GENERALIZATION ANALYSIS OF MUON

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon is a matrix-aware first-order optimizer designed for matrix-structured parameters and has demonstrated strong performance in training large neural networks. Existing analyses mainly study its convergence behavior on training examples, whereas its behavior on testing examples remains less understood. In this paper, we study the convergence of Muon measured by the nuclear norm of the population gradient. We first establish a data-dependent uniform convergence guarantee for the deviation between population and empirical gradients over a ball in $\mathbb{R}^{d_1\times d_2}$ with rank at most r. Our bound depends on $r(d_1+d_2)$ instead of the ambient dimension d_1d_2 and incorporates the empirical second-order moment of the gradients. We then establish high-probability convergence rates for Muon. Our generalization and convergence analyses together give a high-probability excess risk bound of order $\widetilde{O}(\sqrt{r(d_1+d_2)/n})$. We further show that the dimension dependency can be removed for linear networks and shallow neural networks. Our studies give the first high-probability generalization guarantee for Muon.

1 INTRODUCTION

First-order methods remain the workhorse of large-scale neural network training. Modern architectures, however, store the majority of their trainable parameters as matrices: attention projections, fully connected layers, and embedding tables, so that the geometry of the parameter space is intrinsically two-dimensional. Standard vector-space optimizers such as stochastic gradient descent (SGD) treat these matrices as flattened vectors and therefore ignore the spectral structure that governs their updates. This observation has motivated a recent line of *matrix-aware* first-order methods, among which Muon (Jordan et al., 2024) is a particularly effective optimizer for large models.

Muon maintains a momentum matrix and replaces it, at every iteration, by its polar factor: the nearest semi-orthogonal matrix in Frobenius norm, equivalently obtained by setting every nonzero singular value of the momentum to one. The resulting update is a steepest-descent step under the operator-norm geometry (Chen et al., 2025; Pethick et al., 2025), which equalizes the contribution of singular directions and yields substantial spectral benefits (Ma et al., 2026; Lau et al., 2025). A natural consequence of this geometry is that stationarity should be measured by the nuclear norm of the gradient rather than by the Frobenius norm (Li & Hong, 2025; Shen et al., 2026; Chang et al., 2025; Kim & Oh, 2026; Hübler et al., 2026).

The empirical success of Muon motivates a lot of works on the optimization behavior, as measured by the gradient of empirical risks. For example, non-asymptotic convergence rates of order $O(T^{-\frac{1}{4}})$ have been developed for Muon on smooth problems in a stochastic setting (Li & Hong, 2025; Pethick et al., 2025; Shen et al., 2026; Chang et al., 2025), where T is the iteration number. Extensions to Newton–Schulz orthogonalization (Kim & Oh, 2026; Shulgin et al., 2025b; Do et al., 2026), momentum and variance reduction (Li & Hong, 2025; Chang et al., 2025; Sato et al., 2025), and heavy-tailed noise (Zhang & Lin, 2026; Yu et al., 2026; Hübler et al., 2026; Choudhury et al., 2026; Sfyraki & Wang, 2026) have also been studied in the literature. These guarantees describe how well Muon drives the *empirical* nuclear-norm gradient to zero, while do not explain the performance of the output model on test examples.

We often measure the testing performance of models by their population risks, which show their prediction behavior on testing examples. For nonconvex problems, a natural measure is the gradient of population risks (Mei et al., 2018; Foster et al., 2018; Lei, 2026). The difference between the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

population and empirical gradients is called a generalization gap, which quantifies how the training behavior of models generalizes to testing examples.

Uniform convergence of gradients has been a central tool to study the generalization gap for nonconvex stochastic methods (Mei et al., 2018; Foster et al., 2018; Lei, 2026; Li & Liu, 2022; Li et al., 2026; Arora et al., 2023). Existing bounds, however, are typically stated in the Frobenius (or Euclidean) norm $\|\cdot\|_F$ over a Euclidean ball in $\mathbb{R}^{d_1 d_2}$, and therefore scale with the ambient dimension $d_1 d_2$. Two structural facts about Muon make this scaling both loose and mismatched. First, the relevant geometry is the nuclear/operator-norm (i.e., $\|\cdot\|_*$, $\|\cdot\|_\sigma$) pair, which is strictly stronger than the Frobenius geometry since $\|\cdot\|_{\mathrm{F}} \leq \|\cdot\|_*$. Second, gradient-based training of matrix-parameterized models are shown to have a implicit bias toward low-rank solutions (Gunasekar et al., 2017; Arora et al., 2019; Huh et al., 2023; Jacot, 2023b), then the natural hypothesis class is not the full Euclidean ball but the rank-constrained operator-norm ball $\mathcal{B}_{R,r} := \{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{W}\|_\sigma \leq R, \operatorname{rank}(\mathbf{W}) \leq r\}$.

To our knowledge, the only existing generalization analysis of Muon (Huang et al., 2026) uses algorithmic stability to get a bound of order $O(1/(n\kappa^T))$ with κ being a small number, which grows exponentially with T. Meanwhile, all optimization guarantees for Muon hold *in expectation* (Li & Hong, 2025; Shen et al., 2026; Chang et al., 2025; Kim & Oh, 2026; Hübler et al., 2026), which fail to give the high-probability control of the optimization error. These gaps leave us with a basic question:

Does Muon enjoy a high-probability guarantee on the nuclear-norm stationarity of the population gradient which incorporates the low-rank structure of the parameter matrices?

This paper answers the question in the affirmative. Our contributions are as follows.

• **Data-dependent uniform convergence under the nuclear norm.** We establish a high-probability uniform convergence bound for the nuclear-norm gradient deviation over $\mathcal{B}_{R,r}$, which incorporates the empirical second-order moment of the gradients. The bound depends on the rank-dependent complexity $r(d_1+d_2)$ rather than on the ambient dimension $d_1 d_2$, incorporates the empirical second-moment information at the model of interest, and controls the stronger nuclear norm than the Frobenius analysis of Lei (2026).

• **High-probability optimization error.** We establish the first high-probability convergence rates for Muon, matching the best known in-expectation convergence rates (Li & Hong, 2025; Shen et al., 2026; Chang et al., 2025; Sato et al., 2025; Kim & Oh, 2026; Hübler et al., 2026).

• **High-probability excess risk.** Combining both generalization and optimization analyses, we give the first high-probability excess risk bound of Muon, which is of the order $\widetilde{O}(\sqrt{r(d_1+d_2)/n})$, where n is the sample size. Compared to the results of SGD (Lei, 2026), the ambient dimension $d_1 d_2$ is replaced by the rank-dependent complexity $r(d_1+d_2)$, which is advantageous when the iterates have low-rank.

• **Dimension-independent bounds.** For linear networks and shallow neural networks (SNNs), the explicit dependence on d_1 and d_2 can be removed. For linear networks, we derive an excess risk bound governed by the eigenvalue decay of the feature covariance and becomes $\widetilde{O}(n^{-1/2})$ under fast decay; for SNNs, we derive generalization bounds which adapt to the rank of $\mathbf{W}$ and to the empirical loss-derivative moment.

To our best knowledge, we provide the first high-probability generalization guarantees for Muon.

2 RELATED WORK

Convergence of Muon. Muon was introduced as a matrix-aware minimization method with Newton–Schulz orthogonalization (Jordan et al., 2024). Its idealized exact-polar direction was subsequently characterized under norm-steepest-descent and linear-minimization-oracle frameworks (Bernstein & Newhouse, 2024; Pethick et al., 2025), while Chen et al. (2025) established a spectral-norm-constrained interpretation for weight-decayed Muon. For smooth nonconvex objectives, non-asymptotic convergence rates were developed for both deterministic and stochastic muons (Li & Hong, 2025; Shen et al., 2026). Subsequent works studied variance reduction, batch-dependent rate improvements, and generalized smoothness or heavy-tailed noise (Chang et al., 2025;

Nagashima & Iiduka, 2026; Zhang & Lin, 2026; Yu et al., 2026), while structured-model analyses show spectral preconditioning benefits by orthogonalization (Lau et al., 2025; Ma et al., 2026).

To bridge the gap between exact-polar theory and practical Muon, Kim & Oh (2026) analyzed finite Newton–Schulz iterations, whereas Shulgin et al. (2025b) studied approximate polarization via an inexact-oracle model. Related extensions incorporate error feedback (Gruntkowska et al., 2025), Nesterov momentum and randomized polarization (Choudhury et al., 2026), and stochastic Frank–Wolfe or native heavy-tailed analyses (Sfyraki & Wang, 2026; Hübler et al., 2026). Recent results also identify fixed-batch non-convergence phenomena and a potential smoothing benefit of finite Newton–Schulz iterations (Do et al., 2026; Li & Tsuchiya, 2026). These results concern the optimization measured by empirical gradients, while our focus here is the generalization behavior.

Uniform Convergence of Gradients. The seminal work (Mei et al., 2018) pioneered the uniform convergence analysis of gradients, and gave dimension-dependent rates under some assumptions on Hessian statistical noises. The work (Zhang et al., 2017; Xu & Zeevi, 2025) applied some localization analyses to study the uniform convergence of excess gradients, which can imply fast rates for models close to the best model. The recent work (Lei, 2026) incorporated the second-moment information of the gradient at a single model into the uniform convergence of gradients. Dimension-independent uniform convergence of gradients was developed for generalized linear models (Foster et al., 2018). These convergence analyses of gradients were widely used to study the generalization behavior of stochastic optimization methods on nonconvex problems, including SGD (Lei, 2026), Clipped SGD (Li & Liu, 2022), SGD with momentum (Li et al., 2026), differentially private stochastic optimization (Arora et al., 2023) and stochastic gradient descent ascent (Zhang et al., 2024).

Generalization analysis of Muon. The recent work (Huang et al., 2026) used algorithmic stability (Bousquet & Elisseeff, 2002; Hardt et al., 2016) to develop a generalization bound of order $O(1/(n\kappa^T))$ for Muon with T iterations, where n is the sample size and κ is a small number. However, this generalization bound grows exponentially fast as a function of T.

3 PROBLEM FORMULATION

Notations. For a matrix A, let $\|A\|_{\mathrm{F}}, \|A\|_*$ and $\|A\|_\sigma$ denote the Frobenius norm, nuclear norm and the operator norm, respectively. We say $x \lesssim y$ if there exists some universal constant $c > 0$ such that $x \leq cy$. We use the notation $x \asymp y$ if $x \lesssim y$ and $y \lesssim x$. We hide logarithmic factors in $\widetilde{O}(\cdot)$. Define $a \vee b = \max\{a, b\}$.

Let ρ be a probability measure on a compact sample space $\mathcal{Z}$. Let $S = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ be a training dataset of n examples drawn independently from ρ. We aim to learn the parameter matrix $\mathbf{W}$ in the parameter space $\mathcal{W} \subset \mathbb{R}^{d_1 \times d_2}$ based on S. We measure the performance of $\mathbf{W}$ on an example $\mathbf{z}$ by the loss function $f(\mathbf{W}; \mathbf{z})$, where $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$. Then our objective is to find a model with a small population risk

$$\min_{\mathbf{W} \in \mathcal{W}} F(\mathbf{W}) := \mathbb{E}_{\mathbf{z}}[f(\mathbf{W}; \mathbf{z})], \tag{3.1}$$

where $\mathbb{E}_{\mathbf{z}}[\cdot]$ denotes expectation with respect to (w.r.t) $\mathbf{z} \sim \rho$. Since do not know ρ in practice, we can approximate the population risk F by an empirical risk

$$F_S(\mathbf{W}) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{W}; \mathbf{z}_i).$$

We often apply a stochastic learning algorithm $\mathcal{A}$ to the dataset S and build a model $\mathcal{A}(S) \in \mathcal{W}$ as an approximation of $\mathbf{W}^* = \arg\min_{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}} F(\mathbf{W})$. While $\mathcal{A}(S)$ often performs well on the training dataset, it may not generalize well to the test dataset due to the limited number of training examples. This asks for the investigation of gap between the population and risk to understand the generalization behavior of a learning algorithm.

Muon (Jordan et al., 2024) is a matrix-aware first-order optimizer designed for matrix-structured parameters. Upon receiving a stochastic gradient $\mathbf{G}_t = \nabla f(\mathbf{W}_t; \mathbf{z}_{t,i})$ at the t-th iteration, Muon first constructs a momentum matrix

$$\mathbf{M}_t = \beta \mathbf{M}_{t-1} + (1 - \beta)\mathbf{G}_t$$

to stabilize the noisy update direction. It then orthogonalizes $\mathbf{M}_t$ by replacing $\mathbf{M}_t$ with the polar factor $\mathbf{U}_t \mathbf{V}_t^\top$, where $\mathbf{M}_t = \mathbf{U}_t \mathbf{S}_t \mathbf{V}_t^\top$ denotes the singular value decomposition (SVD), i.e., replacing all nonzero singular values of $\mathbf{M}_t$ by one while preserving its singular subspaces. To avoid

the expensive exact SVD, the polar factor is approximated by a small number of Newton–Schulz iterations (Shulgin et al., 2025a; Kim & Oh, 2026; Do et al., 2026). Muon then updates models by

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \mathbf{U}_t \mathbf{V}_t^\top.$$

Geometrically, this procedure corresponds to a normalized steepest-descent step under the operator-norm geometry. As a result, Muon equalizes the contribution of singular directions in the update, exploits the matrix structure of the parameters, and naturally motivates measuring stationarity by $\|\nabla F(\mathbf{W})\|_*$ rather than by the Frobenius norm (Li & Hong, 2025; Shen et al., 2026; Chang et al., 2025; Kim & Oh, 2026; Hübler et al., 2026).

For Muon detailed in Algorithm 1, there is a lot of work on the convergence analysis in terms of the behaviour on the training examples. For example, when F_S is nonconvex and smooth, Li & Hong (2025); Shen et al. (2026); Chang et al. (2025); Kim & Oh (2026); Hübler et al. (2026) show that $\min_{t\in[T]} \mathbb{E}[\|\nabla F_S(\mathbf{W}_t)\|_*] = O(T^{-1/4})$. In this paper, we are interested in studying the convergence of Muon in terms of the behavior on testing examples, measured by $\|\nabla F(\mathcal{A}(S))\|_*$. We call it the *excess risk* and decompose it into two terms as follows:

$$||\nabla F(\mathcal{A}(S)||_* \leq ||\nabla F(\mathcal{A}(S)) - \nabla F_S(\mathcal{A}(S))||_* + ||\nabla F_S(\mathcal{A}(S))||_*,$$

where we refer to $||\nabla F(\mathcal{A}(S)) - \nabla F_S(\mathcal{A}(S))||_*$ as the *generalization gap*, and $||\nabla F_S(\mathcal{A}(S))||_*$ as the *optimization error*. The generalization gap shows how the empirical behavior of $\mathcal{A}(S)$ on S generalizes to a test set, while the optimization error considers the empirical behavior of the models. We consider the following assumptions widely used in the convergence analyses of Muon (Pethick et al., 2025; Kovalev, 2025; Shen et al., 2026; Kim & Oh, 2026).

Algorithm 1 Muon

Input: Initial weights $\mathbf{W}_0$, learning rate schedule $\{\eta_t\}$, $\beta \in [0, 1)$, batch size B
for $t = 0, 1, \ldots, T-1$ **do**
  Sample a batch $\{\mathbf{z}_{t,i}\}_{i=1}^B$ uniformly
  $\mathbf{G}_t = \frac{1}{B}\sum_{i=1}^B \nabla f(\mathbf{W}_t; \mathbf{z}_{t,i})$ (or $\mathbf{G}_t = \nabla f(\mathbf{W}_t)$ in the deterministic setting)
  if $t = 0$ **then**
    $\mathbf{M}_0 = \mathbf{G}_0$
  else
    $\mathbf{M}_t = \beta \mathbf{M}_{t-1} + (1-\beta)\mathbf{G}_t$
  end if
  $(\mathbf{U}_t, \mathbf{S}_t, \mathbf{V}_t) = \mathrm{SVD}(\mathbf{M}_t)$
  $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \mathbf{U}_t \mathbf{V}_t^\top$
end for

Assumption 1 (Bounded variance)**.** We assume $\nabla f(\mathbf{W}; \mathbf{z})$ is an unbiased stochastic estimator of the full gradient $\nabla F_S(\mathbf{W})$ and has a bounded variance, i.e.

$$\mathbb{E}_{\mathbf{z}}[\nabla f(\mathbf{W}; \mathbf{z})] = \nabla F_S(\mathbf{W}), \quad \mathbb{E}_{\mathbf{z}}[\|\nabla f(\mathbf{W}; \mathbf{z}) - \nabla F_S(\mathbf{W})\|_{\mathrm{F}}^2] \leq \sigma^2.$$

Assumption 2 (Spectral-norm Lipschitz smoothness)**.** Let the constant $L_* > 0$. We assume $\nabla f(\mathbf{W}; \mathbf{z})$ is L_*-smooth w.r.t $\mathbf{W}$, i.e.,

$$\|\nabla f(\mathbf{W}; \mathbf{z}) - \nabla f(\mathbf{W}'; \mathbf{z})\|_* \leq L_* \|\mathbf{W} - \mathbf{W}'\|_\sigma, \quad \forall \mathbf{W}, \mathbf{W}' \in \mathcal{W}.$$

4 Main Results

In this section, we present our main theoretical results. Section 4.1 establishes uniform convergence of the gradient deviation $\|\nabla F(\mathbf{W}) - \nabla F_S(\mathbf{W})\|_*$ under the nuclear norm, which controls the generalization gap. Section 4.2 then applies this result to the Muon output and combines it with a high-probability bound on the optimization error $\|\nabla F_S(\mathcal{A}(S))\|_*$, yielding a high-probability guarantee for the excess risk $\|\nabla F(\mathcal{A}(S))\|_*$. The dimension-independent bounds for linear and shallow networks are developed in Section 4.3.

4.1 Uniform Convergence of Gradients

To study the generalization gap of Muon, we need to establish the quantitative connection between population and empirical gradients at the output of the algorithm. However, since the output depends on the training samples, we turn to consider the uniform convergence of gradients over the whole function class. The uniform generalization gap is defined as the deviation between the population and empirical gradients over the whole function class.

Gradient-based training is known to exhibit an implicit bias toward low-rank solutions (Gunasekar et al., 2017; Arora et al., 2019; Dai et al., 2021; Jacot, 2023b; Wang & Jacot, 2024; Timor et al., 2022; Galanti et al., 2022), and such structure is also widely observed in trained networks (Huh et al., 2023; Jacot, 2023a; Papyan et al., 2020; Rangamani et al., 2023; Gur-Ari et al., 2018; Kobayashi et al., 2024; Balzano et al., 2025). Restricting attention to rank-sparse matrices therefore yields a more faithful hypothesis class and tighter complexity control for generalization (Pinto et al., 2025; Ledent et al., 2025a). Accordingly, we study the uniform generalization gap over the operator-norm ball of bounded rank $\mathcal{B}_{R,r} = \{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{W}\|_\sigma \le R, \operatorname{rank}(\mathbf{W}) \le r\}$, as stated in the following theorem (see Section A for the proof).

Theorem 1 (Uniform generalization gap)**.** *Define the empirical second-order moment* $\mathcal{L}_{*,S}(\mathbf{W}) := \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{W};\mathbf{z}_i)\|_*^2$*, and* $G_* = \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \|\nabla f(\mathbf{W};\mathbf{z})\|_*$ *for any* $\mathbf{z}$*. Let* $\Gamma_{n,\delta} := \log\frac{4}{\delta} + r(d_1+d_2)\log(3n) + \log\log(en)$*. Then, for every* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, the following inequality holds simultaneously for all* $\mathbf{W} \in \mathcal{B}_{R,r}$*:*

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* \lesssim \frac{G_*\Gamma_{n,\delta}}{n} + \sqrt{\frac{\mathcal{L}_{*,S}(\mathbf{W})\Gamma_{n,\delta}}{n}}.$$

Remark 1 (Comparison)**.** By using $\widetilde{O}(\cdot)$ to ignore all the logarithmic factors, Theorem 1 implies

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* = \widetilde{O}\bigg(\frac{G_* r(d_1+d_2)}{n} + \sqrt{\frac{\mathcal{L}_{*,S}(\mathbf{W}) r(d_1+d_2)}{n}}\bigg), \quad \forall \mathbf{W} \in \mathcal{B}_{R,r}.$$

Lei (2026) develops the uniform convergence w.r.t the Frobenius norm over the bounded ball $\mathcal{B}_R = \{\mathbf{W} \in \mathbb{R}^d : \|\mathbf{W}\|_{\mathrm{F}} \le R\}$ and shows that

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_{\mathrm{F}} = \widetilde{O}\bigg(\frac{G_F d_1 d_2}{n} + \sqrt{\frac{\mathcal{L}_{F,S}(\mathbf{W}) d_1 d_2}{n}}\bigg), \quad \forall \mathbf{W} \in \mathcal{B}_R,$$

where we define $\mathcal{L}_{F,S}(\mathbf{W}) := \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{W};\mathbf{z}_i)\|_{\mathrm{F}}^2$, $G_F = \sup_{\mathbf{W}\in\mathcal{B}_R} \|\nabla f(\mathbf{W};\mathbf{z})\|_{\mathrm{F}}$. In comparison, Theorem 1 improves their results by incorporating the rank information, and develops a tighter bound when there is a low-rank structure of $\mathbf{W}$ since $r(d_1+d_2) \ll d_1 d_2$ in this case. Moreover, we control the nuclear norm, which is stronger than Frobenius norm since $\|\cdot\|_{\mathrm{F}} \le \|\cdot\|_*$.

4.2 Generalization Analysis of Muon

We now apply Theorem 1 to the Muon iterates. The uniform bound depends on $G_* = \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \|\nabla f(\mathbf{W};\mathbf{z})\|_*$, which can be controlled by the operator-norm radius of the hypothesis class. We therefore first bound the Muon trajectory in operator norm in the following theorem (see proof in Section B.1).

Lemma 2 (Operator-norm bound of Muon)**.** *Let* $\{\mathbf{W}_t\}_{t=0}^T$ *be generated by Algorithm 1. Then, for every* $t \ge 1$ *we have* $\|\mathbf{W}_t\|_\sigma \le \|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{t-1}\eta_k$.

Lemma 2 shows that all Muon iterates lie in the operator-norm ball of radius $\|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{T-1}\eta_k$. Under Assumption 2 and the condition $b_* := \sup_{\mathbf{z}} \|\nabla f(\mathbf{0};\mathbf{z})\|_* < \infty$, this radius yields a concrete surrogate for G_*: for every $\mathbf{W} \in \mathcal{B}_{R,r}$,

$$\|\nabla f(\mathbf{W};\mathbf{z})\|_* \le \|\nabla f(\mathbf{W};\mathbf{z}) - \nabla f(\mathbf{0};\mathbf{z})\|_* + \|\nabla f(\mathbf{0};\mathbf{z})\|_* \le L_*\|\mathbf{W}\|_\sigma + b_* \le L_* R + b_*.$$

Hence one may replace G_* in Theorem 1 by $L_*R + b_*$. Specializing to a constant stepsize η (so that $R \le \|\mathbf{W}_0\|_\sigma + T\eta$) gives the following corollary on high-probability bounds of the generalization gap $\|\nabla F_S(\mathbf{W}_t) - \nabla F(\mathbf{W}_t)\|_*$.

