

$$|h_\theta - h_{\theta'}| = |h_\theta(X) - h_{\theta'}(X)| = |g_\theta(\mathcal{P}_{\mathcal{C}}(X)) - g_{\theta'}(\mathcal{P}_{\mathcal{C}}(X))| \leq Kd(\theta, \theta'). \tag{555}$$

Then for any, X we have that that,

$$h_{\theta'} - Kd(\theta, \theta') \leq h_\theta \leq h_{\theta'} + Kd(\theta, \theta'). \tag{556}$$

Integrating with respect to the measure μ_N, we obtain

$$\int_{\omega} h_{\theta'} d\mu_N(\omega) - Kd(\theta, \theta') \leq \int_{\omega} h_\theta d\mu_N(\omega) \leq \int_{\omega} h_{\theta'} d\mu_N(\omega) + L_X d(\theta, \theta'). \tag{557}$$

Similarly for the measure μ we obtain

$$\implies \int_{\omega} h_{\theta'} d\mu(\omega) - Kd(\theta, \theta') \leq \int_{\omega} h_\theta d\mu(\omega) \leq \int_{\omega} h_{\theta'} d\mu(\omega) + Kd(\theta, \theta'). \tag{558}$$

Now, subtracting the above equations, we obtain

$$\begin{aligned}
&\int_{\omega} h_{\theta'} d\mu_N(\omega) - \int_{\omega} h_{\theta'} d\mu(\omega) - 2Kd(\theta, \theta') \\
&\leq \int_{\omega} h_\theta d\mu_N(\omega) - \int_{\omega} h_\theta d\mu(\omega) \leq \\
&\int_{\omega} h_{\theta'} d\mu_N(\omega) - \int_{\omega} h_{\theta'} d\mu(\omega) + 2Kd(\theta, \theta').
\end{aligned} \tag{559}$$

Now, take the absolute value on both sides. Later on, applying triangular inequality, we obtain

$$\left|\int_{\omega} h_\theta d\mu_N(\omega) - \int_{\omega} h_\theta d\mu(\omega)\right| \leq \left|\int_{\omega} h_{\theta'} d\mu_N(\omega) - \int_{\omega} h_{\theta'} d\mu(\omega)\right| + 2Kd(\theta, \theta'). \tag{560}$$

Now choose, θ^* as $arg\sup_{\theta\in\mathcal{F}_\theta}\left|\int_{\omega} h_{\theta'} d\mu_N(\omega) - \int_{\omega} h_{\theta'} d\mu(\omega)\right|$, then we have that,

$$\sup_{\theta\in\mathcal{F}_\theta}\left|\int_{\omega} h_{\theta'} d\mu_N(\omega) - \int_{\omega} h_{\theta'} d\mu(\omega)\right| \leq \left|\int_{\omega} h_{\theta'} d\mu_N(\omega) - \int_{\omega} h_{\theta'} d\mu(\omega)\right| + 2Kd(\theta^*, \theta'). \tag{561}$$

Now choose any θ' that lies at-most ν from θ^* on the metric, $d(.,.)$, i.e, $d(\theta', \theta^*) \leq \nu$, we have

$$\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \leq \left| \int_\omega h_{\theta'} d\mu_N(\omega) - \int_\omega h_{\theta'} d\mu(\omega) \right| + 2K\nu. \tag{562}$$

By definition, we can bound the right hand term by the supremum,

$$\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \leq 2K\nu + \sup_{\theta' \in \mathcal{N}_\nu(\mathcal{F}_\theta, d(.,.))} \left| \int_\omega h_{\theta'} d\mu_N(\omega) - \int_\omega h_{\theta'} d\mu(\omega) \right|. \tag{563}$$

We apply the probability measure on both side, obtaining,

$$P\left(\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon \right)$$

$$\leq P\left(\sup_{\theta' \in \mathcal{N}_\nu(\mathcal{F}_\theta, d(.,.))} \left| \int_\omega h_{\theta'} d\mu_N(\omega) - \int_\omega h_{\theta'} d\mu(\omega) \right| \geq \epsilon - 2K\nu \right), \tag{564}$$

the inequality is satisfied by the monotonicity of the probability measure. Now we apply the union-argument for the ν-net cover then we have

$$P\left(\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon \right) \leq$$

$$P\left(\bigcup_{\theta' \in \mathcal{N}_\nu(\mathcal{F}_\theta, d(.,.))} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon - 2K\nu \right), \tag{565}$$

Now we upper bound the right side union term with summation, and then we have

$$P\left(\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon \right) \leq$$

$$\sum_{\theta' \in \mathcal{N}_\nu(\mathcal{F}_\theta, d(.,.))} P\left(\left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon - 2K\nu \right), \tag{566}$$