Corollary 3 (High-probability generalization gap of Muon)**.** *Let $\Gamma_{n,\delta}$ be defined in Theorem 1 with $r = \max_t \operatorname{rank}(\mathbf{W}_t)$. Suppose Assumption 2 holds, $b_* = \sup_{\mathbf{z}} \|\nabla f(\mathbf{0}; \mathbf{z})\|_* < \infty$, and $R_T = \|\mathbf{W}_0\|_\sigma + T\eta$. Let $\{\mathbf{W}_t\}_{t=0}^T$ be generated by Algorithm 1 with stepsize η. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $t = 0, \dots, T-1$,*

$$\|\nabla F_S(\mathbf{W}_t) - \nabla F(\mathbf{W}_t)\|_* \lesssim \frac{(L_* R_T + b_*)\Gamma_{n,\delta}}{n} + \sqrt{\frac{\mathcal{L}_{*,S}(\mathbf{W}_t)\Gamma_{n,\delta}}{n}}.$$

Beyond the generalization gap, the excess risk of Muon also depends on the optimization error $\|\nabla F_S(\mathbf{W}_t)\|_*$. Existing analyses of Muon typically control this quantity in expectation (Li & Hong, 2025; Shen et al., 2026; Chang et al., 2025; Sato et al., 2025; Kim & Oh, 2026; Hübler et al., 2026). In contrast, we establish high-probability bounds (see proof in Section B.2).

Theorem 4 (High-probability optimization error of Muon)**.** *Suppose that Assumption 1 holds. Let $\{\mathbf{W}_t, \mathbf{M}_t\}_{t=0}^{T-1}$ be generated by Algorithm 1 with a constant stepsize $\eta_t = \eta$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* \le \frac{F_S(\mathbf{W}_0) - F_S^\star}{T\eta} + \frac{3\eta L_*}{1-\beta} + \frac{2\sqrt{q}G_F}{T(1-\beta)} + 4\sqrt{q}\log(2/\delta)\left[\frac{G_F(1-\beta)}{3} + \frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}\right], \quad (4.1)$$

where $q = \min\{d_1, d_2\}$, $F_S^* = \min_{\mathbf{W}} F_S(\mathbf{W})$, $G_F = \sup_{\mathbf{W}\in\mathcal{B}_R} \|\nabla f(\mathbf{W}; \mathbf{z})\|_{\mathrm{F}}$.

Specializing the stepsize η and the momentum parameter β, Theorem 4 yields the following high-probability convergence rates in the stochastic and deterministic regimes.

Corollary 5 (High-probability optimization error of Muon)**.** *Under the same assumptions as Theorem 4, let $\Delta_S = F_S(\mathbf{W}_0) - F_S^\star$, $\Gamma_\delta = (\frac{\sigma}{\sqrt{B}} + \frac{G_F}{3})\log(2/\delta)$, $q = \min\{d_1, d_2\}$.*

1. ***Stochastic setting.*** *Set $\eta = \sqrt{\frac{(1-\beta)\Delta_S}{TL_*}}$, $1-\beta = \min\{\frac{\sqrt{L_*\Delta_S}}{\Gamma_\delta\sqrt{qT}},\ 1\}$. Then, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* \lesssim \left(\frac{qL_*\Delta_S\Gamma_\delta^2}{T}\right)^{\frac{1}{4}} + \frac{q\Gamma_\delta^2}{\sqrt{L_*\Delta_S T}}. \quad (4.2)$$

2. ***Deterministic setting.*** *Set $\eta = \sqrt{\frac{\Delta_S}{TL_*}}$ and $\beta < 1$, then with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* \lesssim \sqrt{\frac{L_*\Delta_S}{T}}. \quad (4.3)$$

Consequently, the sample complexity to achieve an ϵ-nuclear-norm stationary point with probability at least $1-\delta$ is $O(qL_*\Delta_S\Gamma_\delta^2\epsilon^{-4})$ in the stochastic setting and $O\left(L_*\Delta_S\epsilon^{-2}\right)$ in the deterministic setting.

Remark 2 (Comparison with existing convergence analyses)**.** Unlike existing convergence analyses established in expectation, our results hold with high probability. High-probability analysis is more challenging than in-expectation analysis: Let $\mathbf{C}_t$ denote the full-gradient momentum counterpart of $\mathbf{M}_t$, defined by $\mathbf{C}_0 = \nabla F_S(\mathbf{W}_0)$ and $\mathbf{C}_t = \beta\mathbf{C}_{t-1} + (1-\beta)\nabla F_S(\mathbf{W}_t)$. In the in-expectation analyses, the martingale-difference property of $\mathbf{G}_t - \nabla F_S(\mathbf{W}_t)$ eliminates cross terms after taking expectations, then the accumulated momentum noise can be controlled through its second moment. In contrast, our high-probability analysis must control the deviation $\mathbf{M}_t - \mathbf{C}_t = \beta^t(\mathbf{G}_0 - \nabla F_S(\mathbf{W}_0)) + (1-\beta)\sum_{i=1}^t \beta^{t-i}(\mathbf{G}_i - \nabla F_S(\mathbf{W}_i))$, which requires a matrix-valued martingale concentration inequality to control $\mathbf{M}_t - \mathbf{C}_t$ simultaneously over the iterations, together with a bound on the error between $\mathbf{C}_t$ and $\nabla F_S(\mathbf{W}_t)$.
In the stochastic setting, we choose the same parameters η and β as in Li & Hong (2025); Shen et al.

(2026); Chang et al. (2025); Kim & Oh (2026); Hübler et al. (2026) and obtain the optimal complexity with respect to the orders of q and ϵ (Hübler et al., 2026) for finding an ϵ-nuclear-norm stationary point with high probability, with the noise parameter σ replaced by the high-probability quantity Γ_δ. In the deterministic setting, using the same choices of η and β as in Shen et al. (2026); Kim & Oh (2026), we obtain the optimal complexity with respect to the order of ϵ with high probability, matching the results of the existing analyses in expectation.

We now combine the generalization gap given in Corollary 3 and the optimization error given in Theorem 4 to get the following excess risk bound of Muon (see proof in Section B.3).

Theorem 6 (Excess risk of Muon)**.** *Let the definitions and assumptions in Corollary 3, Theorem 4 and Corollary 5 hold. Let $\{\mathbf{W}_t, \mathbf{M}_t\}_{t=0}^{T-1}$ be generated by Algorithm 1 with a constant stepsize $\eta_t = \eta$. By choosing $\eta = \sqrt{\frac{(1-\beta)\Delta_S}{TL_*}}$, $1-\beta = \min\{\frac{\sqrt{L_*\Delta_S}}{\Gamma_\delta\sqrt{qT}},\ 1\}$ as in Corollary 5, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* \lesssim \left(1+\sqrt{\frac{\Gamma_{n,\delta}}{n}}\right)\left\{\left(\frac{qL_*\Delta_S\Gamma_\delta^2}{T}\right)^{\frac{1}{4}} + \frac{\sqrt{q}\Gamma_\delta^2}{\sqrt{L_*\Delta_S T}}\right\} + \frac{\left(L_*\|\mathbf{W}_0\|_\sigma + \left(\frac{L_*^3\Delta_S^3 T}{\Gamma_\delta^2 q}\right)^{1/4} + b_*\right)\Gamma_{n,\delta}}{n} + \sigma\sqrt{\frac{\Gamma_{n,\delta}}{n}}.$$

Remark 3. If we choose $T \asymp \frac{\Gamma_\delta^2 q n^2}{\Gamma_{n,\delta}^2 L_*\Delta_S}$, Theorem 6 implies the excess risk bounds:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* = \widetilde{O}\left(\left(\sigma+\sqrt{L_*\Delta}\right)\sqrt{\frac{r(d_1+d_2)}{n}} + \frac{r(d_1+d_2)}{n}\left(L_*\|\mathbf{W}_0\|_\sigma + b_*\right)\right).$$

As a comparison, Lei (2026) show that, with $T \asymp n^2/((d_1d_2)^2L^2)$, SGD satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{W}_t)\|_{\mathrm{F}} = \widetilde{O}\left(\widetilde{L}\sqrt{\sigma^2+G_0^2}\sqrt{\frac{d_1d_2}{n}}\right),$$

where $\widetilde{L}$ is the smoothness constant w.r.t the Frobenius norm and G_0 is the constant such that $\sqrt{\eta}\|\nabla f(\mathbf{W}_t;\mathbf{z})\|_{\mathrm{F}} \le G_0$. In comparison, Theorem 6 shows that Muon replaces the parameter dimension d_1d_2 in the SGD rate by the rank-dependent complexity $r(d_1+d_2)$. This is advantageous when there is a low-rank structure of $\mathbf{W}$ since $r(d_1+d_2) \ll d_1d_2$ in this case. Moreover, Muon controls the nuclear norm of the population gradient, which is stronger than Frobenius-norm stationarity since $\|\cdot\|_{\mathrm{F}} \le \|\cdot\|_*$.

4.3 Dimension-independent Bounds

The uniform convergence rates in Section 4.1 and the excess risk bounds in Section 4.2 have explicit dependency on d_1, d_2, r, which are undesirable for overparameterized models. We now consider structured matrix models: linear networks and SNNs (proof in Section C).

Linear Networks. Let $a \in \mathbb{R}^{d_1}$ be a fixed vector. We consider loss functions of the form

$$f(\mathbf{W}; z) = \ell\big(y, a^\top \mathbf{W}\phi(x)\big), \tag{4.4}$$

where $\phi : \mathcal{X} \to \mathbb{R}^{d_2}$ is a feature map and $\ell : \mathbb{R}^2 \to \mathbb{R}_+$. Define the covariance operator $\Sigma_\phi = \mathbb{E}_X[\phi(X)\phi(X)^\top]$ and let $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ be its eigenvalues. For every $\mathbf{W} \in \mathcal{W}$, define $V(\mathbf{W}) = \mathbb{E}_X[(a^\top \mathbf{W}\phi(X))^2]$.

Building on dimension-independent uniform convergence bounds for vector-valued linear models and covariance-eigenvalue localization (Foster et al., 2018; Lei, 2026), we extend the analysis to the matrix-valued linear network (4.4) and control the uniform generalization gap in nuclear norm.

Theorem 7 (Dimension-independent uniform generalization gap for linear networks)**.** *Suppose that f takes the form in Eq. (4.4), the map $t \mapsto \ell(y,t)$ is L_ℓ-smooth and G_ℓ-Lipschitz for every y,*

and that $\sup_x \|\phi(x)\|_2 \le B_x$. *Define* $\nu(\mathbf{W}) = V(\mathbf{W}) \vee \frac{1}{n}$, *and* $\mathfrak{C}_\phi(R, v) = \inf_{h\in\mathbb{N}}\{1 + vh + R^2\|a\|_2^2 \sum_{j>h} \lambda_j\}$. *For every* $\delta \in (0, 1)$, *with probability at least* $1-\delta$, *uniformly for all* $\mathbf{W} \in \mathcal{B}_{R,r}$,

$$\|\nabla F(\mathbf{W}) - \nabla F_S(\mathbf{W})\|_* = \widetilde{\mathcal{O}}\Bigg([1 + \|a\|_2 B_x(1 + L_\ell)]\sqrt{\frac{\mathfrak{C}_\phi(R, \nu(\mathbf{W}))}{n}} + \frac{(L_\ell R\|a\|_2^2 B_x^2 + G_\ell\|a\|_2 B_x)\log(1/\delta)}{n}\Bigg). \quad (4.5)$$

We now apply Theorem 7 to the Muon iterates, which gives the dimension-independent high-probability bound on the generalization gap of Muon for linear networks (4.4).

Corollary 8 (Dimension-independent high-probability generalization gap of Muon for linear networks)**.** *Suppose that* f *takes the form in Eq. (4.4) and that the assumptions of Theorem 7 hold. Let* $\{\mathbf{W}_t\}_{t=0}^T$ *be generated by Algorithm 1 with constant stepsize* η, *and* $R_T = \|\mathbf{W}_0\|_\sigma + T\eta$. *Then, for every* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$, *simultaneously for all* $t = 0, \dots, T - 1$,

$$\|\nabla F(\mathbf{W}_t) - \nabla F_S(\mathbf{W}_t)\|_* = \widetilde{\mathcal{O}}\Bigg([1 + \|a\|_2 B_x(1 + L_\ell)]\sqrt{\frac{\mathfrak{C}_\phi(R_T, \nu(\mathbf{W}_t))}{n}} + \frac{(L_\ell R_T\|a\|_2^2 B_x^2 + G_\ell\|a\|_2 B_x)\log(1/\delta)}{n}\Bigg). \quad (4.6)$$

We now specialize the optimization errors established in Theorem 4 and Corollary 5 to the linear networks (4.4), which gives the dimension-independent high-probability optimization error.

Theorem 9 (Dimension-independent high-probability optimization error of Muon for linear networks)**.** *Suppose* f *takes the form in Eq. (4.4). Theorem 4 and Corollary 5 hold with* $L_* = L_\ell\|a\|_2^2 B_x^2$, $G_F = 2G_\ell\|a\|_2 B_x$, $\sigma = G_\ell\|a\|_2 B_x$, $\Gamma_\delta = G_\ell\|a\|_2 B_x\big(\frac{1}{\sqrt{B}} + \frac{2}{3}\big)\log(2/\delta)$, *and* $q = 1$.

Remark 4. Theorem 9 shows that, for linear networks, Muon achieves a dimension-independent optimization error. Compared with Corollary 5, the factor $q = \min\{d_1, d_2\}$ is removed from both the convergence rate and the iteration complexity.

We now combine the generalization gap given in Corollary 8 and the optimization error given in Theorem 9 to get the following dimension-independent excess risk bound of Muon.

Theorem 10 (Dimension-independent excess risk of Muon for linear networks)**.** *Suppose that* f *takes the form in Eq. (4.4) and that the assumptions and definitions of Corollary 8 and Theorem 9 hold. Let* Δ_S *be defined in Corollary 5. Define* $\nu_T = \frac{1}{T}\sum_{t=0}^{T-1} \nu(\mathbf{W}_t)$ *and*

$$\mathfrak{G}_n(R, v, \delta) = \|a\|_2 B_x L_\ell\left[\frac{1}{n}\inf_{h\in\mathbb{N}}\Big(1 + vh + R^2\|a\|_2^2\sum_{j>h}\lambda_j\Big)\right]^{1/2} + \frac{(L_\ell R\|a\|_2^2 B_x^2 + G_\ell\|a\|_2 B_x)\log(1/\delta)}{n}.$$

By choosing $\eta = \sqrt{\frac{(1-\beta)\Delta_S}{TL_{\mathrm{lin}}}}$, $1 - \beta = \min\{\frac{\sqrt{L_{\mathrm{lin}}\Delta_S}}{\Gamma_\delta\sqrt{T}}, 1\}$, *with probability at least* $1 - \delta$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* \le \widetilde{\mathcal{O}}\Big(\Big(\frac{L_{\mathrm{lin}}\Delta_S\Gamma_\delta^2}{T}\Big)^{1/4} + \frac{\Gamma_\delta^2}{\sqrt{L_{\mathrm{lin}}\Delta_S T}} + \mathfrak{G}_n(R_T, \nu_T, \delta)\Big). \quad (4.7)$$

Remark 5. Assume that $\nu_T = \widetilde{O}(1)$, and that the eigenvalues of the covariance operator $\Sigma_\phi = \mathbb{E}_X[\phi(X)\phi(X)^\top]$ satisfy $\lambda_j \le cj^{-p}$ for some $c > 0$ and $p > 1$, which are used in Lei (2026). If we ignore all factors independent of n and T, Theorem 10 implies

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* = \widetilde{\mathcal{O}}\Bigg(T^{-1/4} + \sqrt{\frac{1}{n}\inf_{h\in\mathbb{N}}\Big\{h + R_T^2\sum_{j>h}\lambda_j\Big\}} + \frac{T^{1/4}}{n}\Bigg).$$

Then we can choose $T \asymp n^{\frac{2p}{p+1}}$ to get the excess risk bound of the order $\widetilde{O}(n^{-\frac{p}{2(p+1)}})$, which becomes $\widetilde{O}(n^{-1/2})$ if $p \to \infty$. As a comparison, Theorem 6 has an explicit dependency on d_1, d_2, and r, while the bound in Theorem 10 is independent of d_1, d_2, and r.

SNNs. Let $a \in \mathbb{R}^{d_1}$ be a fixed vector. We consider loss functions of the form

$$f(\mathbf{W}; z) = \ell\big(y, a^\top \sigma(\mathbf{W}\phi(x))\big), \tag{4.8}$$

where $\phi : \mathcal{X} \to \mathbb{R}^{d_2}$ is a feature map and $\ell : \mathbb{R}^2 \to \mathbb{R}_+$. Define $\mathscr{L}_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell'\big(y_i, a^\top \sigma(\mathbf{W}\phi(x_i))\big)^2$. We develop dimension-independent uniform convergence bounds for model (4.8) under Assumption 3, which is satisfied by a broad class of loss functions, including the logistic, exponential, and linear losses (Taheri & Thrampoulidis, 2024).

Assumption 3. Assume there exists $\gamma > 0$ such that $|\ell''(y,t)| \le \gamma|\ell'(y,t)|$ for all y, t, and that $|\ell'(y,t)| \le \max\{|\ell'(y,t_1)|, |\ell'(y,t_2)|\}$ whenever t lies between t_1 and t_2.

Theorem 11 (Dimension-independent uniform generalization gap for SNNs)**.** *Suppose that f takes the form in Eq. (4.8). Assume that $\sigma(0) = 0$, $\|\sigma'(\cdot)\|_\infty \le L_\sigma$, and $\sigma'(\cdot)$ is $L_{\sigma'}$-Lipschitz. Assume further that, for every y, the map $t \mapsto \ell(y,t)$ is L_ℓ-smooth and G_ℓ-Lipschitz, and that $\sup_x \|\phi(x)\|_2 \le B_x$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, the following inequality holds uniformly for every $\mathbf{W} \in \mathcal{B}_{R,r}$:*

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* = \widetilde{\mathcal{O}}\Bigg(\left[\frac{1 + L_\sigma\|a\|_2 B_x + R\sqrt{r}\,\|a\|_2 B_x^2(L_{\sigma'} + \gamma L_\sigma^2\|a\|_2)}{\sqrt{n}}\right] \times \left[\sqrt{\mathscr{L}_S(\mathbf{W})} + \frac{G_\ell(1 + \beta L_\sigma R\sqrt{r}\,\|a\|_2 B_x)\log(2/\delta)}{\sqrt{n}}\right]\Bigg). \tag{4.9}$$

Remark 6. There is a lot of work on the dimension-independent uniform convergence of function values for SNNs (Golowich et al., 2020; Vardi et al., 2022; Liu et al., 2024; Lei & Xie, 2026), some of which incorporates the rank information (Ledent et al., 2025b). However, there is far less work on the uniform convergence of gradients for SNNs: while local uniform convergence of gradients was studied for one-hidden-layer ReLU networks (Zhang et al., 2019) and sigmoid networks (Fu et al., 2020), their bounds hold in a neighborhood of a ground-truth network and depend explicitly on the input dimension and network width. As a comparison, Theorem 11 holds uniformly over $\mathcal{B}_{R,r}$, has no explicit dependence on d_1 and d_2, and incorporates the empirical loss-derivative moment $\mathscr{L}_S(\mathbf{W})$. It is tighter when $\mathbf{W}$ has a low-rank structure and the empirical risk is small, since the self-bounding property $|\ell'(y,t)|^2 \le 4L_\ell \ell(y,t)$ (Srebro et al., 2010) implies $\mathscr{L}_S(\mathbf{W}) \le 4L_\ell F_S(\mathbf{W})$.

We now apply Theorem 11 to the Muon iterates, which gives the dimension-independent high-probability bound on the generalization gap for SNNs (4.8).

Corollary 12 (Dimension-independent high-probability generalization gap of Muon for SNNs)**.** *Suppose f takes the form in Eq. (4.8). Let assumptions and definitions in Theorem 11 hold, $\{\mathbf{W}_t\}_{t=0}^T$ be generated by Algorithm 1 with stepsize η, $r = \max_t \operatorname{rank}(\mathbf{W}_t)$, and $R_T = \|\mathbf{W}_0\|_\sigma + T\eta$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $t = 0, \dots, T-1$,*

$$\|\nabla F_S(\mathbf{W}_t) - \nabla F(\mathbf{W}_t)\|_* = \widetilde{\mathcal{O}}\Bigg(\left[\frac{1 + L_\sigma\|a\|_2 B_x + R_T\sqrt{r}\,\|a\|_2 B_x^2(L_{\sigma'} + \beta L_\sigma^2\|a\|_2)}{\sqrt{n}}\right] \times \left[\sqrt{\mathscr{L}_S(\mathbf{W}_t)} + \frac{G_\ell(1 + \gamma L_\sigma R_T\sqrt{r}\,\|a\|_2 B_x)\log(2/\delta)}{\sqrt{n}}\right]\Bigg). \tag{4.10}$$

5 Conclusion and Future Work

In this paper, we establish the first high-probability generalization guarantee for Muon under population nuclear-norm stationarity. Our data-dependent uniform bound replaces the ambient dimension $d_1 d_2$ with the rank-aware complexity $r(d_1 + d_2)$ and, together with a high-probability optimization analysis, yields an excess-risk rate of $\widetilde{O}(\sqrt{r(d_1+d_2)/n})$. We further obtain a dimension-independent excess-risk bound for linear networks, approaching $\widetilde{O}(n^{-1/2})$ under fast eigenvalue decay, and a dimension-independent generalization bound for SNNs.

Several directions remain open. First, extending the analysis to the inexact polar factors produced by finite Newton–Schulz iterations would better reflect practical implementations of Muon (Shulgin et al., 2025b; Kim & Oh, 2026). Second, it would be valuable to develop dimension-independent guarantees for deep networks and blockwise Muon updates (Jordan et al., 2024; Pethick et al., 2025). Finally, replacing worst-case rank and trajectory-radius bounds with localized quantities, such as effective rank and singular-value decay, may yield sharper guarantees (Ledent et al., 2025a).

REFERENCES

Raman Arora, Raef Bassily, Tomás González, Cristóbal A Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. In *International Conference on Machine Learning*, pp. 1060–1092. PMLR, 2023.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7413–7424, 2019.

Anonymous Authors. Towards optimistic convergence of gradients for nonconvex learning. *preprint*, 2026.

Laura Balzano, Tianjiao Ding, Benjamin D. Haeffele, Soo Min Kwon, Qing Qu, Peng Wang, Zhangyang Wang, and Can Yaras. An overview of low-rank structures in the training and adaptation of large models, 2025.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026.

Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In *Advances in Neural Information Processing Systems*, 2021.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On muon optimization: From non-convergence to an error analysis with polar express and the newton-schulz polynomial from implementations. *arXiv preprint arXiv:2608.04607*, 2026.

Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for nonconvex learning and optimization. In *Advances in Neural Information Processing Systems*, pp. 8759–8770, 2018.

Haoyu Fu, Yuejie Chi, and Yingbin Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 68:3225–3235, 2020. doi: 10.1109/TSP.2020.2993153.

Tomer Galanti, Zachary S. Siegel, Aparna Gupte, and Tomaso Poggio. SGD and weight decay secretly minimize the rank of your neural network. 2022.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference on learning theory*, pp. 297–299. PMLR, 2018.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020. doi: 10.1093/imaiai/iaz007.

Kaja Gruntkowska, Alexander Gaponov, Zhirayr Tovmasyan, and Peter Richtárik. Error feedback for Muon and friends. *arXiv preprint arXiv:2510.00643*, 2025.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

Feihu Huang, Yuning Luo, and Songcan Chen. Mimuon: Mixed muon optimizer with improved generalization for large models. *arXiv preprint arXiv:2605.19619*, 2026.

Florian Hübler, Thomas Pethick, and Suvrit Sra. Free heavy-tailed lunch for muon: A theoretical justification of empirical success. *arXiv preprint arXiv:2606.14560*, 2026.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023.

Arthur Jacot. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff. In *Advances in Neural Information Processing Systems*, 2023a.

Arthur Jacot. Implicit bias of large depth networks: A notion of rank for nonlinear functions. In *International Conference on Learning Representations*, 2023b.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. In *International Conference on Learning Representations*, volume 2026, pp. 45672–45718, 2026.

Seijin Kobayashi, Yassir Akram, and Johannes von Oswald. Weight decay induces low-rank attention layers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4481–4510, 2024.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Antoine Ledent, Rodrigo Alves, and Yunwen Lei. Generalization bounds for rank-sparse neural networks. In *Advances in Neural Information Processing Systems*, 2025a.

Antoine Ledent, Rodrigo Alves, and Yunwen Lei. Generalization bounds for rank-sparse neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL `https://proceedings.neurips.cc/paper_files/paper/2025/hash/d95564314a9cd449603ed4b9d7dda02c-Abstract-Conference.html`.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Yunwen Lei. Towards better generalization bounds of stochastic optimization for nonconvex learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2104–2117, 2026.

Yunwen Lei and Yufeng Xie. Towards initialization-dependent and non-vacuous generalization bounds for overparameterized shallow neural networks, 2026. URL `https://arxiv.org/abs/2604.00505`.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025.

Mingyi Li and Taira Tsuchiya. Muon with finite Newton–Schulz: The smoothing benefit in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2608.26288*, 2026.

Shaojie Li and Yong Liu. High probability guarantees for nonconvex stochastic gradient descent with heavy tails. In *International Conference on Machine Learning*, pp. 12931–12963. PMLR, 2022.

Shaojie Li, Pengwei Tang, Bowei Zhu, and Yong Liu. High probability bounds for non-convex stochastic optimization with momentum. In *International Conference on Learning Representations*, volume 2026, pp. 25489–25524, 2026.

Fanghui Liu, Leello Dadi, and Volkan Cevher. Learning with norm constrained, over-parameterized, two-layer neural networks. *Journal of Machine Learning Research*, 25(138):1–42, 2024. URL https://www.jmlr.org/papers/v25/22-1250.html.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Shuntaro Nagashima and Hideaki Iiduka. Improved convergence rates of Muon optimizer for nonconvex optimization. *arXiv preprint arXiv:2601.19400*, 2026.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Andrea Pinto, Akshay Rangamani, and Tomaso A. Poggio. On generalization bounds for neural networks with low rank layers. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 921–936. PMLR, 2025.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A. Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic Frank–Wolfe under heavy-tailed noise. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2506.04192.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=4nH4CulGaP.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact Muon update. *arXiv preprint arXiv:2510.19933*, 2025a.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact Muon update. *arXiv preprint arXiv:2510.19933*, 2025b.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in ReLU networks. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 1429–1459. PMLR, 2022.

Gal Vardi, Ohad Shamir, and Nati Srebro. The sample complexity of one-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:9139–9150, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Zihan Wang and Arthur Jacot. Implicit bias of SGD in L_2-regularized linear DNNs: One-way jumps from high to low rank. In *International Conference on Learning Representations*, 2024.

Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 50(1):40–67, 2025.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. *arXiv preprint arXiv:2602.07425*, 2026.

Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: $O(1/n)$-and $O(1/n^2)$-type of risk bounds. In *Conference on Learning Theory*, pp. 1954–1979, 2017.

Siqi Zhang, Yifan Hu, Liang Zhang, and Niao He. Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 694–702. PMLR, 2024.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer ReLU networks via gradient descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1524–1534. PMLR, 2019.

Yubo Zhang and Junhong Lin. On convergence of muon for nonconvex stochastic optimization under generalized smoothness. 2026.

A Proof of Uniform Convergence of Gradients

In this section, we present the proofs of uniform convergence of gradients. We give some definitions and useful lemmas in Section A.1, define and establish the connection between empirical second-order moment and the population second-order moment in Section A.2, and give the proof of the main results (Theorem 1) in Section A.3.

A.1 Definitions and Useful Lemmas

The uniform convergence depends on the complexity of function classes. We now give the definitions of covering number and Rademacher complexity, which are two complexity measures of function classes.

Definition 1 (Covering number)**.** Let $(\mathcal{G}, d)$ be a metric space and set $\mathcal{F} \subseteq \mathcal{G}$. For any $\epsilon > 0$, a set $\mathcal{F}^{\triangle} \subset \mathcal{F}$ is called an ϵ-cover of $\mathcal{F}$ if for every $f \in \mathcal{F}$ we can find an element $g \in \mathcal{F}^{\triangle}$ satisfying $d(f, g) \leq \epsilon$. The covering number $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the cardinality of the minimal ϵ-cover of $\mathcal{F}$:

$$\mathcal{N}(\epsilon, \mathcal{F}, d) := \min \left\{ \left| \mathcal{F}^{\triangle} \right| : \mathcal{F}^{\triangle} \text{ is an } \epsilon\text{-cover of } \mathcal{F} \right\},$$

where $|\cdot|$ denotes the cardinality of a set.

Definition 2 (Rademacher complexity)**.** Let $\mathcal{F} = \{f : \mathcal{Z} \mapsto \mathbb{R}\}$ be a function class. Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher variables with $\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$. We define the empirical Rademacher complexity as

$$\Re_{\tilde{S}}(\widetilde{\mathcal{F}}) = \mathbb{E}_{\epsilon}\Big[\sup_{f \in \tilde{\mathcal{F}}} \frac{1}{n} \sum_{i \in [n]} \epsilon_i f(\mathbf{z}_i) \Big].$$

The following lemma provides a connection between these two complexity measures.

Lemma 13 (Dudley's entropy integral Srebro et al. (2010))**.** *Let $\mathcal{F} = \{f : \mathcal{Z} \mapsto \mathbb{R}\}$ be a function class with $\sup_{f \in \mathcal{F}} d_S(f, 0) \leq D$ and $S = \{z_1, \ldots, z_n\}$, where d_S is a pseudometric on $\mathcal{F}$ defined as follows*

$$d_S(f, g) := \Big(\frac{1}{n} \sum_{i=1}^{n} (f(z_i) - g(z_i))^2 \Big)^{\frac{1}{2}}.$$

Then, there holds

$$\Re_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + 12 \int_{\alpha}^{D} \sqrt{\frac{\log \mathcal{N}(r, \mathcal{F}, d_S)}{n}} \, \mathrm{d}r \right\}.$$

We use the following classical result to estimate the covering numbers of the ball with bounded radius and rank.

Lemma 14 (Wainwright (2019))**.** *Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$, and write $B(R) := \{x \in \mathbb{R}^d : \|x\| \leq R\}$. Then for every $\delta > 0$, there holds*

$$\mathcal{N}(\delta; B(R), \|\cdot\|) \leq \big(1 + \frac{2R}{\delta}\big)^d.$$

Using Lemma 14, the following lemma gives the covering numbers of the ball with bounded radius and rank.

Lemma 15. *Recall $\mathcal{B}_{R,r} = \{\mathbf{W} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{W}\|_{\sigma} \leq R, \mathit{rank}(\mathbf{W}) \leq r\}$. Define the metric $d_{op}(\mathbf{W}, \mathbf{W}') := \|\mathbf{W} - \mathbf{W}'\|_{\sigma}$ over ball $\mathcal{B}_{R,r}$. Then $\log \mathcal{N}(\epsilon, \mathcal{B}_{R,r}, d_{op}) \leq Cr(d_1 + d_2) \log(3R/\epsilon)$, where C is a constant.*

Proof of Lemma 15. For any matrix $\mathbf{W} \in \mathcal{B}_{R,r}$, since $\mathrm{rank}(\mathbf{W}) \leq r$, we can consider its Singular Value Decomposition (SVD):

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top},$$

where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ have orthonormal columns, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing the singular values. The condition $\|\mathbf{W}\|_{\sigma} \leq R$ implies $\|\mathbf{\Sigma}\|_{\sigma} \leq R$.
We can factorize $\mathbf{W}$ into $\mathbf{W} = \mathbf{L}\mathbf{R}^{\top}$, where $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}^{1/2} \in \mathbb{R}^{m \times r}$ and $\mathbf{R} = \mathbf{V}\mathbf{\Sigma}^{1/2} \in \mathbb{R}^{n \times r}$. Notice that:

$$\|\mathbf{L}\|_{\sigma}^2 = \|\mathbf{L}^{\top}\mathbf{L}\|_{\sigma} = \|\mathbf{\Sigma}^{1/2}\mathbf{U}^{\top}\mathbf{U}\mathbf{\Sigma}^{1/2}\|_{\sigma} = \|\mathbf{\Sigma}\|_{\sigma} \leq R.$$

Thus, $\|\mathbf{L}\|_{\sigma} \leq \sqrt{R}$, and similarly $\|\mathbf{R}\|_{\sigma} \leq \sqrt{R}$. This means $\mathbf{L}$ and $\mathbf{R}$ belong to the bounded sets:

$$S_L = \{\mathbf{L} \in \mathbb{R}^{d_1 \times r} : \|\mathbf{L}\|_{\sigma} \leq \sqrt{R}\}, \quad S_R = \{\mathbf{R} \in \mathbb{R}^{d_2 \times r} : \|\mathbf{R}\|_{\sigma} \leq \sqrt{R}\}.$$

By Lemma 14, identifying $\mathbb{R}^{d_1 \times r}$ with the $d_1 r$-dimensional vector space $\mathbb{R}^{d_1 r}$ (on which $\|\cdot\|_{\sigma}$ remains a norm), the set S_L is a ball of radius $\sqrt{R}$. Thus, for any $\delta > 0$, there exists a δ-net $\mathcal{N}_L$ covering S_L in the operator norm such that $|\mathcal{N}_L| \leq \left(1 + \frac{2\sqrt{R}}{\delta}\right)^{d_1 r}$. Similarly, identifying $\mathbb{R}^{d_2 \times r}$

with $\mathbb{R}^{d_2 r}$, there exists a δ-net $\mathcal{N}_R$ for S_R with $|\mathcal{N}_R| \leq \left(1 + \frac{2\sqrt{R}}{\delta}\right)^{d_2 r}$. We construct a candidate net for $\mathcal{B}_{R,r}$ by taking all pairwise products

$$\mathcal{N}_W = \{\hat{\mathbf{L}}\hat{\mathbf{R}}^\top : \hat{\mathbf{L}} \in \mathcal{N}_L, \hat{\mathbf{R}} \in \mathcal{N}_R\}$$

For any $\mathbf{W} = \mathbf{L}\mathbf{R}^\top \in \mathcal{B}_{R,r}$, we can find $\hat{\mathbf{L}} \in \mathcal{N}_L$ and $\hat{\mathbf{R}} \in \mathcal{N}_R$ such that $\|\mathbf{L} - \hat{\mathbf{L}}\|_\sigma \leq \delta$ and $\|\mathbf{R} - \hat{\mathbf{R}}\|_\sigma \leq \delta$. The operator norm distance to the composite matrix is bounded by:

$$\begin{aligned}
&\left\|\mathbf{W} - \hat{\mathbf{L}}\hat{\mathbf{R}}^\top\right\|_\sigma = \left\|\mathbf{L}\mathbf{R}^\top - \hat{\mathbf{L}}\mathbf{R}^\top + \hat{\mathbf{L}}\mathbf{R}^\top - \hat{\mathbf{L}}\hat{\mathbf{R}}^\top\right\|_\sigma \\
&\leq \left\|(\mathbf{L} - \hat{\mathbf{L}})\mathbf{R}^\top\right\|_\sigma + \left\|\hat{\mathbf{L}}(\mathbf{R} - \hat{\mathbf{R}})^\top\right\|_\sigma \\
&\leq \|\mathbf{L} - \hat{\mathbf{L}}\|_\sigma \|\mathbf{R}\|_\sigma + \|\hat{\mathbf{L}}\|_\sigma \|\mathbf{R} - \hat{\mathbf{R}}\|_\sigma \leq \delta\sqrt{R} + \sqrt{R}\delta = 2\delta\sqrt{R}.
\end{aligned}$$

To obtain an ϵ-net for $\mathcal{B}_{R,r}$, we set $2\delta\sqrt{R} = \epsilon$, which gives $\delta = \frac{\epsilon}{2\sqrt{R}}$. The size of the net $\mathcal{N}_W$ is then bounded by:

$$|\mathcal{N}_W| \leq |\mathcal{N}_L| \cdot |\mathcal{N}_R| \leq \left(1 + \frac{2\sqrt{R}}{\epsilon/(2\sqrt{R})}\right)^{d_1 r} \left(1 + \frac{2\sqrt{R}}{\epsilon/(2\sqrt{R})}\right)^{d_2 r} = \left(1 + \frac{4R}{\epsilon}\right)^{r(d_1+d_2)}.$$

Taking the logarithm yields:

$$\log \mathcal{N}(\epsilon, \mathcal{B}_{R,r}, d_{op}) \leq r(d_1 + d_2) \log\left(1 + \frac{4R}{\epsilon}\right)$$

For high-precision covering where $\epsilon \leq R$, we can bound $(1 + 4R/\epsilon) \leq 5R/\epsilon$. There always exists an absolute constant C such that $\log(1 + 4R/\epsilon) \leq C \log(3R/\epsilon)$. Thus, the stated bound holds. □

Our analysis also requires the following result on integral estimation.

Lemma 16 (Lei (2026))**.** *Let $a, b > 0$. Then*

$$\int_0^b \log^{\frac{1}{2}}(a/\epsilon) \mathrm{d}\epsilon \leq b \log^{\frac{1}{2}}(a/b) + 2^{-1} b\sqrt{\pi}.$$

A.2 Bounds on Squared Norm of Gradients

In this subsection, we define the empirical second-order moment $\mathcal{L}_{*,S}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{W}; z_i)\|_*^2$, and the population second-order moment $\mathcal{L}_*(\mathbf{W}) = \mathbb{E}_z\left[\|\nabla f(\mathbf{W}; z)\|_*^2\right]$, and give their connection in the following lemma, which will be used in the proof of Theorem 1.

Lemma 17. *Let Assumption 2 hold. Define $G_* = \sup_{\mathbf{W}:\|\mathbf{W}-\mathbf{W}_0\|_\sigma} \|\nabla f(\mathbf{W}; \mathbf{z})\|_*$ for any $\mathbf{z}$. For any $x > 0$, with probability at least $1 - 2\exp(-x)$, the following inequalities hold simultaneously for all $\mathbf{W} \in \mathcal{B}_{R,r}$:*

$$\begin{aligned}
\mathcal{L}_{*,S}(\mathbf{W}) - 2\mathcal{L}_*(\mathbf{W}) &\lesssim \frac{G_*^2 \cdot r(d_1 + d_2)\log(L_* R/G_*)}{n} + \frac{G_*^2(x + \log\log n)}{n}, \\
\mathcal{L}_*(\mathbf{W}) - 2\mathcal{L}_{*,S}(\mathbf{W}) &\lesssim \frac{G_*^2 \cdot r(d_1 + d_2)\log(L_* R/G_*)}{n} + \frac{G_*^2(x + \log\log n)}{n}.
\end{aligned}$$

The proof of Lemma 17 requires the following lemma.

Lemma 18 (Lei (2026))**.** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ to $[0, b]$. Let ϕ_n be a sub-root function such that*

$$\mathbb{E}_\epsilon\left[\sup_{\substack{f\in\mathcal{F}: \\ \frac{1}{n}\sum_{i=1}^n f(z_i) \leq \rho}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i)\right] \leq \phi_n(\rho),$$

where ϵ_i are Rademacher variables. Define ρ_n^* as the solution of the equation $\phi_n(\rho) = \rho$, and denote $\rho_0 = \frac{b(x + 6\log\log n)}{n}$. Then, for any $x > 0$, with probability at least $1 - 2\exp(-x)$, the

following inequalities hold simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^{n} f(z_i) \le 2\mathbb{E}_z[f(z)] + 108r_n^* + 20\rho_0 + 38\sqrt{\rho_0\rho_n^*}, \tag{A.1}$$

$$\mathbb{E}_z[f(z)] \le \frac{2}{n}\sum_{i=1}^{n} f(z_i) + 106\rho_n^* + 48\rho_0. \tag{A.2}$$

Proof of Lemma 17. The proof follows exactly the structure of Lemma A.7 in Lei (2026), with Euclidean norms replaced by the operator-norm / nuclear-norm dual pair and covering numbers replaced by the low-rank matrix version.

Step 1: Define function classes and empirical metric. For any $\mathbf{W} \in \mathcal{B}_{R,r}$, define the scalar function

$$\tilde{f}_{\mathbf{W}}(z) := \|\nabla f(\mathbf{W}; z)\|_*^2.$$

We introduce the full function class and its localized version:

$$\tilde{\mathcal{F}}_{R,r} = \left\{ z \mapsto \tilde{f}_{\mathbf{W}}(z) : \mathbf{W} \in \mathcal{B}_{R,r} \right\}, \quad \tilde{\mathcal{F}}_{R,r}(\rho) = \left\{ \tilde{f}_{\mathbf{W}} \in \tilde{\mathcal{F}}_{R,r} : L_S(\mathbf{W}) \le \rho \right\}.$$

Define the empirical pseudometric over $\tilde{\mathcal{F}}_{R,r}$:

$$\tilde{d}_S(\tilde{f}_{\mathbf{W}}, \tilde{f}_{\mathbf{W}'}) = \Big(\frac{1}{n}\sum_{i=1}^{n} \big(\|\nabla f(\mathbf{W}; z_i)\|_*^2 - \|\nabla f(\mathbf{W}'; z_i)\|_*^2 \big)^2 \Big)^{\frac{1}{2}}.$$

Step 2: Bound the metric via smoothness. For any $\mathbf{W}, \mathbf{W}' \in \mathcal{B}_{R,r}$ with $L_S(\mathbf{W}) \le \rho$ and $L_S(\mathbf{W}') \le \rho$, by the reverse triangle inequality for nuclear norm and the identity $a^2 - b^2 = (a-b)(a+b)$, we have

$$\begin{aligned}\big| \|\nabla f(\mathbf{W}; z)\|_*^2 - \|\nabla f(\mathbf{W}'; z)\|_*^2 \big| &= \big| \|\nabla f(\mathbf{W}; z)\|_* - \|\nabla f(\mathbf{W}'; z)\|_* \big| \cdot \big(\|\nabla f(\mathbf{W}; z)\|_* + \|\nabla f(\mathbf{W}'; z)\|_* \big) \\ &\le \|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z)\|_* \cdot \big(\|\nabla f(\mathbf{W}; z)\|_* + \|\nabla f(\mathbf{W}'; z)\|_* \big).\end{aligned}$$

Squaring both sides and applying $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$\big(\|\nabla f(\mathbf{W}; z)\|_*^2 - \|\nabla f(\mathbf{W}'; z)\|_*^2 \big)^2 \le 2 \cdot \|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z)\|_*^2 \cdot \big(\|\nabla f(\mathbf{W}; z)\|_*^2 + \|\nabla f(\mathbf{W}'; z)\|_*^2 \big).$$

Taking the sample average and invoking the L_*-smoothness assumption

$$\|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z)\|_* \le L_* \cdot \|\mathbf{W} - \mathbf{W}'\|_\sigma,$$

we obtain

$$\begin{aligned}\tilde{d}_S^2(\tilde{f}_{\mathbf{W}}, \tilde{f}_{\mathbf{W}'}) &\le \frac{2L_*^2 \|\mathbf{W} - \mathbf{W}'\|_\sigma^2}{n} \sum_{i=1}^{n} \big(\|\nabla f(\mathbf{W}; z_i)\|_*^2 + \|\nabla f(\mathbf{W}'; z_i)\|_*^2 \big) \\ &= 2L_*^2 \|\mathbf{W} - \mathbf{W}'\|_\sigma^2 \cdot \big(L_S(\mathbf{W}) + L_S(\mathbf{W}') \big) \\ &\le 4L_*^2 \rho \cdot \|\mathbf{W} - \mathbf{W}'\|_\sigma^2.\end{aligned}$$

Equivalently,

$$\tilde{d}_S(\tilde{f}_{\mathbf{W}}, \tilde{f}_{\mathbf{W}'}) \le 2L_* \sqrt{\rho} \cdot \|\mathbf{W} - \mathbf{W}'\|_\sigma.$$

Step 3: Covering number of the function class. By the metric contraction above, an ϵ-cover of $\mathcal{B}_{R,r}$ under d_{op} induces an $\epsilon \cdot 2L_*\sqrt{\rho}$-cover of $\tilde{\mathcal{F}}_{R,r}(\rho)$ under $\tilde{d}_S$. Thus

$$\log \mathcal{N}\left(\epsilon, \tilde{\mathcal{F}}_{R,r}(\rho), \tilde{d}_S \right) \le \log \mathcal{N}\left(\frac{\epsilon}{2L_*\sqrt{\rho}}, \mathcal{B}_{R,r}, d_{op} \right).$$

Substitute the given covering number bound in Lemma 15

$$\log \mathcal{N}\left(\epsilon, \mathcal{B}_{R,r}, d_{op} \right) \le C \cdot r(d_1 + d_2) \log\left(\frac{3R}{\epsilon} \right),$$

we get

$$\log \mathcal{N}\left(\epsilon, \tilde{\mathcal{F}}_{R,r}(\rho), \tilde{d}_S\right) \leq C \cdot r(d_1+d_2) \log\left(\frac{6L_* R\sqrt{\rho}}{\epsilon}\right).$$