Now we replace the summation with the ν-covering number, $\mathcal{N}(\mathcal{F}_\theta, d(.,.), \nu)$ obtaining

$$P\left(\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon \right) \leq \mathcal{N}(\mathcal{F}_\theta, d(.,.), \nu)\delta(\epsilon - 2K\nu). \tag{567}$$

Now set $\nu = \epsilon/(2K)$ then we have

$$\implies P\left(\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| \geq \epsilon \right) \leq \mathcal{N}(\mathcal{F}_\theta, d(.,.), \epsilon/(2K))\delta(\epsilon/2). \tag{568}$$

Now, we have established the uniform concentration for h_θ. Next, we move onto relating h_θ with the desired function h_θ.

Recall that

$$|\mathbb{E}\left[h_\theta(X) - g_\theta(X)\right]| \leq B. \tag{569}$$

As $P(\cap_{i=1}^N X_i \in \mathcal{C}) \geq 1 - \delta_\mathcal{C}$, we can safely claim $\int_\omega g_\theta d\mu_N(\omega) = \int_\omega h_\theta d\mu_N(\omega)$ with probability at least $1 - \delta_\mathcal{C}$. We have

$$\left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| = \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right|$$

$$\leq \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| + \left| \int_\omega h_\theta d\mu(\omega) - \int_\omega h_\theta d\mu(\omega) \right|, \tag{570}$$

$$\implies \left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| \leq \left| \int_\omega h_\theta d\mu_N(\omega) - \int_\omega h_\theta d\mu(\omega) \right| + B, \tag{571}$$

with probability at least $1 - \delta_{\mathcal{C}}$. Now we check the Lipschitzness of function h_θ in θ we have

$$|h_{\theta_1}(X) - h_{\theta_2}(X)| = |g_{\theta_1}(\mathcal{P}_{\mathcal{C}}(X)) - g_{\theta_2}(\mathcal{P}_{\mathcal{C}}(X))| \leq K d(\theta_1, \theta_2). \tag{572}$$

Similarly, in expectation measure, we have that

$$\begin{aligned} |\mathbb{E}[h_{\theta_1}(X)] - \mathbb{E}[h_{\theta_2}(X)]| = |\mathbb{E}[g_{\theta_1}(\mathcal{P}_{\mathcal{C}}(X))] - \mathbb{E}[g_{\theta_2}(\mathcal{P}_{\mathcal{C}}(X))]| &\leq \mathbb{E}\left[|g_{\theta_1}(\mathcal{P}_{\mathcal{C}}(X)) - g_{\theta_2}(\mathcal{P}_{\mathcal{C}}(X))|\right] \\ &\leq K d(\theta_1, \theta_2). \end{aligned} \tag{573}$$

Consequently for any $\theta' \in \{\theta' : d(\theta, \theta') \leq \epsilon/(2K)\}$, we have that

$$\left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| \leq \left| \int_\omega h_{\theta'} d\mu_N(\omega) - \int_\omega h_{\theta'} d\mu(\omega) \right| + \nu + B. \tag{574}$$

Now we choose $\theta = \theta^* = \sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right|$ then,

$$\begin{aligned} \sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| &= \left| \int_\omega g_{\theta^*} d\mu_N(\omega) - \int_\omega g_{\theta^*} d\mu(\omega) \right| \\ &\leq \left| \int_\omega h_{\theta'} d\mu_N(\omega) - \int_\omega h_{\theta'} d\mu(\omega) \right| + \epsilon + B. \end{aligned} \tag{575}$$

We can take a supremum over θ' in the upper bound of the right side term; we have

$$\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| \leq B + \epsilon + \sup_{\theta' \in \mathcal{F}_{\theta'}} \left| \int_\omega h_{\theta'} d\mu_N(\omega) - \int_\omega h_{\theta'} d\mu(\omega) \right|. \tag{576}$$

Then we use the inequality (568) and (571) we have that,

$$\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| \leq 2\epsilon + B, \tag{577}$$

with probability at least $1 - [\mathcal{N}(\mathcal{F}_\theta, d(.,.), \epsilon/(2K))\delta(\epsilon/2) + \delta_{\mathcal{C}}]$. Now rescaling we obtain that

$$P\left(\sup_{\theta \in \mathcal{F}_\theta} \left| \int_\omega g_\theta d\mu_N(\omega) - \int_\omega g_\theta d\mu(\omega) \right| \geq \epsilon + B \right) \leq \mathcal{N}(\mathcal{F}_\theta, d(.,.), \epsilon/(2K))\delta(\epsilon/4) + \delta_{\mathcal{C}}. \tag{578}$$

□