Step 4: Rademacher complexity via Dudley's entropy integral. Note that for any $\tilde{f}_{\mathbf{W}} \in \tilde{\mathcal{F}}_{R,r}(\rho)$,

$$\tilde{d}_S^2(\tilde{f}_{\mathbf{W}}, 0) = \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{W}; z_i)\|_*^4 \leq G_*^2 \cdot \frac{1}{n}\sum_{i=1}^n \|\nabla f(\mathbf{W}; z_i)\|_*^2 = G_*^2 L_S(\mathbf{W}) \leq G_*^2 \rho,$$

where we use the definition of G_* for the second step above. So the diameter of $\tilde{\mathcal{F}}_{R,r}(\rho)$ under $\tilde{d}_S$ is at most $D = G_*\sqrt{\rho}$. Applying Dudley's entropy integral (Lemma 13) and evaluating the integral via Lemma 16, we bound the empirical Rademacher complexity:

$$\begin{aligned}
\mathfrak{R}_S\big(\tilde{\mathcal{F}}_{R,r}(\rho)\big) &\leq \frac{12}{\sqrt{n}} \int_0^{G_*\sqrt{\rho}} \sqrt{\log \mathcal{N}\left(\epsilon, \tilde{\mathcal{F}}_{R,r}(\rho), \tilde{d}_S\right)} d\epsilon \\
&= \frac{12\sqrt{Cr(d_1+d_2)}}{\sqrt{n}} \int_0^{G_*\sqrt{\rho}} \sqrt{\log\left(\frac{6L_* R\sqrt{\rho}}{\epsilon}\right)} d\epsilon \\
&\leq \frac{12\sqrt{Cr(d_1+d_2)}}{\sqrt{n}} \left(G_*\sqrt{\rho}\log^{\frac{1}{2}}(6L_*R/G_*) + 2^{-1}G_*\sqrt{\rho\pi}\right) := \phi_n(\rho).
\end{aligned}$$

Step 5: Fixed point of the sub-root function. The function $\phi_n(\rho)$ is sub-root: it is non-decreasing, and $\phi_n(\rho)/\sqrt{\rho}$ is non-increasing. Its unique fixed point ρ_n^* solving $\phi_n(\rho_n^*) = \rho_n^*$ satisfies

$$\rho_n^* = \frac{12^2 G_*^2 Cr(d_1+d_2)}{n}\left(\log^{\frac{1}{2}}(6L_*R/G_*) + 2^{-1}\sqrt{\pi}\right)^2 \lesssim \frac{G_*^2 \cdot r(d_1+d_2)\log(L_*R/G_*)}{n}.$$

Step 6: Apply local concentration lemma. We apply Lemma 18 to the bounded function class $\tilde{\mathcal{F}}_{R,r}$ with uniform bound $b = G_*^2$ (since $\|\nabla f(\mathbf{W}; z)\|_*^2 \leq G_*^2$). Let

$$\rho_0 = \frac{b(x + 6\log\log n)}{n} = \frac{G_*^2(x+6\log\log n)}{n}.$$

With probability at least $1 - 2\exp(-x)$, the following hold simultaneously for all $\mathbf{W} \in \mathcal{B}_{R,r}$:

$$\begin{aligned}
L_{*,S}(\mathbf{W}) &\leq 2L_*(\mathbf{W}) + 108\rho_n^* + 20\rho_0 + 38\sqrt{\rho_0\rho_n^*}, \\
L_*(\mathbf{W}) &\leq 2L_{*,S}(\mathbf{W}) + 106\rho_n^* + 48\rho_0.
\end{aligned}$$

Under the assumption $\log\log n \lesssim r(d_1+d_2)\log(L_*R/G_*)$, the cross term $\sqrt{\rho_0\rho_n^*}$ is of lower order and can be absorbed. □

A.3 Proof of Theorem 1

Before presenting the proof of Theorem 1, we need to establish the concentration inequality in nuclear norm, which requires the following lemma.

Lemma 19 (Theorem 2.3 in Bousquet (2002))**.** *Assume the X_i are identically distributed according to P. Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ are P-measurable, square-integrable and satisfy $\mathbb{E}[f] = 0$. If $\sup_{f\in\mathcal{F}} \operatorname{ess\,sup} f \leqslant 1$ then we denote $Z = \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i)$. Let σ be a positive real number such that $\sigma^2 \geqslant \sup_{f\in\mathcal{F}} \operatorname{Var}[f(X_1)]$ almost surely, then for all $x \geqslant 0$, with $v = n\sigma^2 + 2\mathbb{E}[Z]$, we have*

$$\mathbb{P}\left[Z \geqslant \mathbb{E}[Z] + \sqrt{2xv} + \frac{x}{3}\right] \leqslant \mathrm{e}^{-x}.$$

Using Lemma 19, we establish the following concentration inequality in nuclear norm.

Lemma 20 (Concentration in nuclear norm)**.** *Let ξ be a random matrix taking values in $\mathbb{R}^{d_1\times d_2}$ and $\sigma^2 = \operatorname{Tr}(\operatorname{Cov}(\xi))$, where $\operatorname{Tr}(A) = \sum_i A_{ii}$ for the matrix A, and $\operatorname{Cov}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])(X-\mathbb{E}[X])^\top\big]$. Assume that $\|\xi\|_* \leq \widetilde{M} < \infty$ almost surely. Let $\xi_1, \ldots, \xi_n$ be independent copies of ξ. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$*

$$\left\|\frac{1}{n}\sum_{i=1}^n (\xi_i - \mathbb{E}\xi_i)\right\|_* \leq \sigma\sqrt{\frac{q}{n}} + \left(\frac{2\big(\mathbb{E}[\|\xi\|_*^2] + 4\widetilde{M}\sigma\sqrt{q/n}\big)\log(1/\delta)}{n}\right)^{\frac{1}{2}} + \frac{2\widetilde{M}\log(1/\delta)}{3n},$$

where $q = \min\{d_1, d_2\}$.

Proof of Lemma 20. Set

$$Y_i := \xi_i - \mathbb{E}\xi_i$$

and define

$$S_n := \Big\| \frac{1}{n} \sum_{i=1}^{n} Y_i \Big\|_*, \qquad R_n := \mathbb{E} S_n.$$

We first bound R_n. Since the rank of any matrix $A \in \mathbb{R}^{d_1 \times d_2}$ is at most $q = \min\{d_1, d_2\}$, and $\|A\|_* \le \sqrt{q}\|A\|_{\mathrm{F}}$,

$$R_n \le \frac{\sqrt{q}}{n} \mathbb{E} \Big\| \sum_{i=1}^{n} Y_i \Big\|_{\mathrm{F}} \le \frac{\sqrt{q}}{n} \Big(\mathbb{E} \Big\| \sum_{i=1}^{n} Y_i \Big\|_{\mathrm{F}}^2 \Big)^{1/2},$$

where we use Jensen's inequality for the second step. Since $Y_1, \dots, Y_n$ are independent and centered, we know

$$\mathbb{E} \Big\| \sum_{i=1}^{n} Y_i \Big\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \mathbb{E}\|Y_i\|_{\mathrm{F}}^2.$$

Moreover, we know

$$\mathbb{E}\|Y_i\|_{\mathrm{F}}^2 = \mathbb{E}\|\xi_i - \mathbb{E}\xi_i\|_{\mathrm{F}}^2 = \mathrm{Tr}(\mathrm{Cov}(\xi_i)).$$

It then follows that

$$R_n \le \frac{1}{n} \Big(q \sum_{i=1}^{n} \mathrm{Tr}(\mathrm{Cov}(\xi_i)) \Big)^{\frac{1}{2}}.$$

We next apply an empirical-process concentration inequality (Lemma 19). By the duality between the nuclear norm and the operator norm, we know

$$S_n = \sup_{\|W\|_\sigma \le 1} \frac{1}{n} \sum_{i=1}^{n} \langle W, Y_i \rangle_F,$$

where

$$\langle A, B \rangle_F := \mathrm{tr}(A^\top B).$$

Since $\|\xi_i\|_* \le \widetilde{M}$ almost surely, Jensen's inequality gives

$$\|\mathbb{E}\xi_i\|_* \le \mathbb{E}\|\xi_i\|_* \le \widetilde{M}.$$

Therefore, we get

$$\|Y_i\|_* \le \|\xi_i\|_* + \|\mathbb{E}\xi_i\|_* \le 2\widetilde{M}$$

almost surely. Hence, for every $\|W\|_\sigma \le 1$, we know

$$|\langle W, Y_i \rangle_F| \le \|W\|_\sigma \|Y_i\|_* \le 2\widetilde{M}.$$

The weak variance parameter satisfies

$$\begin{aligned} v^2 &:= \sup_{\|W\|_\sigma \le 1} \mathbb{E}\langle W, Y_i \rangle_F^2 = \sup_{\|W\|_\sigma \le 1} \mathbb{E}\big[\langle \mathbf{w}, \xi_i - \mathbb{E}[\xi_i] \rangle^2\big] \\ &\le \sup_{\|W\|_\sigma \le 1} \mathbb{E}\langle W, \xi_i \rangle_F^2 \le \mathbb{E}\|\xi_i\|_*^2. \end{aligned}$$

Bousquet's inequality for bounded empirical processes (Lemma 19) now yields, for every $t > 0$,

$$\mathbb{P}\Big(S_n \ge R_n + \Big(\frac{2(v^2 + 4\widetilde{M} R_n)t}{n} \Big)^{\frac{1}{2}} + \frac{2\widetilde{M}t}{3n} \Big) \le e^{-t}.$$

Using the inequalities

$$R_n \le \frac{1}{n} \Big(q \sum_{i=1}^{n} \mathrm{Tr}(\mathrm{Cov}(\xi_i)) \Big)^{\frac{1}{2}}, \qquad v^2 \le \mathbb{E}\|\xi_i\|_*^2,$$

and taking $t = \log(1/\delta)$ proves the result. □

We now present the poorf of Theorem 1.

Proof of Theorem 1. Set

$$q := \min\{d_1, d_2\}, \qquad \epsilon := \frac{R}{n}.$$

Let $\{\mathbf{W}_1, \dots, \mathbf{W}_N\}$ be an ϵ-cover of $\mathcal{B}_{R,r}$ under the operator norm. By Lemma 15,

$$\log N \lesssim r(d_1 + d_2) \log\left(\frac{3R}{\epsilon}\right) = r(d_1 + d_2)\log(3n).$$

Assumption 2 implies that, for every $\mathbf{W} \in \mathcal{B}_{R,r}$ and every z,

$$\begin{aligned} G_* = \|\nabla f(\mathbf{W}; z)\|_* &\le \|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{0}; z)\|_* + \|\nabla f(\mathbf{0}; z)\|_* \\ &\le L_* \|\mathbf{W}\|_\sigma + b_* \le L_* R + b_*. \end{aligned} \tag{A.3}$$

Fix a covering point $\mathbf{W}_j$ and define $\xi_i^{(j)} := \nabla f(\mathbf{W}_j; z_i)$. Then $\xi_1^{(j)}, \dots, \xi_n^{(j)}$ are independent copies of $\xi^{(j)} := \nabla f(\mathbf{W}_j; z)$ and satisfy $\|\xi^{(j)}\|_* \le G_*$ almost surely, $\sigma^2(\xi^{(j)}) \le \mathbb{E}_z \|\nabla f(\mathbf{W}_j; z)\|_*^2 = L_*(\mathbf{W}_j)$, and $\operatorname{rank}(\frac{1}{n}\sum_{i=1}^n (\xi_i^{(j)} - \mathbb{E}\xi_i^{(j)})) \le q$. Applying Lemma 20 with failure probability $\delta/(2N)$ and taking a union bound over $j = 1, \dots, N$, we obtain, with probability at least $1 - \delta/2$, simultaneously for all j,

$$\begin{aligned} &\|\nabla F_S(\mathbf{W}_j) - \nabla F(\mathbf{W}_j)\|_* \\ &\le \sqrt{\frac{qL_*(\mathbf{W}_j)}{n}} + \sqrt{\frac{2\left(L_*(\mathbf{W}_j) + 4G_*\sqrt{L_*(\mathbf{W}_j)}\sqrt{q/n}\right)a}{n}} + \frac{2G_* a}{3n} \\ &\lesssim \frac{G_* a}{n} + \sqrt{\frac{L_*(\mathbf{W}_j)a}{n}} + \sqrt{\frac{qL_*(\mathbf{W}_j)}{n}} \\ &\lesssim \frac{G_*\Gamma_{n,\delta}}{n} + \sqrt{\frac{L_*(\mathbf{W}_j)(\Gamma_{n,\delta} + q)}{n}} \\ &\lesssim \frac{G_*\Gamma_{n,\delta}}{n} + \sqrt{\frac{L_*(\mathbf{W}_j)\Gamma_{n,\delta}}{n}}, \end{aligned} \tag{A.4}$$

where $a := \log\frac{2N}{\delta} \lesssim \Gamma_{n,\delta}$. The second inequality follows from $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and $2\sqrt{xy} \le x + y$ for $x, y \ge 0$. The last inequality follows from $\Gamma_{n,\delta} \ge r(d_1 + d_2)\log(3n) \gtrsim q$, since $r \ge 1$ and $d_1 + d_2 \ge 2q$. Applying Lemma 17 with the corresponding confidence parameter gives, with probability at least $1 - \delta/2$, simultaneously for all $\mathbf{W} \in \mathcal{B}_{R,r}$, there holds

$$L_*(\mathbf{W}) \lesssim L_{*,S}(\mathbf{W}) + \frac{G_*^2\Gamma_{n,\delta}}{n}. \tag{A.5}$$

We work on the intersection of the events in Eq. (A.4) and Eq. (A.5), which has probability at least $1 - \delta$. Fix any $\mathbf{W} \in \mathcal{B}_{R,r}$, and choose $\mathbf{W}_j$ such that

$$\|\mathbf{W} - \mathbf{W}_j\|_\sigma \le \epsilon.$$

By Assumption 2, we know

$$\begin{aligned} &\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* \\ &\le \|\nabla F_S(\mathbf{W}_j) - \nabla F(\mathbf{W}_j)\|_* + \|\nabla F_S(\mathbf{W}) - \nabla F_S(\mathbf{W}_j)\|_* + \|\nabla F(\mathbf{W}) - \nabla F(\mathbf{W}_j)\|_* \\ &\le \|\nabla F_S(\mathbf{W}_j) - \nabla F(\mathbf{W}_j)\|_* + 2L_*\epsilon. \end{aligned} \tag{A.6}$$

The same smoothness assumption gives

$$\|\nabla f(\mathbf{W}_j; z_i)\|_* \le \|\nabla f(\mathbf{W}; z_i)\|_* + L_*\epsilon.$$

Therefore, there holds

$$L_{*,S}(\mathbf{W}_j) \le 2L_{*,S}(\mathbf{W}) + 2L_*^2\epsilon^2. \tag{A.7}$$

Combining Eq. (A.5) and Eq. (A.7), we obtain

$$L_*(\mathbf{W}_j) \lesssim L_{*,S}(\mathbf{W}) + L_*^2\epsilon^2 + \frac{G_*^2\Gamma_{n,\delta}}{n}. \tag{A.8}$$

Substituting Eq. (A.8) into Eq. (A.4) and using $\sqrt{x+y+z} \le \sqrt{x} + \sqrt{y} + \sqrt{z}$ gives

$$\|\nabla F_S(\mathbf{W}_j) - \nabla F(\mathbf{W}_j)\|_* \lesssim \frac{G_* \Gamma_{n,\delta}}{n} + \sqrt{\frac{L_{*,S}(\mathbf{W})\Gamma_{n,\delta}}{n}} + L_* \epsilon \sqrt{\frac{\Gamma_{n,\delta}}{n}}. \tag{A.9}$$

Here, the contribution from $G_*^2 \Gamma_{n,\delta}/n$ has been absorbed into $G_* \Gamma_{n,\delta}/n$. Combining Eq. (A.6) and Eq. (A.9), we obtain

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* \lesssim \frac{G_* \Gamma_{n,\delta}}{n} + \sqrt{\frac{L_{*,S}(\mathbf{W})\Gamma_{n,\delta}}{n}} + L_* \epsilon \sqrt{\frac{\Gamma_{n,\delta}}{n}} + 2L_* \epsilon. \tag{A.10}$$

Finally, since $\epsilon = R/n$ and $L_* R \le G_*$, we know

$$L_* \epsilon = \frac{L_* R}{n} \le \frac{G_*}{n}.$$

Thus, the last two terms in Eq. (A.10) are absorbed into $G_* \Gamma_{n,\delta}/n$. Therefore,

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* \lesssim \frac{G_* \Gamma_{n,\delta}}{n} + \sqrt{\frac{L_{*,S}(\mathbf{W})\Gamma_{n,\delta}}{n}}.$$

Substituting the fact that $G_* \le L_* R + b_*$ (Eq. (A.3)) completes the proof. □

B Proof of Generalization Analysis of Muon

In this section, we present the proofs of the generalization analysis of Muon. We give the proof of the generalization gap in Section B.1, the proof of the optimization error in Section B.2, and the proof of the excess risk in Section B.3.

B.1 Proof of Generalization Gap of Muon

We first give the proof of Lemma 2, which gives the operator-norm bound of Muon.

Proof of Lemma 2. Unrolling the momentum recursion gives

$$\mathbf{M}_t = \beta^t \mathbf{G}_0 + (1-\beta) \sum_{i=1}^{t} \beta^{t-i} \mathbf{G}_i.$$

Thus, the momentum parameter β affects the singular-vector matrices $\mathbf{U}_t$ and $\mathbf{V}_t$. However, because the update direction is orthogonalized, it does not increase the operator norm of a single update.

Indeed, since $\mathbf{M}_t = \mathbf{U}_t \mathbf{S}_t \mathbf{V}_t^\top$ is a compact singular value decomposition, the columns of $\mathbf{U}_t$ and $\mathbf{V}_t$ are orthonormal. Therefore,

$$\|\mathbf{U}_t \mathbf{V}_t^\top\|_\sigma = 1$$

whenever $\mathbf{M}_t \neq 0$. If $\mathbf{M}_t = 0$, we define the update direction to be zero, in which case

$$\|\mathbf{U}_t \mathbf{V}_t^\top\|_\sigma \le 1.$$

It follows that

$$\|\mathbf{W}_{t+1} - \mathbf{W}_t\|_\sigma = \eta_t \|\mathbf{U}_t \mathbf{V}_t^\top\|_\sigma \le \eta_t.$$

Unrolling the parameter recursion yields

$$\mathbf{W}_t = \mathbf{W}_0 - \sum_{k=0}^{t-1} \eta_k \mathbf{U}_k \mathbf{V}_k^\top.$$

By the triangle inequality for the operator norm,

$$\|\mathbf{W}_t\|_\sigma \le \|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{t-1} \eta_k \|\mathbf{U}_k \mathbf{V}_k^\top\|_\sigma \le \|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{t-1} \eta_k.$$

The proof is complete. □

We then apply Lemma 2 to derive the following generalization gap of Muon shown in Corollary 3.

Proof of Corollary 3. By Lemma 2, with the constant learning rate $\eta_t = \eta$, the operator norm of the iterates $\mathbf{W}_t$ for any $t \leq T$ is bounded by

$$\|\mathbf{W}_t\|_\sigma \leq \|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{t-1} \eta \leq \|\mathbf{W}_0\|_\sigma + T\eta.$$

This implies that for all $0 \leq t \leq T$, the iterates $\mathbf{W}_t$ reside within a parameter ball $\mathcal{B}_{R,r}$ with radius $R = \|\mathbf{W}_0\|_\sigma + T\eta$.
By Theorem 1, the uniform generalization gap bound holds simultaneously for all $\mathbf{W} \in \mathcal{B}_{R,r}$ with probability at least $1-\delta$. Substituting $R = \|\mathbf{W}_0\|_\sigma + T\eta$ into the bound from Theorem 1 directly yields

$$\|\nabla F_S(\mathbf{W}_t) - \nabla F(\mathbf{W}_t)\|_* \lesssim \frac{(L_*\|\mathbf{W}_0\|_\sigma + L_* T\eta + b_*)\Gamma_{n,\delta}}{n} + \sqrt{\frac{L_{*,S}(\mathbf{W}_t)\Gamma_{n,\delta}}{n}}.$$

This completes the proof. □

B.2 Proof of Convergence Analysis of Muon

We now gives the proof of the high-probability optimization bound in Theorem 4.
First, we control the deviation between the stochastic momentum $\mathbf{M}_t$ and its deterministic counterpart $\mathbf{C}_t$, which requires the following martingale concentration inequality.

Lemma 21 (Tarres & Yao (2014))**.** *Let $\{\xi_k\}_{k\in\mathbb{N}}$ be a martingale difference sequence in $\mathbb{R}^d$. Suppose that almost surely $\|\xi_k\|_2 \leq B$ and $\sum_{k=1}^t \mathbb{E}[\|\xi_k\|_2^2 \mid \xi_1, \ldots, \xi_{k-1}] \leq \sigma_t^2$. Then, for any $0 < \delta < 1$, the following inequality holds with probability at least $1-\delta$*

$$\max_{1\leq j\leq t} \Big\| \sum_{k=1}^{j} \xi_k \Big\|_2 \leq 2\Big(\frac{B}{3} + \sigma_t\Big) \log(2/\delta).$$

Applying Lemma 21, we derive the high-probability momentum concentration.

Lemma 22 (High-probability momentum concentration)**.** *For every $T \geq 1$ and $\delta \in (0,1)$, $\mathbf{M}_t$ and $\mathbf{W}_t$ generated by Algorithm 1, defining $\mathbf{C}_0 = \nabla F_S(\mathbf{W}_0)$, $\mathbf{C}_t = \beta\mathbf{C}_{t-1} + (1-\beta)\nabla F_S(\mathbf{W}_t)$. Define $G_F = \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \|\nabla f(\mathbf{W};\mathbf{z}) - \nabla F_S(\mathbf{W})\|_{\mathrm{F}}$ for any $\mathbf{z}$. Then with probability at least $1-\delta$, simultaneously for all $t = 0, \ldots, T-1$,*

$$\|\mathbf{M}_t - \mathbf{C}_t\|_{\mathrm{F}} \leq G_F\beta^t + 2\left(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\right)\log\left(\frac{2}{\delta}\right), \tag{B.1}$$

where $\omega_t^2 = \frac{1-\beta}{1+\beta}\left(1-\beta^{2t}\right)$, $b_t = 1-\beta$ if $t \geq 1$; 0 if t=0.

Proof of Lemma 22. Let $Z_t = \mathbf{G}_t - \nabla F_S(\mathbf{W}_t)$, and let $\mathcal{F}_{t-1}$ denote the history before sampling the t-th mini-batch. We know for every $t \geq 0$, $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$, $\|Z_t\|_{\mathrm{F}} \leq G_F$ almost surely, and $\mathbb{E}[\|Z_t\|_{\mathrm{F}}^2 \mid \mathcal{F}_{t-1}] \leq \frac{\sigma^2}{B}$. Note that $\mathbf{M}_0 = \mathbf{G}_0$, $\mathbf{M}_t = \beta\mathbf{M}_{t-1} + (1-\beta)\mathbf{G}_t$. By definition of Z_t, expanding the momentum recursions gives

$$\mathbf{M}_t - \mathbf{C}_t = \beta^t Z_0 + (1-\beta)\sum_{i=1}^{t} \beta^{t-i} Z_i.$$

By triangle inequality and the almost sure bound $\|Z_0\|_{\mathrm{F}} \leq G_F$, we have

$$\|\mathbf{M}_t - \mathbf{C}_t\|_{\mathrm{F}} \leq G_F\beta^t + \Big\|(1-\beta)\sum_{i=1}^{t} \beta^{t-i} Z_i\Big\|_{\mathrm{F}}.$$

For each fixed t, define $a_{t,i} = (1-\beta)\beta^{t-i}$, $i = 1, \ldots, t$. Then the summation part is $\sum_{i=1}^t a_{t,i} Z_i$, and

$$\sum_{i=1}^{t} a_{t,i}^2 = (1-\beta)^2 \sum_{i=1}^{t} \beta^{2(t-i)} = \frac{1-\beta}{1+\beta}\left(1-\beta^{2t}\right) = \omega_t^2.$$

For each fixed t, the sequence $\{a_{t,i}Z_i\}_{i=1}^t$ is a matrix-valued martingale difference sequence. After vectorizing the matrices, it is a martingale difference sequence in $\mathbb{R}^{d_1 d_2}$, with Euclidean norm equal to the Frobenius norm. Furthermore, by the definition of b_t, we know $|a_{t,i}| \le b_t$, and hence $\|a_{t,i}Z_i\|_{\mathrm{F}} \le G_F b_t$. Therefore,

$$\sum_{i=1}^t \mathbb{E}\left[\|a_{t,i}Z_i\|_{\mathrm{F}}^2 \big| \mathcal{F}_{i-1}\right] = \sum_{i=1}^t a_{t,i}^2 \mathbb{E}\left[\|Z_i\|_{\mathrm{F}}^2 \big| \mathcal{F}_{i-1}\right] \le \frac{\sigma^2}{B}\sum_{i=1}^t a_{t,i}^2 = \frac{\sigma^2\omega_t^2}{B}.$$

Applying Lemma 21 to $\{a_{t,i}\,\mathrm{vec}(Z_i)\}_{i=1}^t$, with probability at least $1-\delta$,

$$\max_{1\le t\le T-1}\Big\|\sum_{i=1}^t a_{t,i}Z_i\Big\|_{\mathrm{F}} \le 2\left(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\right)\log\left(\frac{2}{\delta}\right).$$

Then, substituting this back, with probability at least $1-\delta$, simultaneously for all $t = 0, \dots, T-1$,

$$\|\mathbf{M}_t - \mathbf{C}_t\|_{\mathrm{F}} \le G_F\beta^t + 2\left(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\right)\log\left(\frac{2}{\delta}\right).$$

The proof is complete. □

Second, we recall a descent inequality adapted to the operator norm.

Lemma 23 (Operator-norm descent lemma Beck (2017))**.** *Let $g : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}$ be differentiable and L_*-smooth with respect to the operator norm, i.e., $\|\nabla g(\mathbf{W}) - \nabla g(\mathbf{W}')\|_* \le L_*\|\mathbf{W}-\mathbf{W}'\|_\sigma$ for any $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{d_1\times d_2}$. Then,*

$$g(\mathbf{W}') \le g(\mathbf{W}) + \langle \nabla g(\mathbf{W}), \mathbf{W}'-\mathbf{W}\rangle + \frac{L_*}{2}\|\mathbf{W}'-\mathbf{W}\|_\sigma^2.$$

We are now ready to prove Theorem 4. The idea is to apply Lemma 23 along the Muon trajectory and then use Lemma 22 to replace the stochastic momentum terms by their deterministic counterparts, up to a high-probability residual.

Proof of Theorem 4. Since f is L_* spectral norm Lipschitz smooth (Assumption 2), we know F_S is L_* spectral norm Lipschitz smooth, then by Lemma 23, we have

$$F_S(\mathbf{W}_{t+1}) \le F_S(\mathbf{W}_t) + \langle \nabla F_S(\mathbf{W}_t), \mathbf{W}_{t+1}-\mathbf{W}_t\rangle + \frac{L_*}{2}\|\mathbf{W}_{t+1}-\mathbf{W}_t\|_\sigma^2.$$

The Muon step with constant stepsize $\eta_t = \eta$ is $\mathbf{W}_{t+1} = \mathbf{W}_t - \eta\mathbf{U}_t\mathbf{V}_t^\top$, hence

$$\begin{aligned}
F_S(\mathbf{W}_t) - F_S(\mathbf{W}_{t+1}) &\ge \eta\left\langle \nabla F_S(\mathbf{W}_t), \mathbf{U}_t\mathbf{V}_t^\top\right\rangle - \frac{L_*}{2}\eta^2\|\mathbf{U}_t\mathbf{V}_t^\top\|_\sigma^2 \\
&\ge \eta\left\langle \mathbf{M}_t, \mathbf{U}_t\mathbf{V}_t^\top\right\rangle - \frac{L_*}{2}\eta^2 - \eta\left\langle \nabla F_S(\mathbf{W}_t) - \mathbf{M}_t, \mathbf{U}_t\mathbf{V}_t^\top\right\rangle \\
&\ge \eta\left\langle \mathbf{M}_t, \mathbf{U}_t\mathbf{V}_t^\top\right\rangle - \frac{L_*}{2}\eta^2 - \eta\|\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t\|_*\|\mathbf{U}_t\mathbf{V}_t^\top\|_\sigma \\
&\ge \eta\|\mathbf{M}_t\|_* - \frac{L_*}{2}\eta^2 - \eta\|\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t\|_* \\
&\ge \eta\|\nabla F_S(\mathbf{W}_t)\|_* - \frac{L_*}{2}\eta^2 - 2\eta\|\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t\|_*, \qquad \text{(B.2)}
\end{aligned}$$

where we use $\|\mathbf{U}_t\mathbf{V}_t^\top\|_\sigma \le 1$ for the second step, $\left\langle \mathbf{M}_t, \mathbf{U}_t\mathbf{V}_t^\top\right\rangle = \|\mathbf{M}_t\|_*$ for the last second step, and $\|\mathbf{M}_t\|_* \ge \|\nabla F_S(\mathbf{W}_t)\|_* - \|\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t\|_*$ for the last step. Then, we need to analyze and bound the error $\|\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t\|_*$. Note that $\mathbf{M}_0 = \mathbf{G}_0$, and, for $t > 0$,

$$\mathbf{M}_t = \beta\mathbf{M}_{t-1} + (1-\beta)\mathbf{G}_t = (1-\beta)\sum_{i=1}^t \beta^{t-i}\mathbf{G}_i + \beta^t\mathbf{G}_0. \qquad \text{(B.3)}$$

By Lemma 22 and let $\mathbf{C}_t$ defined there, there exists an event $\mathcal{E}_\delta$ with $\mathbb{P}(\mathcal{E}_\delta) \geq 1-\delta$ such that, simultaneously for every $t = 0, \ldots, T-1$,

$$\|\mathbf{M}_t - \mathbf{C}_t\|_{\mathrm{F}} \leq G_F \beta^t + 2\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big)\log(2/\delta). \tag{B.4}$$

Since the rank of any matrix $A \in \mathbb{R}^{d_1 \times d_2}$ is at most $q = \min\{d_1, d_2\}$, and $\|A\|_* \leq \sqrt{q}\|A\|_{\mathrm{F}}$, on $\mathcal{E}_\delta$, simultaneously for every $t = 0, \ldots, T-1$,

$$\|\mathbf{M}_t - \mathbf{C}_t\|_* \leq \sqrt{q}\,\|\mathbf{M}_t - \mathbf{C}_t\|_{\mathrm{F}} \leq \sqrt{q} G_F \beta^t + 2\sqrt{q}\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big)\log(2/\delta). \tag{B.5}$$

Moreover, when $t > 0$, we note that

$$\begin{aligned}
\|\nabla F_S(\mathbf{W}_t) - \mathbf{C}_t\|_* &= \|\nabla F_S(\mathbf{W}_t) - (\beta\mathbf{C}_{t-1} + (1-\beta)\nabla F_S(\mathbf{W}_t))\|_* \\
&= \beta\|\nabla F_S(\mathbf{W}_t) - \mathbf{C}_{t-1}\|_* \\
&\leq \beta\|\nabla F_S(\mathbf{W}_{t-1}) - \mathbf{C}_{t-1}\|_* + \beta\|\nabla F_S(\mathbf{W}_{t-1}) - \nabla F_S(\mathbf{W}_t)\|_* \\
&\leq \beta\|\nabla F_S(\mathbf{W}_{t-1}) - \mathbf{C}_{t-1}\|_* + \beta L_* \|\mathbf{W}_{t-1} - \mathbf{W}_t\|_\sigma \\
&= \beta\|\nabla F_S(\mathbf{W}_{t-1}) - \mathbf{C}_{t-1}\|_* + \beta L_* \eta \|\mathbf{U}_{t-1}\mathbf{V}_{t-1}^\top\|_\sigma \\
&\leq \beta\|\nabla F_S(\mathbf{W}_{t-1}) - \mathbf{C}_{t-1}\|_* + \beta L_* \eta,
\end{aligned} \tag{B.6}$$

where the second inequality follows from Assumption 2. Since $\|\nabla F_S(\mathbf{W}_0) - C_0\|_* = 0$, recursively applying (B.6) gives

$$\|\nabla F_S(\mathbf{W}_t) - \mathbf{C}_t\|_* \leq \sum_{i=1}^{t} \beta^i L_* \eta = \frac{\beta(1-\beta^t)}{1-\beta} L_* \eta \leq \frac{\beta L_* \eta}{1-\beta}. \tag{B.7}$$

Thus, on $\mathcal{E}_\delta$, simultaneously for every $t = 0, \ldots, T-1$, we have

$$\begin{aligned}
\|\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t\|_* &\leq \|\mathbf{C}_t - \mathbf{M}_t\|_* + \|\nabla F_S(\mathbf{W}_t) - \mathbf{C}_t\|_* \\
&\leq \sqrt{q} G_F \beta^t + 2\sqrt{q}\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big)\log(2/\delta) + \frac{\beta L_* \eta}{1-\beta}.
\end{aligned} \tag{B.8}$$

Substituting (B.8) into (B.2), we obtain, on $\mathcal{E}_\delta$,

$$\begin{aligned}
F_S(\mathbf{W}_t) - F_S(\mathbf{W}_{t+1}) \geq \eta\|\nabla F_S(\mathbf{W}_t)\|_* - \frac{L_*}{2}\eta^2 - \frac{2\beta L_* \eta^2}{1-\beta} - 2\eta\sqrt{q} G_F \beta^t \\
- 4\eta\sqrt{q}\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big)\log(2/\delta).
\end{aligned} \tag{B.9}$$

Summing (B.9) over $t = 0, 1, \ldots, T-1$, and using $\sum_{t=0}^{T-1} \beta^t \leq \frac{1}{1-\beta}$, we get

$$\begin{aligned}
F_S(\mathbf{W}_0) - F_S(\mathbf{W}_T) \geq \eta \sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* - \frac{TL_*}{2}\eta^2 - \frac{2T\beta L_* \eta^2}{1-\beta} - \frac{2\eta\sqrt{q} G_F}{1-\beta} \\
- 4\eta\sqrt{q}\log(2/\delta)\sum_{t=0}^{T-1}\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big).
\end{aligned} \tag{B.10}$$

Rearranging terms and dividing both sides by $T\eta$ yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* \leq \frac{F_S(\mathbf{W}_0) - F_S(\mathbf{W}_T)}{T\eta} + \frac{L_*\eta}{2} + \frac{2\eta\beta L_*}{1-\beta} + \frac{2\sqrt{q} G_F}{T(1-\beta)} \\
+ \frac{4\sqrt{q}}{T}\log(2/\delta)\sum_{t=0}^{T-1}\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big).
\end{aligned} \tag{B.11}$$

Since $F_S(\mathbf{W}_T) \geq F_S^\star$, we have $F_S(\mathbf{W}_0) - F_S(\mathbf{W}_T) \leq F_S(\mathbf{W}_0) - F_S^\star$. Consequently, on $\mathcal{E}_\delta$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* \leq \frac{F_S(\mathbf{W}_0) - F_S^*}{T\eta} + \frac{L_*\eta}{2} + \frac{2\eta\beta L_*}{1-\beta} + \frac{2\sqrt{q} G_F}{T(1-\beta)} \\
+ \frac{4\sqrt{q}}{T}\log(2/\delta)\sum_{t=0}^{T-1}\Big(\frac{G_F b_t}{3} + \frac{\sigma\omega_t}{\sqrt{B}}\Big).
\end{aligned} \tag{B.12}$$

By the definition of b_t and ω_t,

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\frac{G_F b_t}{3}+\frac{\sigma\omega_t}{\sqrt{B}}\Big)=\frac{G_F}{3T}\sum_{t=1}^{T-1}(1-\beta)+\frac{\sigma}{T\sqrt{B}}\sum_{t=0}^{T-1}\sqrt{\frac{1-\beta}{1+\beta}(1-\beta^{2t})}$$
$$\le\frac{G_F(1-\beta)}{3}+\frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}. \tag{B.13}$$

Plugging Eq. (B.13) into Eq. (B.12) yields Eq. (4.1). The proof is complete by noting $\frac{L_*\eta}{2}+\frac{2\eta\beta L_*}{1-\beta}\le\frac{3\eta L_*}{1-\beta}$. □

Corollary 5 follows by substituting the stated choices of η and β into Theorem 4.

Proof of Corollary 5. With probability at least $1-\delta$, we start from Theorem 4 and bound the terms as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_S(\mathbf{W}_t)\|_*\le\frac{\Delta_S}{T\eta}+\frac{3\eta L_*}{1-\beta}+\frac{2\sqrt{q}\,G_F}{T(1-\beta)}+4\sqrt{q}\,\log(2/\delta)\left[\frac{G_F(1-\beta)}{3}+\frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}\right]$$
$$\le\frac{\Delta_S}{T\eta}+\frac{3\eta L_*}{1-\beta}+4\sqrt{q}\Gamma_\delta\left(\sqrt{1-\beta}+\frac{1}{T(1-\beta)}\right),$$

where we use $\frac{2\sqrt{q}\,G_F}{T(1-\beta)}\le\frac{4\sqrt{q}\Gamma_\delta}{T(1-\beta)}$, and the elementary bounds $\sqrt{\frac{1-\beta}{1+\beta}}\le\sqrt{1-\beta}$, $\beta\le1$, and $1-\beta\le\sqrt{1-\beta}$. Substituting $\eta=\sqrt{\frac{(1-\beta)\Delta_S}{TL_*}}$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_S(\mathbf{W}_t)\|_*\le4\sqrt{\frac{L_*\Delta_S}{T(1-\beta)}}+4\sqrt{q}\Gamma_\delta\sqrt{1-\beta}+\frac{4\sqrt{q}\Gamma_\delta}{T(1-\beta)}.$$

We now substitute the parameter: $1-\beta=\min\{\frac{\sqrt{L_*\Delta_S}}{\Gamma_\delta\sqrt{qT}},\ 1\}$, then $\sqrt{\frac{L_*\Delta_S}{T(1-\beta)}}=\big(\frac{qL_*\Delta_S\Gamma_\delta^2}{T}\big)^{\frac14}$, $\sqrt{1-\beta}=\big(\frac{L_*\Delta_S}{q\Gamma_\delta^2T}\big)^{\frac14}$, and hence

$$4\sqrt{\frac{L_*\Delta_S}{T(1-\beta)}}+4\sqrt{q}\Gamma_\delta\sqrt{1-\beta}+\frac{4\sqrt{q}\Gamma_\delta}{T(1-\beta)}$$
$$\le4\Big(\frac{qL_*\Delta_S\Gamma_\delta^2}{T}\Big)^{\frac14}+4\sqrt{q}\Gamma_\delta\Big(\frac{L_*\Delta_S}{q\Gamma_\delta^2T}\Big)^{\frac14}+4\sqrt{q}\Gamma_\delta\cdot\frac{\Gamma_\delta\sqrt{q}}{\sqrt{L_*\Delta_ST}}$$
$$\le8\Big(\frac{qL_*\Delta_S\Gamma_\delta^2}{T}\Big)^{\frac14}+\frac{4q\Gamma_\delta^2}{\sqrt{L_*\Delta_ST}}.$$

Combining the two inequalities above yields (4.2). In the deterministic setting, the stochastic momentum error is zero ($\mathbf{M}_t=\mathbf{C}_t$), so $\sigma=G_F=0$. Taking $\beta<1$ and $\eta=\sqrt{\Delta_S/(TL_*)}$, Theorem 4 reduces to

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_S(\mathbf{W}_t)\|_*\le\frac{\Delta_S}{T\eta}+\frac{3\eta L_*}{1-\beta}=\sqrt{\frac{L_*\Delta_S}{T}}+\frac{3}{1-\beta}\sqrt{\frac{L_*\Delta_S}{T}}\lesssim\sqrt{\frac{L_*\Delta_S}{T}},$$

which proves (4.3). The stated iteration complexities follow by requiring each term in the corresponding convergence bound to be at most ϵ. □

B.3 Proof of Excess Risk Analysis of Muon

Theorem 6 combines the generalization gap given in Corollary 3 and the optimization error given in Theorem 4.

Proof of Theorem 6. We know

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* = \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t) - \nabla F_S(\mathbf{W}_t) + \nabla F_S(\mathbf{W}_t)\|_* \\ \leq \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t) - \nabla F_S(\mathbf{W}_t)\|_* + \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_S(\mathbf{W}_t)\|_*. \quad \text{(B.14)}$$

It was shown in Corollary 3 that with probability at least $1-\delta$,

$$\|\nabla F_S(\mathbf{W}_t) - \nabla F(\mathbf{W}_t)\|_* \lesssim \frac{(L_* R_T + b_*)\Gamma_{n,\delta}}{n} + \sqrt{\frac{L_{*,S}(\mathbf{W}_t)\Gamma_{n,\delta}}{n}}$$

And the following optimization error bound was shown in Theorem 4 with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_S(\mathbf{W}_t)\|_* \leq \frac{F_S(\mathbf{W}_0) - F_S^\star}{T\eta} + \frac{3\eta L_*}{1-\beta} + \frac{2\sqrt{q}G_F}{T(1-\beta)} \\ + 4\sqrt{q}\log\left(\frac{2}{\delta}\right)\left[\frac{G_F(1-\beta)}{3} + \frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}\right].$$

We combine the above two inequalities and Eq. (B.14) together, and derive

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* \lesssim \frac{F_S(\mathbf{W}_0) - F_S^\star}{T\eta} + \frac{\eta L_*}{1-\beta} + \frac{\sqrt{q}G_F}{T(1-\beta)} \\ + \sqrt{q}\log\left(\frac{2}{\delta}\right)\left[G_F(1-\beta) + \frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}\right] + \frac{(L_* R_T + b_*)\Gamma_{n,\delta}}{n} + \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\frac{L_{*,S}(\mathbf{W}_t)\Gamma_{n,\delta}}{n}}.$$

By Assumption 1, we know

$$L_{*,S}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(\mathbf{W}; z_i)\|_*^2 = \mathbb{E}_{i_t}[\|\nabla f(\mathbf{W}; z_{i_t})\|_*^2] \\ \leq 2\mathbb{E}_{i_t}[\|\nabla f(\mathbf{W}; z_{i_t}) - \nabla F_S(\mathbf{W})\|_*^2] + 2\|\nabla F_S(\mathbf{W})\|_*^2 \leq 2\sigma^2 + 2\|\nabla F_S(\mathbf{W})\|_*^2.$$

It follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and Eq. (4.1) that

$$\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{L_{*,S}(\mathbf{W}_t)} \leq \sqrt{2}\sigma + \frac{\sqrt{2}}{T}\sum_{t=0}^{T-1}\|\nabla F_S(\mathbf{W}_t)\|_* \\ \leq \sqrt{2}\sigma + \frac{\sqrt{2}(F_S(\mathbf{W}_0) - F_S^\star)}{T\eta} + \frac{\sqrt{2}L_*\eta}{2} + \frac{2\sqrt{2}\eta\beta L_*}{1-\beta} + \frac{2\sqrt{2q}G_F}{T(1-\beta)} \\ + 4\sqrt{2q}\log\left(\frac{2}{\delta}\right)\left[\frac{G_F(1-\beta)}{3} + \frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}\right].$$

Then,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{W}_t)\|_* \lesssim \left(1 + \sqrt{\frac{\Gamma_{n,\delta}}{n}}\right)\left\{\frac{F_S(\mathbf{W}_0) - F_S^\star}{T\eta} + \frac{\eta L_*}{1-\beta} + \frac{\sqrt{q}G_F}{T(1-\beta)} \right. \\ \left. + \sqrt{q}\log(2/\delta)\left[G_F(1-\beta) + \frac{\sigma}{\sqrt{B}}\sqrt{\frac{1-\beta}{1+\beta}}\right]\right\} + \frac{(L_* R_T + b_*)\Gamma_{n,\delta}}{n} + \sigma\sqrt{\frac{\Gamma_{n,\delta}}{n}}.$$

The proof is complete by choosing $\eta = \sqrt{\frac{(1-\beta)\Delta_S}{TL_*}}$, $1-\beta = \min\{\frac{\sqrt{L_*\Delta_S}}{\Gamma_\delta\sqrt{qT}},\ 1\}$. □

C Proofs of Dimension-independent Bounds

In this section, we present the proofs of dimension-independent bounds for both linear networks (Section C.1) and SNNs (Section C.2).

C.1 Proof of Dimension-independent Bounds for Linear Networks

The following lemma gives a classical result on the covering numbers for linear function classes.

Lemma 24 (Zhang (2002)). *If $\|\phi(x)\|_2 \le B_\phi$, then for any $\epsilon > 0$, we have*

$$\log \mathcal{N}(\epsilon,\, \{x \to \langle \phi(x), \mathbf{w}\rangle : \|\mathbf{w}\|_2 \le R\},\, d_{S,\infty}) \le \frac{36R^2 B_\phi^2}{\epsilon^2} \log_2(6n + 8B_\phi R/\epsilon),$$

where $d_{S,\infty}(f,g) = \max_{i\in[n]} |f(x_i) - g(x_i)|$.

Applying Lemma 24 to $x \to a^\top \mathbf{W}\phi(x)$ yields the following covering number for low-rank linear networks.

Lemma 25 (Covering number for low-rank linear networks). *Let $S = (x_1, \dots, x_n)$ be a sample satisfying $\max_{1\le i\le n} \|\phi(x_i)\|_2 \le B_\phi$, and define $\mathcal{F}_{a,R,r} = \{f_{\mathbf{W}} : \mathbb{R}^{d_2} \to \mathbb{R},\ f_{\mathbf{W}}(x) = a^\top \mathbf{W}\phi(x) : \mathbf{W} \in \mathcal{B}_{R,r}\}$ and the metric $d_{S,\infty}(f,g) := \max_{1\le i\le n} |f(x_i) - g(x_i)|$. Then, for every $\epsilon > 0$,*

$$\log \mathcal{N}(\epsilon, \mathcal{F}_{a,R,r}, d_{S,\infty}) \le \frac{36R^2 \|a\|_2^2 B_\phi^2}{\epsilon^2} \log_2\big(6n + \frac{8R\|a\|_2 B_\phi}{\epsilon}\big). \tag{C.1}$$

Proof. If $a = 0$, then $\mathcal{F}_{a,R,r} = \{0\}$ and $\mathcal{N}(\epsilon, \mathcal{F}_{a,R,r}, d_{S,\infty}) = 1$, so (C.1) holds trivially. Henceforth assume $a \neq 0$.

For every $\mathbf{W} \in \mathcal{B}_{R,r}$, set

$$v := \mathbf{W}^\top a \in \mathbb{R}^{d_2}.$$

Then

$$a^\top \mathbf{W}\phi(x) = (\mathbf{W}^\top a)^\top \phi(x) = v^\top \phi(x)$$

and

$$\|v\|_2 = \|\mathbf{W}^\top a\|_2 \le \|\mathbf{W}\|_\sigma \|a\|_2 \le R\|a\|_2.$$

Consequently,

$$\mathcal{F}_{a,R,r} \subseteq \Big\{x \mapsto v^\top \phi(x) : \|v\|_2 \le R\|a\|_2\Big\}. \tag{C.2}$$

We now show that the reverse inclusion also holds when $r \ge 1$. Take an arbitrary vector $v \in \mathbb{R}^{d_2}$ with $\|v\|_2 \le R\|a\|_2$, and define

$$\mathbf{W}_v := \frac{a v^\top}{\|a\|_2^2}.$$

The matrix $\mathbf{W}_v$ is an outer product, and therefore $\mathrm{rank}(\mathbf{W}_v) \le 1 \le r$. Its operator norm satisfies

$$\|\mathbf{W}_v\|_\sigma = \frac{\|a\|_2 \|v\|_2}{\|a\|_2^2} = \frac{\|v\|_2}{\|a\|_2} \le R,$$

so $\mathbf{W}_v \in \mathcal{B}_{R,r}$. Moreover,

$$\mathbf{W}_v^\top a = \frac{v a^\top a}{\|a\|_2^2} = v,$$

and therefore

$$a^\top \mathbf{W}_v \phi(x) = v^\top \phi(x).$$

Combining this construction with (C.2) yields the identity

$$\mathcal{F}_{a,R,r} = \Big\{x \mapsto v^\top \phi(x) : \|v\|_2 \le R\|a\|_2\Big\}. \tag{C.3}$$

The right-hand side of (C.3) is the Euclidean linear class appearing in Lemma 24, with effective radius $R\|a\|_2$, and feature bound $\|\phi(x)\|_2 \le B_\phi$. Applying that result gives

$$\log \mathcal{N}\big(\epsilon, \mathcal{F}_{a,R,r}, d_{S,\infty}\big) \le \frac{36R^2 \|a\|_2^2 B_\phi^2}{\epsilon^2} \log_2\left(6n + \frac{8R\|a\|_2 B_\phi}{\epsilon}\right),$$

which is precisely (C.1). □

Our analyses require the following two lemmas on the properties of Gaussian complexities.

Lemma 26 (Sudakov minoration inequality Ledoux & Talagrand (1991))**.** *Let $\mathcal{F}$ be a class of real-valued functions, $S = \{z_1, \ldots, z_n\}$ and g_i be a sequence of $N(0,1)$ Gaussian random variables. Then*

$$\mathbb{E}_{\mathbf{g}}\Big[\sup_{f\in\mathcal{F}}\sum_{i\in[n]} g_i f(z_i)\Big] \geq \sqrt{n}\sup_{\epsilon>0}\epsilon\log^{\frac{1}{2}}\mathcal{N}(\epsilon,\mathcal{F},d_S),$$

where

$$d_S(f,g) = \Big(\frac{1}{n}\sum_{i=1}^{n}\big(f(z_i)-g(z_i)\big)^2\Big)^{\frac{1}{2}}.$$

Lemma 27 (Gaussian contraction principle Ledoux & Talagrand (1991))**.** *Let $T \subset \mathbb{R}^n$ and let $g_1, \ldots, g_n$ be i.i.d. $N(0,1)$. Suppose that for each $i \in [n]$, the map $\varphi_i : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz. Then*

$$\mathbb{E}_g \sup_{t\in T}\sum_{i=1}^{n} g_i\,\varphi_i(t_i) \leq L\,\mathbb{E}_g\sup_{t\in T}\sum_{i=1}^{n} g_i\,t_i. \tag{C.4}$$

The following lemma provides estimates on vector Rademacher complexities for linear networks.

Lemma 28. *Define $b_{\mathcal{W}} := \sup_{\mathbf{W}\in\mathcal{W}}\Big(\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i, a^\top\mathbf{W}\phi(x_i))^2\Big)^{1/2}$. Then*

$$\frac{1}{n}\,\mathbb{E}_\epsilon\sup_{\mathbf{W}\in\mathcal{W}}\Big\|\sum_{i=1}^{n}\epsilon_i\nabla_{\mathbf{W}}f(\mathbf{W};z_i)\Big\|_* \leq \frac{12}{\sqrt{n}} + \left(\frac{24\sqrt{2}\,\|a\|_2 B_x L_\ell}{n}\mathbb{E}_g\left[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i=1}^{n} g_i\,a^\top\mathbf{W}\phi(x_i)\right]\right.$$
$$\left. + \frac{144\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\sqrt{n}}\log_2^{1/2}\Big(2\big(6n+8\|a\|_2 B_x G_\ell\sqrt{n}\big)\Big)\right)\log\left(\frac{G_\ell\|a\|_2 B_x\sqrt{n}}{3}\right). \tag{C.5}$$

Furthermore,

$$\mathbb{E}_g\left[\sup_{\mathbf{W}\in\mathcal{W}}\sum_{i=1}^{n} g_i\,a^\top\mathbf{W}\phi(x_i)\right] \leq R\|a\|_2 B_x\sqrt{n}. \tag{C.6}$$

Consequently,

$$\frac{1}{n}\,\mathbb{E}_\epsilon\sup_{\mathbf{W}\in\mathcal{W}}\Big\|\sum_{i=1}^{n}\epsilon_i\nabla_{\mathbf{W}}f(\mathbf{W};z_i)\Big\|_* \leq \frac{12}{\sqrt{n}} + \tag{C.7}$$
$$\frac{\log\frac{G_\ell\|a\|_2 B_x\sqrt{n}}{3}}{\sqrt{n}}\left(24\sqrt{2}\,R\|a\|_2^2 B_x^2 L_\ell + 144\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}\log_2^{1/2}\Big(2\big(6n+8\|a\|_2 B_x G_\ell\sqrt{n}\big)\Big)\right).$$

Proof. If $a = 0$, then $\nabla_{\mathbf{W}}f(\mathbf{W};z) = 0$ for every $\mathbf{W}$ and z, and the conclusion is immediate. Henceforth assume $a \neq 0$. For every $\mathbf{W} \in \mathcal{W}$, differentiating with respect to $\mathbf{W}$ gives

$$\nabla_{\mathbf{W}}f(\mathbf{W};z) = \ell'\big(y, a^\top\mathbf{W}\phi(x)\big)\,a\phi(x)^\top. \tag{C.8}$$

For $\mathbf{W} \in \mathcal{W}$ and $\mathbf{C} \in \mathbb{R}^{d_1\times d_2}$ with $\|\mathbf{C}\|_{\mathrm{op}} \leq 1$, define

$$\begin{aligned} h_{\mathbf{W},\mathbf{C}}(z) &:= \langle\nabla_{\mathbf{W}}f(\mathbf{W};z),\mathbf{C}\rangle_F \\ &= \ell'\big(y,a^\top\mathbf{W}\phi(x)\big)\langle a\phi(x)^\top,\mathbf{C}\rangle_F \\ &= \ell'\big(y,a^\top\mathbf{W}\phi(x)\big)a^\top\mathbf{C}\phi(x). \end{aligned} \tag{C.9}$$

By the dual representation of the nuclear norm,

$$\|\mathbf{M}\|_* = \sup_{\|\mathbf{C}\|_{\mathrm{op}}\leq 1}\langle\mathbf{M},\mathbf{C}\rangle_F,$$

we have

$$\frac{1}{n}\,\mathbb{E}_\epsilon\sup_{\mathbf{W}\in\mathcal{W}}\left\|\sum_{i=1}^{n}\epsilon_i\nabla_{\mathbf{W}}f(\mathbf{W};z_i)\right\|_* = \frac{1}{n}\,\mathbb{E}_\epsilon\sup_{\substack{\mathbf{W}\in\mathcal{W}\\ \|\mathbf{C}\|_{\mathrm{op}}\leq 1}}\sum_{i=1}^{n}\epsilon_i h_{\mathbf{W},\mathbf{C}}(z_i). \tag{C.10}$$

Step 1: Metric comparison. Define

$$h^{(1)}_{\mathbf{W}}(z) := \ell'\big(y, a^\top \mathbf{W}\phi(x)\big), \\ h^{(2)}_{\mathbf{C}}(x) := \langle a\phi(x)^\top, \mathbf{C}\rangle_F = a^\top \mathbf{C}\phi(x). \tag{C.11}$$

For any $\mathbf{W}, \mathbf{W}' \in \mathcal{W}$ and $\|\mathbf{C}\|_{\mathrm{op}}, \|\mathbf{C}'\|_{\mathrm{op}} \le 1$, write

$$h_{\mathbf{W},\mathbf{C}} - h_{\mathbf{W}',\mathbf{C}'} = \big(h^{(1)}_{\mathbf{W}} - h^{(1)}_{\mathbf{W}'}\big)h^{(2)}_{\mathbf{C}} + h^{(1)}_{\mathbf{W}'}\big(h^{(2)}_{\mathbf{C}} - h^{(2)}_{\mathbf{C}'}\big).$$

Using $(u+v)^2 \le 2u^2 + 2v^2$, we obtain

$$\sum_{i=1}^n \big(h_{\mathbf{W},\mathbf{C}}(z_i) - h_{\mathbf{W}',\mathbf{C}'}(z_i)\big)^2 \\ \le 2\sum_{i=1}^n \big(h^{(1)}_{\mathbf{W}}(z_i) - h^{(1)}_{\mathbf{W}'}(z_i)\big)^2 h^{(2)}_{\mathbf{C}}(x_i)^2 + 2\sum_{i=1}^n h^{(1)}_{\mathbf{W}'}(z_i)^2\big(h^{(2)}_{\mathbf{C}}(x_i) - h^{(2)}_{\mathbf{C}'}(x_i)\big)^2. \tag{C.12}$$

Since

$$\|a\phi(x_i)^\top\|_* = \|a\|_2\|\phi(x_i)\|_2 \le \|a\|_2 B_x,$$

duality implies

$$|h^{(2)}_{\mathbf{C}}(x_i)| = |\langle a\phi(x_i)^\top, \mathbf{C}\rangle_F| \le \|a\phi(x_i)^\top\|_*\|\mathbf{C}\|_{\mathrm{op}} \le \|a\|_2 B_x. \tag{C.13}$$

Using the definition of $b_{\mathcal{W}}$ and dividing (C.12) by n, we get

$$d_S(h_{\mathbf{W},\mathbf{C}}, h_{\mathbf{W}',\mathbf{C}'}) \le \sqrt{2}\,\|a\|_2 B_x\, d_S\big(h^{(1)}_{\mathbf{W}}, h^{(1)}_{\mathbf{W}'}\big) + \sqrt{2}\, b_{\mathcal{W}}\, d_{S,\infty}\big(h^{(2)}_{\mathbf{C}}, h^{(2)}_{\mathbf{C}'}\big), \tag{C.14}$$

where

$$d_S(q, q') = \Big(\frac{1}{n}\sum_{i=1}^n (q(z_i) - q'(z_i))^2\Big)^{1/2}$$

and

$$d_{S,\infty}(q, q') = \max_{i\in[n]} |q(z_i) - q'(z_i)|.$$

Step 2: Covering-number decomposition. Let

$$\mathcal{H} := \{h_{\mathbf{W},\mathbf{C}} : \mathbf{W} \in \mathcal{W},\ \|\mathbf{C}\|_{\mathrm{op}} \le 1\},$$

and define

$$\mathcal{H}_1 := \Big\{h^{(1)}_{\mathbf{W}} : \mathbf{W} \in \mathcal{W}\Big\}, \qquad \mathcal{H}_2 := \Big\{h^{(2)}_{\mathbf{C}} : \|\mathbf{C}\|_{\mathrm{op}} \le 1\Big\}.$$

It follows from (C.14) that

$$\log \mathcal{N}(\epsilon, \mathcal{H}, d_S) \le \log \mathcal{N}\left(\frac{\epsilon}{2\sqrt{2}\,\|a\|_2 B_x}, \mathcal{H}_1, d_S\right) + \log \mathcal{N}\left(\frac{\epsilon}{2\sqrt{2}\, b_{\mathcal{W}}}, \mathcal{H}_2, d_{S,\infty}\right). \tag{C.15}$$

Step 3: Covering the derivative class. By Lemma 26, for every $\eta > 0$,

$$\log^{1/2} \mathcal{N}(\eta, \mathcal{H}_1, d_S) \le \frac{1}{\sqrt{n}\,\eta}\mathbb{E}_g\Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i \ell'\big(y_i, a^\top \mathbf{W}\phi(x_i)\big)\Big]. \tag{C.16}$$

Since $t \mapsto \ell'(y,t)$ is L_ℓ-Lipschitz, Lemma 27 yields

$$\mathbb{E}_g\Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i \ell'\big(y_i, a^\top \mathbf{W}\phi(x_i)\big)\Big] \le L_\ell \mathbb{E}_g\Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i)\Big]. \tag{C.17}$$

Taking $\eta = \frac{\epsilon}{2\sqrt{2}\,\|a\|_2 B_x}$ in (C.16) gives

$$\log^{1/2} \mathcal{N}\Big(\frac{\epsilon}{2\sqrt{2}\,\|a\|_2 B_x}, \mathcal{H}_1, d_S\Big) \le \frac{2\sqrt{2}\,\|a\|_2 B_x L_\ell}{\sqrt{n}\,\epsilon}\mathbb{E}_g\Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i)\Big]. \tag{C.18}$$

Step 4: Dimension-free covering of the dual class.

Recall that

$$\mathcal{H}_2 = \left\{ x \mapsto a^\top \mathbf{C}\phi(x) : \|\mathbf{C}\|_{\mathrm{op}} \le 1 \right\}.$$

We first observe that

$$\mathcal{H}_2 = \mathcal{F}_{a,1,r}. \tag{C.19}$$

Applying Lemma 25 with $R = 1$ and $B_\phi = B_x$, we obtain, for every $\eta > 0$,

$$\log \mathcal{N}(\eta, \mathcal{H}_2, d_{S,\infty}) \le \frac{36\|a\|_2^2 B_x^2}{\eta^2} \log_2 \Big(6n + \frac{8\|a\|_2 B_x}{\eta}\Big). \tag{C.20}$$

Taking $\eta = \frac{\epsilon}{2\sqrt{2}\, b_{\mathcal{W}}}$ yields

$$\log^{1/2} \mathcal{N}\Big(\frac{\epsilon}{2\sqrt{2}\, b_{\mathcal{W}}}, \mathcal{H}_2, d_{S,\infty}\Big) \le \frac{12\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\epsilon} \log_2^{1/2} \Big(6n + \frac{16\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\epsilon}\Big). \tag{C.21}$$

Step 5: Combined entropy bound. Using $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$, and combining (C.15), (C.18), and (C.21), we obtain

$$\begin{aligned}\log^{1/2} \mathcal{N}(\epsilon, \mathcal{H}, d_S) \le{}& \frac{2\sqrt{2}\,\|a\|_2 B_x L_\ell}{\sqrt{n}\,\epsilon} \mathbb{E}_g \left[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i) \right] \\ &+ \frac{12\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\epsilon} \log_2^{1/2} \left(6n + \frac{16\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\epsilon} \right). \end{aligned} \tag{C.22}$$

Step 6: Dudley's entropy integral. Since $\ell(y,\cdot)$ is G_ℓ-Lipschitz, we know $|\ell'(y,t)| \le G_\ell$. Together with (C.13), this gives $|h_{\mathbf{W},\mathbf{C}}(z)| \le G_\ell \|a\|_2 B_x$. Applying Lemma 13 with $\alpha = \frac{3}{\sqrt{n}}$ gives

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\substack{\mathbf{W}\in\mathcal{W} \\ \|\mathbf{C}\|_{\mathrm{op}}\le 1}} \sum_{i=1}^n \epsilon_i h_{\mathbf{W},\mathbf{C}}(z_i) \le \frac{12}{\sqrt{n}} + \frac{12}{\sqrt{n}} \int_{3/\sqrt{n}}^{G_\ell \|a\|_2 B_x} \log^{1/2} \mathcal{N}(\epsilon, \mathcal{H}, d_S)\, d\epsilon. \tag{C.23}$$

Substituting (C.22) into (C.23), and using $b_{\mathcal{W}} \le G_\ell$, $\epsilon \ge \frac{3}{\sqrt{n}}$, we obtain

$$6n + \frac{16\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\epsilon} \le 6n + \frac{16\sqrt{2}}{3} \|a\|_2 B_x G_\ell \sqrt{n} \le 2\left(6n + 8\|a\|_2 B_x G_\ell \sqrt{n}\right).$$

Furthermore, we know

$$\int_{3/\sqrt{n}}^{G_\ell \|a\|_2 B_x} \frac{d\epsilon}{\epsilon} = \log\left(\frac{G_\ell \|a\|_2 B_x \sqrt{n}}{3}\right).$$

It follows that

$$\begin{aligned}\frac{1}{n}\mathbb{E}_\epsilon \sup_{\substack{\mathbf{W}\in\mathcal{W} \\ \|\mathbf{C}\|_{\mathrm{op}}\le 1}} \sum_{i=1}^n \epsilon_i h_{\mathbf{W},\mathbf{C}}(z_i) \le{}& \frac{12}{\sqrt{n}} + \Bigg(\frac{24\sqrt{2}\,\|a\|_2 B_x L_\ell}{n} \mathbb{E}_g \Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i) \Big] \\ &+ \frac{144\sqrt{2}\,\|a\|_2 B_x b_{\mathcal{W}}}{\sqrt{n}} \log_2^{1/2}\left(2\left(6n + 8\|a\|_2 B_x G_\ell \sqrt{n}\right)\right) \Bigg) \log\Big(\frac{G_\ell \|a\|_2 B_x \sqrt{n}}{3}\Big). \end{aligned}$$

Combining this estimate with (C.10) proves (C.5). Finally, since $\mathcal{W} \subseteq \mathcal{B}_{R,r}$,

$$\|\mathbf{W}^\top a\|_2 \le \|\mathbf{W}\|_{\mathrm{op}} \|a\|_2 \le R\|a\|_2.$$

Therefore,

$$\begin{aligned}\mathbb{E}_g \Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i) \Big] &\le R\|a\|_2\, \mathbb{E}_g \Big\| \sum_{i=1}^n g_i \phi(x_i) \Big\|_2 \le R\|a\|_2 \Big(\sum_{i=1}^n \|\phi(x_i)\|_2^2 \Big)^{1/2} \\ &\le R\|a\|_2 B_x \sqrt{n}. \end{aligned}$$

This proves (C.6). Substituting it into (C.5) gives (C.7). □

The following two lemmas can be found in Bartlett et al. (2005). Lemma 29 controls the uniform deviation of empirical averages from their expectations, while Lemma 30 gives a control on the local Gaussian complexities.

Lemma 29 (Bartlett et al. (2005)). *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ into $[-b, b]$. Assume there is some $r > 0$ such that $\mathrm{Var}[f(Z)] \le r$ for any $f \in \mathcal{F}$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have*

$$\sup_{f\in\mathcal{F}} \Big(\mathbb{E}_Z[f(Z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\Big) \le 4\mathbb{E}[\mathfrak{R}_S(\mathcal{F})] + \Big(\frac{2r\log(2/\delta)}{n}\Big)^{\frac{1}{2}} + \frac{8b\log(2/\delta)}{3n}.$$

Lemma 30 (Bartlett et al. (2005)). *Let $V(\mathbf{w}) = \mathbb{E}_Z[\langle \mathbf{w}, \phi(x)\rangle^2]$. Let $(\lambda_i)_i$ be the eigenvalue of the operator $\mathbf{v} \mapsto \mathbb{E}_X[\langle \mathbf{v}, \phi(X)\rangle \phi(X)]$ arranged in a nonincreasing order. Then*

$$\frac{1}{n}\,\mathbb{E}\Big[\sup_{\mathbf{w}:\, V(\mathbf{w})\le r,\, \|\mathbf{w}\|\le R} \sum_{i=1}^n g_i \langle \mathbf{w}, \phi(x_i)\rangle\Big] \le \Big(\frac{2}{n}\min_{h\in\mathbb{N}}\Big(rh + R^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big)^{\frac{1}{2}}, \tag{C.24}$$

where g_i are standard Gaussian random variables.

Reducing $x \to a^\top \mathbf{W}\phi(x)$ to a linear class via $v = \mathbf{W}^\top a$ yields the following scalarized form of Lemma 30.

Lemma 31 (Localized Gaussian complexity for linear networks). *For every $\rho \ge 0$,*

$$\frac{1}{n}\mathbb{E}_{S,g}\Big[\sup_{\substack{\mathbf{W}\in\mathcal{W}\\ V(\mathbf{W})\le\rho}} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i)\Big] \le \Big[\frac{2}{n}\inf_{h\in\mathbb{N}}\Big(\rho h + R^2\|a\|_2^2 \sum_{j=h+1}^{\infty}\lambda_j\Big)\Big]^{1/2}. \tag{C.25}$$

Proof. For every $\mathbf{W} \in \mathcal{W}$, set

$$v_{\mathbf{W}} := \mathbf{W}^\top a.$$

Then

$$a^\top \mathbf{W}\phi(x) = v_{\mathbf{W}}^\top \phi(x), \qquad \|v_{\mathbf{W}}\|_2 \le R\|a\|_2,$$

and

$$V(\mathbf{W}) = \mathbb{E}_X\left[\big(v_{\mathbf{W}}^\top \phi(X)\big)^2\right].$$

Consequently,

$$\begin{aligned}&\big\{x \mapsto a^\top \mathbf{W}\phi(x) : \mathbf{W}\in\mathcal{W},\ V(\mathbf{W}) \le \rho\big\}\\ &\quad\subseteq \big\{x \mapsto v^\top\phi(x) : \|v\|_2 \le R\|a\|_2,\ \mathbb{E}_X[(v^\top\phi(X))^2] \le \rho\big\}.\end{aligned}$$

Applying Lemma 30 for linear function classes gives (C.25). □

The following lemma is standard and we omit the proof for brevity.

Lemma 32. *Let $a, b \ge 0$. If $x^2 \le ax + b$, then $x^2 \le a^2 + 2b$. On the other hand, if $x = a + \sqrt{b}$ for $a, b \ge 0$, then $ax + b \le x^2$.*

Theorem 33 (Localized dimension-independent gradient bound). *Assume that, for every y, the map $t \mapsto \ell(y,t)$ is L_ℓ-smooth and G_ℓ-Lipschitz, and that $\sup_x \|\phi(x)\|_2 \le B_x$. Define $B_\ell := (\mathbb{E}_Y[\ell'(Y,0)^2])^{1/2}$. Then, for every $\rho \ge 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\begin{aligned}\sup_{\substack{\mathbf{W}\in\mathcal{W}\\ V(\mathbf{W})\le\rho}} \|\nabla F(\mathbf{W}) - \nabla F_S(\mathbf{W})\|_* &= \widetilde{\mathcal{O}}\Bigg(\|a\|_2 B_x L_\ell\Big[\frac{1}{n}\inf_{h\in\mathbb{N}}\Big(\rho h + R^2\|a\|_2^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big]^{1/2}\\ &+ \frac{L_\ell R\|a\|_2^2 B_x^2}{n} + \|a\|_2 B_x\big(B_\ell + L_\ell\sqrt{\rho}\big)\Big(\frac{\log(2/\delta)}{n}\Big)^{1/2} + \frac{G_\ell\|a\|_2 B_x\log(2/\delta)}{n} + \frac{1}{\sqrt{n}} + \frac{\|a\|_2 B_x}{n^{3/4}}\Bigg),\end{aligned} \tag{C.26}$$

where $\widetilde{\mathcal{O}}$ hides logarithmic factors except those depending on δ.

Proof. For convenience, denote

$$\mathcal{W}_\rho := \{\mathbf{W} \in \mathcal{W} : V(\mathbf{W}) \le \rho\}.$$

For every $\mathbf{W} \in \mathcal{W}_\rho$ and $\mathbf{C} \in \mathbb{R}^{d_1 \times d_2}$ satisfying $\|\mathbf{C}\|_{\mathrm{op}} \le 1$, define

$$h_{\mathbf{W},\mathbf{C}}(z) := \langle \nabla_{\mathbf{W}} f(\mathbf{W}; z), \mathbf{C} \rangle_F.$$

By (C.8),

$$h_{\mathbf{W},\mathbf{C}}(z) = \ell'\big(y, a^\top \mathbf{W}\phi(x)\big) a^\top \mathbf{C}\phi(x).$$

We first estimate the variance of $h_{\mathbf{W},\mathbf{C}}$. Since

$$|a^\top \mathbf{C}\phi(x)| \le \|a\|_2 \|\mathbf{C}\|_{\mathrm{op}} \|\phi(x)\|_2 \le \|a\|_2 B_x,$$

we have

$$\mathrm{Var}_Z[h_{\mathbf{W},\mathbf{C}}(Z)] \le \mathbb{E}_Z[h_{\mathbf{W},\mathbf{C}}(Z)^2] \le \|a\|_2^2 B_x^2 \mathbb{E}_Z[\ell'\big(Y, a^\top \mathbf{W}\phi(X)\big)^2]. \tag{C.27}$$

The L_ℓ-smoothness of ℓ implies

$$|\ell'(y,t) - \ell'(y,0)| \le L_\ell |t|.$$

Using $(u+v)^2 \le 2u^2 + 2v^2$, we obtain

$$\begin{aligned}\mathbb{E}_Z\left[\ell'\big(Y, a^\top \mathbf{W}\phi(X)\big)^2\right] &\le 2\mathbb{E}_Y[\ell'(Y,0)^2] + 2L_\ell^2 \mathbb{E}_X\left[\big(a^\top \mathbf{W}\phi(X)\big)^2\right] \\ &\le 2\big(B_\ell^2 + L_\ell^2 \rho\big).\end{aligned} \tag{C.28}$$

Combining (C.27) and (C.28) gives

$$\mathrm{Var}_Z[h_{\mathbf{W},\mathbf{C}}(Z)] \le 2\|a\|_2^2 B_x^2 \big(B_\ell^2 + L_\ell^2 \rho\big). \tag{C.29}$$

Furthermore,

$$|h_{\mathbf{W},\mathbf{C}}(z)| \le G_\ell \|a\|_2 B_x. \tag{C.30}$$

Applying Lemma 29 to the class

$$\mathcal{H}_\rho := \{z \mapsto h_{\mathbf{W},\mathbf{C}}(z) : \mathbf{W} \in \mathcal{W}_\rho,\ \|\mathbf{C}\|_{\mathrm{op}} \le 1\},$$

we find that, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{\mathbf{W}\in\mathcal{W}_\rho} \|\nabla F(\mathbf{W}) - \nabla F_S(\mathbf{W})\|_* = \sup_{\substack{\mathbf{W}\in\mathcal{W}_\rho \\ \|\mathbf{C}\|_{\mathrm{op}}\le 1}} \langle \nabla F(\mathbf{W}) - \nabla F_S(\mathbf{W}), \mathbf{C}\rangle \\ &\le \frac{4}{n}\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho} \Big\| \sum_{i=1}^n \epsilon_i \nabla_{\mathbf{W}} f(\mathbf{W}; z_i)\Big\|_*\Big] + 2\|a\|_2 B_x (B_\ell + L_\ell\sqrt{\rho})\Big(\frac{\log(2/\delta)}{n}\Big)^{1/2} \\ &+ \frac{8G_\ell \|a\|_2 B_x \log(2/\delta)}{3n}.\end{aligned} \tag{C.31}$$

It remains to estimate the expected Rademacher complexity. Define

$$b_\rho := \sup_{\mathbf{W}\in\mathcal{W}_\rho} \left(\frac{1}{n}\sum_{i=1}^n \ell'\big(y_i, a^\top \mathbf{W}\phi(x_i)\big)^2\right)^{1/2}. \tag{C.32}$$

Applying Lemma 28 to $\mathcal{W}_\rho$, and absorbing its logarithmic factors into $\widetilde{\mathcal{O}}$, gives

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho} \Big\| \sum_{i=1}^n \epsilon_i \nabla_{\mathbf{W}} f(\mathbf{W}; z_i)\Big\|_*\Big] \\ &\qquad = \widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{n}} + \frac{\|a\|_2 B_x L_\ell}{n}\mathbb{E}_{S,g}\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho} \sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i)\Big] + \frac{\|a\|_2 B_x}{\sqrt{n}}\mathbb{E}_S[b_\rho]\Big).\end{aligned} \tag{C.33}$$

By Lemma 31,

$$\frac{\|a\|_2 B_x L_\ell}{n}\mathbb{E}_{S,g}\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho}\sum_{i=1}^n g_i a^\top \mathbf{W}\phi(x_i)\Big] \le \sqrt{2}\|a\|_2 B_x L_\ell\left[\frac{1}{n}\inf_{h\in\mathbb{N}}\Big(\rho h + R^2\|a\|_2^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\right]^{1/2}. \tag{C.34}$$

We next estimate $\mathbb{E}_S[b_\rho]$. By symmetrization in Bartlett et al. (2005),

$$\mathbb{E}_S[b_\rho^2] \le \sup_{\mathbf{W}\in\mathcal{W}_\rho}\mathbb{E}_Z[\ell'(Y, a^\top\mathbf{W}\phi(X))^2] + 2\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho}\frac{1}{n}\sum_{i=1}^n \epsilon_i \ell'\big(y_i, a^\top\mathbf{W}\phi(x_i)\big)^2\Big]. \tag{C.35}$$

The first term in Eq. (C.35) is bounded by $2(B_\ell^2 + L_\ell^2\rho)$ according to (C.28). For the second term in Eq. (C.35), define $q_{\mathbf{W}}(z) := \ell'\big(y, a^\top\mathbf{W}\phi(x)\big)^2$. For $\mathbf{W}, \mathbf{W}' \in \mathcal{W}_\rho$, we have

$$\begin{aligned} d_S(q_{\mathbf{W}}, q_{\mathbf{W}'})^2 &= \frac{1}{n}\sum_{i=1}^n\Big[\ell'\big(y_i, a^\top\mathbf{W}\phi(x_i)\big)^2 - \ell'\big(y_i, a^\top\mathbf{W}'\phi(x_i)\big)^2\Big]^2 \\ &\le \frac{1}{n}\sum_{i=1}^n\Big[\big|\ell'\big(y_i, a^\top\mathbf{W}\phi(x_i)\big)\big| + \big|\ell'\big(y_i, a^\top\mathbf{W}'\phi(x_i)\big)\big|\Big]^2 \\ &\qquad\times\Big[\ell'\big(y_i, a^\top\mathbf{W}\phi(x_i)\big) - \ell'\big(y_i, a^\top\mathbf{W}'\phi(x_i)\big)\Big]^2 \\ &\le 4b_\rho^2 L_\ell^2 \max_{i\in[n]}\big|a^\top(\mathbf{W}-\mathbf{W}')\phi(x_i)\big|^2. \end{aligned}$$

Therefore,

$$d_S(q_{\mathbf{W}}, q_{\mathbf{W}'}) \le 2b_\rho L_\ell d_{S,\infty}\big(a^\top\mathbf{W}\phi(\cdot), a^\top\mathbf{W}'\phi(\cdot)\big). \tag{C.36}$$

Applying Lemma 25 and (C.36), we obtain

$$\log\mathcal{N}\left(\epsilon, \{q_{\mathbf{W}} : \mathbf{W}\in\mathcal{W}_\rho\}, d_S\right) \le \frac{144 b_\rho^2 L_\ell^2 R^2\|a\|_2^2 B_x^2}{\epsilon^2}\log_2\left(6n + \frac{16 b_\rho L_\ell R\|a\|_2 B_x}{\epsilon}\right). \tag{C.37}$$

Since $0 \le q_{\mathbf{W}}(z) \le G_\ell^2$, Lemma 13 with $\alpha = 3/\sqrt{n}$ gives

$$\begin{aligned} &\mathbb{E}_\epsilon\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho}\frac{1}{n}\sum_{i=1}^n\epsilon_i q_{\mathbf{W}}(z_i)\Big] \\ &\quad\le \frac{12}{\sqrt{n}} + \frac{144 b_\rho L_\ell R\|a\|_2 B_x}{\sqrt{n}}\log_2^{1/2}(6n + 6G_\ell L_\ell R\|a\|_2 B_x\sqrt{n})\log\Big(\frac{G_\ell^2\sqrt{n}}{3}\Big), \end{aligned} \tag{C.38}$$

where we use $b_\rho \le G_\ell$.

Combining (C.35), (C.28), and (C.38), and then taking expectation over S, gives

$$\mathbb{E}_S[b_\rho^2] \le 2(B_\ell^2 + L_\ell^2\rho) + \frac{24}{\sqrt{n}} + \widetilde{\mathcal{O}}\Big(\frac{L_\ell R\|a\|_2 B_x}{\sqrt{n}}\mathbb{E}_S[b_\rho]\Big). \tag{C.39}$$

Since

$$\mathbb{E}_S[b_\rho] \le \big(\mathbb{E}_S[b_\rho^2]\big)^{1/2},$$

(C.39) is a quadratic inequality in $\big(\mathbb{E}_S[b_\rho^2]\big)^{1/2}$. Applying Lemma 32, we obtain

$$\mathbb{E}_S[b_\rho^2] = \widetilde{\mathcal{O}}\Big(B_\ell^2 + L_\ell^2\rho + \frac{1}{\sqrt{n}} + \frac{L_\ell^2 R^2\|a\|_2^2 B_x^2}{n}\Big), \tag{C.40}$$

and

$$\mathbb{E}_S[b_\rho] = \widetilde{\mathcal{O}}\Big(B_\ell + L_\ell\sqrt{\rho} + \frac{1}{n^{1/4}} + \frac{L_\ell R\|a\|_2 B_x}{\sqrt{n}}\Big). \tag{C.41}$$

Substituting (C.34) and (C.41) into (C.33), we obtain

$$\frac{1}{n}\mathbb{E}_{S,\epsilon}\Big[\sup_{\mathbf{W}\in\mathcal{W}_\rho}\Big\|\sum_{i=1}^{n}\epsilon_i\nabla_{\mathbf{W}}f(\mathbf{W};z_i)\Big\|_*\Big]=\widetilde{\mathcal{O}}\Bigg(\|a\|_2B_xL_\ell\Big[\frac{1}{n}\inf_{h\in\mathbb{N}}\Big(\rho h+R^2\|a\|_2^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big]^{1/2}$$
$$+\frac{1}{\sqrt{n}}+\frac{\|a\|_2B_x(B_\ell+L_\ell\sqrt{\rho})}{\sqrt{n}}+\frac{\|a\|_2B_x}{n^{3/4}}+\frac{L_\ell R\|a\|_2^2B_x^2}{n}\Bigg). \quad \text{(C.42)}$$

Finally, substituting (C.42) into (C.31) proves (C.26). Under the additional nondegeneracy and envelope assumptions stated in the theorem, the lower-order terms can be absorbed into the remaining terms, completing the proof. □

To conduct a localization analysis, we require the following uniform localized convergence argument based on the peeling trick.

Lemma 34 (Uniform localized convergence argument (Xu & Zeevi, 2025))**.** *For a function class $\mathcal{F}=\{f_{\mathbf{W}}:\mathbf{W}\in\mathcal{W}\}$ and the functional $\widetilde{V}:\mathcal{W}\mapsto[0,\widetilde{R}]$, assume there is a function $\psi(s;\delta)$, which is non-decreasing w.r.t. r and satisfies that $\forall\delta\in(0,1)$, $\forall s\in[0,\widetilde{R}]$, with probability at least $1-\delta$*

$$\sup_{\mathbf{W}\in\mathcal{W}:\widetilde{V}(\mathbf{W})\le r}\Big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_{\mathbf{W}}(z_i)-\mathbb{E}_Z[\nabla f_{\mathbf{W}}(Z)]\Big\|\le\psi(s;\delta).$$

Then, given any $\delta\in(0,1)$ and $s_0\in(0,\widetilde{R}]$, with probability at least $1-\delta$, for all $\mathbf{W}\in\mathcal{W}$

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_{\mathbf{W}}(z_i)-\mathbb{E}_Z[\nabla f_{\mathbf{W}}(Z)]\Big\|\le\psi\big(2\widetilde{V}(\mathbf{W})\vee s_0;\,\delta/\log_2(2\widetilde{R}/s_0)\big). \quad \text{(C.43)}$$

We now give the proof of Theorem 7 using Theorem 33 and Lemma 34.

Proof of Theorem 7. We define

$$\psi(s,\delta)=\widetilde{C}\Bigg(\|a\|_2B_xL_\ell\Big[\frac{1}{n}\inf_{h\in\mathbb{N}}\Big(sh+R^2\|a\|_2^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big]^{1/2}+\frac{L_\ell R\|a\|_2^2B_x^2\log(1/\delta)}{n}$$
$$+\|a\|_2B_x\big(B_\ell+L_\ell\sqrt{s}\big)\Big(\frac{\log(1/\delta)}{n}\Big)^{1/2}+\frac{G_\ell\|a\|_2B_x\log(1/\delta)}{n}+\frac{1}{\sqrt{n}}+\frac{\|a\|_2B_x}{n^{3/4}}\Bigg),$$

where $\widetilde{C}$ is a constant which only has a logarithmic dependency on R, n and other parameters. Then, by Theorem 33, ψ satisfies the condition in Eq. (C.43) with $\widetilde{V}=V$. Therefore, we can apply Lemma 34 with $\widetilde{V}=V$, $\widetilde{R}=\sup_{\mathbf{W}\in\mathcal{W}}V(\mathbf{W})$ and $s_0=1/n$ to derive the following inequality uniformly for all $\mathbf{W}\in\mathcal{W}$

$$\|\nabla F(\mathbf{W})-\nabla F_S(\mathbf{W})\|_*\le\widetilde{C}\Bigg(\|a\|_2B_xL_\ell\Big[\frac{1}{n}\inf_{h\in\mathbb{N}}\Big(2\big(V(\mathbf{W})\vee1/n\big)h+R^2\|a\|_2^2\sum_{j=h+1}^{\infty}\lambda_j\Big)\Big]^{1/2}$$
$$+\frac{L_\ell R\|a\|_2^2B_x^2\log\big(\log_2(2\widetilde{R}n)/\delta\big)}{n}+\|a\|_2B_x\Big(B_\ell+L_\ell\sqrt{2\big(V(\mathbf{W})\vee1/n\big)}\Big)\Big(\frac{\log\big(\log_2(2\widetilde{R}n)/\delta\big)}{n}\Big)^{1/2}$$
$$+\frac{G_\ell\|a\|_2B_x\log\big(\log_2(2\widetilde{R}n)/\delta\big)}{n}+\frac{1}{\sqrt{n}}+\frac{\|a\|_2B_x}{n^{3/4}}\Bigg).$$

The lower-order terms can be absorbed into the remaining terms, completing the proof. □

We now apply Theorem 7 to the Muon iterates, which gives the dimension-independent high-probability bound on the generalization gap of Muon for linear networks.

Proof of Corollary 8. By Lemma 2, with the constant learning rate $\eta_t = \eta$, the operator norm of the iterates $\mathbf{W}_t$ for any $t \le T$ is bounded by

$$\|\mathbf{W}_t\|_\sigma \le \|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{t-1} \eta \le \|\mathbf{W}_0\|_\sigma + T\eta.$$

This implies that for all $0 \le t \le T$, the iterates $\mathbf{W}_t$ reside within a parameter ball $\mathcal{B}_{R,r}$ with radius $R = \|\mathbf{W}_0\|_\sigma + T\eta$.
By Theorem 7, the uniform generalization gap bound holds simultaneously for all $\mathbf{W} \in \mathcal{B}_{R,r}$ with probability at least $1-\delta$. Substituting $R_T = \|\mathbf{W}_0\|_\sigma + T\eta$ into the bound from Theorem 7 directly completes the proof. □

We now specialize the optimization errors established in Theorem 4 and Corollary 5 to the linear networks (4.4), which gives the dimension-independent high-probability optimization error.

Proof of Theorem 9. We calculate separately all model-dependent parameters appearing in Theorem 4.
Gradient formula and uniform gradient bound. For $f(\mathbf{W}; z) = \ell\big(y, a^\top \mathbf{W}\phi(x)\big)$, direct differentiation gives

$$\nabla f(\mathbf{W}; z) = \ell'\big(y, a^\top \mathbf{W}\phi(x)\big) a\phi(x)^\top. \tag{C.44}$$

Since $t \mapsto \ell(y,t)$ is G_ℓ-Lipschitz, $|\ell'(y,t)| \le G_\ell$. Together with $\|\phi(x)\|_2 \le B_x$, this yields

$$\|\nabla f(\mathbf{W}; z)\|_{\mathrm{F}} = \big|\ell'\big(y, a^\top \mathbf{W}\phi(x)\big)\big| \|a\|_2 \|\phi(x)\|_2 \le G_\ell \|a\|_2 B_x. \tag{C.45}$$

Computation of L_*. For any $\mathbf{W}, \mathbf{W}'$,

$$\begin{aligned}
&\|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z)\|_* \\
&= \big|\ell'\big(y, a^\top \mathbf{W}\phi(x)\big) - \ell'\big(y, a^\top \mathbf{W}'\phi(x)\big)\big| \|a\|_2 \|\phi(x)\|_2 \\
&\le L_\ell \big|a^\top (\mathbf{W} - \mathbf{W}')\phi(x)\big| \|a\|_2 \|\phi(x)\|_2 \le L_\ell \|a\|_2^2 B_x^2 \|\mathbf{W} - \mathbf{W}'\|_{\mathrm{op}}.
\end{aligned}$$

Thus, we can take $L_* = L_\ell \|a\|_2^2 B_x^2$.
Computation of b_*. Every sample gradient in Eq. (C.44) is an outer product and hence has rank at most one. Its nuclear norm therefore equals its Frobenius norm. Using $|\ell'(y,0)| \le G_\ell$ and $\|\phi(x)\|_2 \le B_x$, we obtain

$$b_* = \sup_z \|\nabla f(\mathbf{0}; z)\|_* = \sup_{(x,y)} |\ell'(y,0)| \|a\|_2 \|\phi(x)\|_2 \le G_\ell \|a\|_2 B_x.$$

Computation of σ. Since

$$\begin{aligned}
\mathbb{E}_z \left[\|\nabla f(\mathbf{W}; z) - \nabla F_S(\mathbf{W})\|_{\mathrm{F}}^2\right] &= \mathbb{E}_z \|\nabla f(\mathbf{W}; z)\|_{\mathrm{F}}^2 - \|\nabla F_S(\mathbf{W})\|_{\mathrm{F}}^2 \\
&\le \mathbb{E}_z \|\nabla f(\mathbf{W}; z)\|_{\mathrm{F}}^2 \le G_\ell^2 \|a\|_2^2 B_x^2,
\end{aligned}$$

the variance parameter can be taken as $\sigma = G_\ell \|a\|_2 B_x$.
Computation of G_F. By Eq. (C.45), $\|\nabla F_S(\mathbf{W})\|_{\mathrm{F}} \le G_\ell \|a\|_2 B_x$, then by the triangle inequality,

$$\|\nabla f(\mathbf{W}; z) - \nabla F_S(\mathbf{W})\|_{\mathrm{F}} \le \|\nabla f(\mathbf{W}; z)\|_{\mathrm{F}} + \|\nabla F_S(\mathbf{W})\|_{\mathrm{F}} \le 2G_\ell \|a\|_2 B_x.$$

Thus, we can take $G_F = 2G_\ell \|a\|_2 B_x$.
Computation of q. Define

$$\mathbf{v}_S(\mathbf{W}) := \frac{1}{n} \sum_{i=1}^n \ell'\big(y_i, a^\top \mathbf{W}\phi(x_i)\big)\phi(x_i).$$

Then $\nabla F_S(\mathbf{W}) = a\mathbf{v}_S(\mathbf{W})^\top$. Likewise, every mini-batch gradient can be written as $\mathbf{G}_t = a\mathbf{g}_t^\top$ for some random vector $\mathbf{g}_t \in \mathbb{R}^{d_2}$. Since the momentum recursion is linear and $\mathbf{M}_0 = \mathbf{G}_0$, there exists $\mathbf{m}_t \in \mathbb{R}^{d_2}$ such that $\mathbf{M}_t = a\mathbf{m}_t^\top$. Similarly, the deterministic momentum defined by $\mathbf{C}_0 = \nabla F_S(\mathbf{W}_0)$, $\mathbf{C}_t = \beta\mathbf{C}_{t-1} + (1-\beta)\nabla F_S(\mathbf{W}_t)$ can be written as $\mathbf{C}_t = a\mathbf{c}_t^\top$ for some $\mathbf{c}_t \in \mathbb{R}^{d_2}$.

Consequently, the matrices $\mathbf{M}_t - \mathbf{C}_t$, $\nabla F_S(\mathbf{W}_t) - \mathbf{C}_t$, and $\nabla F_S(\mathbf{W}_t) - \mathbf{M}_t$ all have rank at most one. Therefore, we can take $q = 1$.

Substitution into the general result. We plug in all the model-dependent parameters

$$L_* = L_\ell \|a\|_2^2 B_x^2, \qquad b_* \le G_\ell \|a\|_2 B_x,$$
$$G_F = 2G_\ell \|a\|_2 B_x, \qquad \sigma = G_\ell \|a\|_2 B_x, \qquad q = 1.$$

into Theorem 4 and Corollary 5 to complete the proof. □

We now combine the generalization gap given in Corollary 8 and the optimization error given in Theorem 9 to get the following dimension-independent excess risk bound of Muon for linear networks (4.4).

Proof of Theorem 10. It follows from Corollary 8 that, with probability at least $1 - \delta/2$, simultaneously for every $t = 0, \dots, T-1$,

$$\|\nabla F(\mathbf{W}_t) - \nabla F_S(\mathbf{W}_t)\|_* = \widetilde{\mathcal{O}}\Bigg([1 + \|a\|_2 B_x (1 + L_\ell)] \sqrt{\frac{\mathfrak{C}_\phi(R_T, \nu(\mathbf{W}_t))}{n}} + \frac{(L_\ell R_T \|a\|_2^2 B_x^2 + G_\ell \|a\|_2 B_x) \log(2/\delta)}{n} \Bigg). \tag{C.46}$$

For any fixed $h \in \mathbb{N}$, the definition of $\mathfrak{C}_\phi$ and Jensen's inequality give

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \sqrt{\mathfrak{C}_\phi(R_T, \nu(\mathbf{W}_t))} &\le \frac{1}{T} \sum_{t=0}^{T-1} \left[1 + \nu(\mathbf{W}_t) h + R_T^2 \|a\|_2^2 \sum_{j>h} \lambda_j \right]^{1/2} \\ &\le \left[1 + \frac{h}{T} \sum_{t=0}^{T-1} \nu(\mathbf{W}_t) + R_T^2 \|a\|_2^2 \sum_{j>h} \lambda_j \right]^{1/2} \\ &= \left[1 + \nu_T h + R_T^2 \|a\|_2^2 \sum_{j>h} \lambda_j \right]^{1/2}. \end{aligned}$$

Taking the infimum over $h \in \mathbb{N}$ yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \sqrt{\mathfrak{C}_\phi(R_T, \nu(\mathbf{W}_t))} \le \sqrt{\mathfrak{C}_\phi(R_T, \nu_T)}.$$

Therefore, Eq. (C.46) implies

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\mathbf{W}_t) - \nabla F_S(\mathbf{W}_t)\|_* \le \widetilde{\mathcal{O}}(\mathfrak{G}_n(R_T, \nu_T, \delta)). \tag{C.47}$$

By Theorem 9 and Corollary 5, for the choices $\eta = \sqrt{\frac{(1-\beta)\Delta_S}{T L_{\mathrm{lin}}}}$, $1 - \beta = \min\{\frac{\sqrt{L_{\mathrm{lin}} \Delta_S}}{\Gamma_\delta \sqrt{T}}, 1\}$, with probability at least $1 - \delta/2$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F_S(\mathbf{W}_t)\|_* \le \widetilde{\mathcal{O}}\Bigg(\Big(\frac{L_{\mathrm{lin}} \Delta_S \Gamma_\delta^2}{T} \Big)^{1/4} + \frac{\Gamma_\delta^2}{\sqrt{L_{\mathrm{lin}} \Delta_S T}} \Bigg). \tag{C.48}$$

By a union bound, the events in Eqs. (C.47) and (C.48) hold simultaneously with probability at least $1 - \delta$. Combining them with Eq. (B.14), we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\mathbf{W}_t)\|_* \le \widetilde{\mathcal{O}}\Bigg(\Big(\frac{L_{\mathrm{lin}} \Delta_S \Gamma_\delta^2}{T} \Big)^{1/4} + \frac{\Gamma_\delta^2}{\sqrt{L_{\mathrm{lin}} \Delta_S T}} + \mathfrak{G}_n(R_T, \nu_T, \delta) \Bigg).$$

The proof is completed □

C.2 Proof of Dimension-independent Bounds for Shallow Networks

We require the following lemmas in our proofs.

Lemma 35 (Lipschitz composition of Rademacher complexity Golowich et al. (2018)). *Let $\mathcal{H}$ be a class of real-valued functions mapping into $[-R, R]$. For fixed $a \in \mathbb{R}$ and $L > 0$, define*

$$\mathcal{F}_{L,a} = \{f : [-R, R] \to \mathbb{R} : f(0) = a,\ |f(s) - f(t)| \le L|s - t|\}$$

and

$$\mathcal{F}_{L,a} \circ \mathcal{H} = \{x \mapsto f(h(x)) : f \in \mathcal{F}_{L,a},\ h \in \mathcal{H}\}.$$

Then there exists a universal constant $c > 0$ such that, for any sample $\tilde{S} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$,

$$\mathfrak{R}_{\tilde{S}}\big(\mathcal{F}_{L,a} \circ \mathcal{H}\big) \le cL\Big(\frac{R}{\sqrt{n}} + \log^{3/2}(n)\, \mathfrak{R}_{\tilde{S}}(\mathcal{H})\Big).$$

Lemma 36 (Rademacher complexity of SNNs (Vardi et al., 2022, Theorem 2)). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be applied element-wise. Assume that $\sigma(0) = 0$ and that σ is L_σ-Lipschitz. Let $u_1, \dots, u_n \in \mathbb{R}^d$ satisfy $\|u_i\|_2 \le B_x$. Then there exists a universal constant $c > 0$ such that*

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\|\mathbf{W}\|_F \le R_F} \sum_{i=1}^n \epsilon_i a^\top \sigma(\mathbf{W} u_i) \le \frac{cL_\sigma \|a\|_2 R_F B_x}{\sqrt{n}}\big(1 + \log^{3/2}(2n)\big). \tag{C.49}$$

Lemma 37 (Vector contraction for Rademacher complexities Maurer (2016)). *Let $\mathcal{X}$ be any set and let $(x_1, \dots, x_n) \in \mathcal{X}^n$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \ell_2$, and write $f_k(x)$ for the k-th coordinate of $f(x)$. Suppose that for each $i = 1, \dots, n$, the map $h_i : \ell_2 \to \mathbb{R}$ is L-Lipschitz with respect to the Euclidean norm. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i\, h_i\big(f(x_i)\big) \le \sqrt{2}\, L\, \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sum_{k \ge 1} \epsilon_{ik}\, f_k(x_i),$$

where $\{\epsilon_i\}$ and $\{\epsilon_{ik}\}$ are independent Rademacher variables.

Lemma 38 (Rademacher contraction principle Ledoux & Talagrand (1991)). *Let $T \subset \mathbb{R}^n$ be symmetric, and let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher variables. Suppose that $\varphi_i : \mathbb{R} \to \mathbb{R}$ satisfies $\varphi_i(0) = 0$ and is L-Lipschitz for every $i \in [n]$. Then*

$$\mathbb{E}_\epsilon \sup_{t \in T} \sum_{i=1}^n \epsilon_i \varphi_i(t_i) \le 2L\, \mathbb{E}_\epsilon \sup_{t \in T} \sum_{i=1}^n \epsilon_i t_i.$$

We first consider the function class

$$\mathcal{S} = \{x \mapsto s_{\mathbf{W}}(x) : \mathbf{W} \in \mathcal{W}\}$$

and establish the following upper bound on its empirical Gaussian complexity.

Lemma 39. *Let $g_1, \dots, g_n$ be standard Gaussian random variables. Assume that $\|\phi(x)\|_2 \le B_x$, that $\sigma(0) = 0$, and that σ is L_σ-Lipschitz. Then*

$$n\mathfrak{R}_S(\mathcal{S}) = \mathbb{E}_g \sup_{\mathbf{W} \in \mathcal{B}_{R,r}} \sum_{i=1}^n g_i a^\top \sigma(\mathbf{W}\phi(x_i)) \le cL_\sigma R\sqrt{r}\, \|a\|_2 B_x \sqrt{n}\, \Lambda_n, \tag{C.50}$$

where $c > 0$ is a constant and $\Lambda_n = \sqrt{\log(2n)}(1 + \log^{3/2}(2n))$.

Proof. Since g_i has the same distribution as $\epsilon_i |g_i|$, where ϵ_i is an independent Rademacher variable, we have

$$n\mathfrak{R}_S(\mathcal{S}) = \mathbb{E}_{|g|}\mathbb{E}_\epsilon \sup_{\mathbf{W} \in \mathcal{B}_{R,r}} \sum_{i=1}^n \epsilon_i |g_i| a^\top \sigma(\mathbf{W}\phi(x_i)).$$

Conditionally on $(|g_1|, \dots, |g_n|)$, define $\varphi_i(t) := |g_i| t$. Then $\varphi_i(0) = 0$ and each φ_i is $\max_{j \in [n]} |g_j|$-Lipschitz. Moreover, the set $\{(a^\top \sigma(\mathbf{W}\phi(x_1)), \dots, a^\top \sigma(\mathbf{W}\phi(x_n))) : \mathbf{W} \in \mathcal{B}_{R,r}\}$

is symmetric, since replacing a by $-a$ changes every coordinate's sign. Then applying Lemma 38 conditioned on $|g|$ gives

$$n\Re_S(\mathcal{S}) \le 2\mathbb{E}_{|g|}\big[\max_{1\le i\le n}|g_i|\big]\mathbb{E}_\epsilon \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \sum_{i=1}^n \epsilon_i a^\top \sigma(\mathbf{W}\phi(x_i))$$
$$\le 2\sqrt{2\log(2n)}\mathbb{E}_\epsilon \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \sum_{i=1}^n \epsilon_i a^\top \sigma(\mathbf{W}\phi(x_i)), \tag{C.51}$$

where the last step follows from the standard Gaussian maximal inequality $\mathbb{E}_g \max_{1\le i\le n}|g_i| \le \sqrt{2\log(2n)}$. Applying Lemma 36 with $b = \|a\|_2$, $B = R\sqrt{r}$, and with its inputs x_i replaced by $\phi(x_i)$. Its input condition holds since $\|\phi(x_i)\|_2 \le B_x$. Hence

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \sum_{i=1}^n \epsilon_i a^\top \sigma(\mathbf{W}\phi(x_i)) \le \frac{1}{n}\mathbb{E}_\epsilon \sup_{\|\mathbf{W}\|_F\le R\sqrt{r}} \sum_{i=1}^n \epsilon_i a^\top \sigma(\mathbf{W}\phi(x_i))$$
$$\le \frac{cL_\sigma R\sqrt{r}\,\|a\|_2 B_x}{\sqrt{n}}\big(1+\log^{3/2}(2n)\big). \tag{C.52}$$

Combining (C.51) and (C.52) completes the proof. □

We next bound $K_2(S)$, which captures the Gaussian complexity associated with the gradients of $s_{\mathbf{W}}$ with respect to $\mathbf{W}$.

Lemma 40. *Let $g_1,\dots,g_n$ be independent standard Gaussian random variables. Assume that $\|\phi(x)\|_2 \le B_x$, $\|\sigma'(\cdot)\|_\infty \le L_\sigma$, and that σ' is $L_{\sigma'}$-Lipschitz. Then*

$$K_2(S) = \mathbb{E}_g \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \Big\| \sum_{i=1}^n g_i\big(a\odot\sigma'(\mathbf{W}\phi(x_i))\big)\phi(x_i)^\top\Big\|_*$$
$$\le c\|a\|_2 B_x\sqrt{n}(L_\sigma + L_{\sigma'}R\sqrt{r}\,B_x\Lambda_n), \tag{C.53}$$

where $c>0$ is a constant and $\Lambda_n = \sqrt{\log(2n)}\big(1+\log^{3/2}(2n)\big)$.

Proof. Separating the value of σ' at the origin gives

$$K_2(S) \le \mathbb{E}_g\Big\|\sigma'(0)a\Big(\sum_{i=1}^n g_i\phi(x_i)\Big)^\top\Big\|_* + \mathbb{E}_g \sup_{\mathbf{W}\in\mathcal{B}_{R,r}} \Big\|\sum_{i=1}^n g_i\Big(a\odot\big(\sigma'(\mathbf{W}\phi(x_i))-\sigma'(0)\mathbf{1}\big)\Big)\phi(x_i)^\top\Big\|_*. \tag{C.54}$$

For the first term in Eq. (C.54), since the outer product has rank one,

$$\mathbb{E}_g\Big\|\sigma'(0)a\Big(\sum_{i=1}^n g_i\phi(x_i)\Big)^\top\Big\|_* = |\sigma'(0)|\|a\|_2\mathbb{E}_g\Big\|\sum_{i=1}^n g_i\phi(x_i)\Big\|_2$$
$$\le L_\sigma\|a\|_2\Big(\sum_{i=1}^n\|\phi(x_i)\|_2^2\Big)^{1/2} \le L_\sigma\|a\|_2 B_x\sqrt{n}. \tag{C.55}$$

We now control the second term in Eq. (C.54). Let $w_j^\top$ be the j-th row of $\mathbf{W}$. By the triangle inequality for the nuclear norm and the rank-one identity $\|uv^\top\|_* = \|u\|_2\|v\|_2$,

$$\Big\|\sum_{i=1}^n g_i\Big(a\odot\big(\sigma'(\mathbf{W}\phi(x_i))-\sigma'(0)\mathbf{1}\big)\Big)\phi(x_i)^\top\Big\|_*$$
$$\le \sum_{j=1}^{d_1}|a_j|\Big\|\sum_{i=1}^n g_i\big(\sigma'(w_j^\top\phi(x_i))-\sigma'(0)\big)\phi(x_i)\Big\|_2$$
$$\le \|a\|_2\left[\sum_{j=1}^{d_1}\Big\|\sum_{i=1}^n g_i\big(\sigma'(w_j^\top\phi(x_i))-\sigma'(0)\big)\phi(x_i)\Big\|_2^2\right]^{1/2}.$$

Since

$$a \odot \big(\sigma'(\mathbf{W}\phi(x_i)) - \sigma'(0)\mathbf{1}\big) = \sum_{j=1}^{d_1} a_j \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big) e_j,$$

where e_j denotes the j-th standard basis vector in $\mathbb{R}^{d_1}$. Then

$$\begin{aligned}
&\sum_{i=1}^{n} g_i \Big(a \odot \big(\sigma'(\mathbf{W}\phi(x_i)) - \sigma'(0)\mathbf{1}\big)\Big)\phi(x_i)^\top \\
&= \sum_{i=1}^{n} g_i \sum_{j=1}^{d_1} a_j \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big) e_j \phi(x_i)^\top \\
&= \sum_{j=1}^{d_1} e_j \Big[a_j \sum_{i=1}^{n} g_i \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big)\phi(x_i)\Big]^\top.
\end{aligned}$$

Applying the triangle inequality for the nuclear norm and the identity $\|uv^\top\|_* = \|u\|_2\|v\|_2$, we obtain

$$\begin{aligned}
&\Big\| \sum_{i=1}^{n} g_i \Big(a \odot \big(\sigma'(\mathbf{W}\phi(x_i)) - \sigma'(0)\mathbf{1}\big)\Big)\phi(x_i)^\top \Big\|_* \\
&= \Big\| \sum_{j=1}^{d_1} e_j \Big[a_j \sum_{i=1}^{n} g_i \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big)\phi(x_i)\Big]^\top \Big\|_* \\
&= \sum_{j=1}^{d_1} \|e_j\|_2 \, |a_j| \Big\| \sum_{i=1}^{n} g_i \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big)\phi(x_i) \Big\|_2 \\
&\le \|a\|_2 \left[\sum_{j=1}^{d_1} \Big\| \sum_{i=1}^{n} g_i \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big)\phi(x_i) \Big\|_2^2\right]^{1/2},
\end{aligned} \tag{C.56}$$

where the last step follows from $\|e_j\|_2 = 1$ and the Cauchy–Schwarz inequality.

Define

$$\theta = \frac{w_j}{\|w_j\|_2} \quad \text{and} \qquad \psi(t) = \frac{1}{\|w_j\|_2}\big(\sigma'(\|w_j\|_2\, t) - \sigma'(0)\big),$$

we know

$$\frac{1}{\|w_j\|_2}\big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big) = \frac{1}{\|w_j\|_2}\big(\sigma'(\|w_j\|_2 \cdot \theta^\top \phi(x_i)) - \sigma'(0)\big) = \psi(\theta^\top \phi(x_i)),$$

then

$$\begin{aligned}
&\Big\| \sum_{i=1}^{n} g_i \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big)\phi(x_i) \Big\|_2 \\
&= \|w_j\|_2 \Big\| \sum_{i=1}^{n} g_i \frac{1}{\|w_j\|_2} \big(\sigma'(w_j^\top \phi(x_i)) - \sigma'(0)\big)\phi(x_i) \Big\|_2 \\
&= \|w_j\|_2 \Big\| \sum_{i=1}^{n} g_i \psi(\theta^\top \phi(x_i))\phi(x_i) \Big\|_2 \\
&\le \|w_j\|_2 \sup_{\psi \in \Psi,\ \|\theta\|_2 \le 1} \Big\| \sum_{i=1}^{n} g_i \psi(\theta^\top \phi(x_i))\phi(x_i) \Big\|_2,
\end{aligned}$$

where

$$\Psi := \{\psi : \psi(0) = 0,\ \mathrm{Lip}(\psi) \le L_{\sigma'}\}.$$

Plugging the above inequality into Eq. (C.56) and using $\sum_{j=1}^{d_1} \|w_j\|_2^2 = \|\mathbf{W}\|_F^2$, we get

$$\Big\| \sum_{i=1}^{n} g_i \Big(a \odot \big(\sigma'(\mathbf{W}\phi(x_i)) - \sigma'(0)\mathbf{1} \big) \Big) \phi(x_i)^\top \Big\|_* \\ \leq \|a\|_2 \|\mathbf{W}\|_F \sup_{\psi \in \Psi,\ \|\theta\|_2 \leq 1} \Big\| \sum_{i=1}^{n} g_i \psi(\theta^\top \phi(x_i)) \phi(x_i) \Big\|_2. \tag{C.57}$$

Combining Eq. (C.57) and Eq. (C.56), noting the fact that $\|\mathbf{W}\|_F \leq \sqrt{\mathrm{rank}(\mathbf{W})} \|\mathbf{W}\|_{\mathrm{op}} \leq R\sqrt{r}$, we obtain

$$\mathbb{E}_g \sup_{\mathbf{W} \in \mathcal{B}_{R,r}} \Big\| \sum_{i=1}^{n} g_i \Big(a \odot \big(\sigma'(\mathbf{W}\phi(x_i)) - \sigma'(0)\mathbf{1} \big) \Big) \phi(x_i)^\top \Big\|_* \\ \leq R\sqrt{r}\, \|a\|_2 \mathbb{E}_g \sup_{\psi \in \Psi,\ \|\theta\|_2 \leq 1} \Big\| \sum_{i=1}^{n} g_i \psi(\theta^\top \phi(x_i)) \phi(x_i) \Big\|_2. \tag{C.58}$$

It remains to control the single-neuron process in Eq. (C.58). By Euclidean duality,

$$\mathbb{E}_g \sup_{\psi \in \Psi,\ \|\theta\|_2 \leq 1} \Big\| \sum_{i=1}^{n} g_i \psi(\theta^\top \phi(x_i)) \phi(x_i) \Big\|_2 \\ = \mathbb{E}_g \sup_{\psi \in \Psi,\ \|\theta\|_2 \leq 1, \|z\|_2 \leq 1} \sum_{i=1}^{n} g_i \psi(\theta^\top \phi(x_i)) z^\top \phi(x_i) \\ = \mathbb{E}_{|g|} \mathbb{E}_\epsilon \sup_{\psi \in \Psi,\ \|\theta\|_2 \leq 1, \|z\|_2 \leq 1} \sum_{i=1}^{n} \epsilon_i |g_i| \psi(\theta^\top \phi(x_i)) z^\top \phi(x_i), \tag{C.59}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher variables. We know

$$\frac{1}{n} \mathbb{E}_\epsilon \sup_{\|z\|_2 \leq 1} \sum_{i=1}^{n} \epsilon_i z^\top \phi(x_i) = \frac{1}{n} \mathbb{E}_\epsilon \Big\| \sum_{i=1}^{n} \epsilon_i \phi(x_i) \Big\|_2 \leq \frac{1}{n} \Big(\mathbb{E}_\epsilon \Big\| \sum_{i=1}^{n} \epsilon_i \phi(x_i) \Big\|_2^2 \Big)^{1/2} \\ = \frac{1}{n} \Big(\sum_{i=1}^{n} \|\phi(x_i)\|_2^2 \Big)^{1/2} = \frac{B_x}{\sqrt{n}}. \tag{C.60}$$

Every map $x \mapsto z^\top \phi(x)$ with $\|z\|_2 \leq 1$ takes values in $[-B_x, B_x]$. Applying Lemma 35 with $R = B_x$, $a = 0$ and $L = L_{\sigma'}$ therefore yields

$$\frac{1}{n} \mathbb{E}_\epsilon \sup_{\psi \in \Psi,\ \|\theta\|_2 \leq 1} \sum_{i=1}^{n} \epsilon_i \psi(\theta^\top \phi(x_i)) \\ \leq c L_{\sigma'} \left(\frac{B_x}{\sqrt{n}} + \log^{3/2}(2n)\, \frac{1}{n} \mathbb{E}_\epsilon \sup_{\|z\|_2 \leq 1} \sum_{i=1}^{n} \epsilon_i z^\top \phi(x_i) \right) \\ \leq \frac{c L_{\sigma'} B_x}{\sqrt{n}} \big(1 + \log^{3/2}(2n) \big). \tag{C.61}$$

Next, for every $\psi \in \Psi$ and $\|\theta\|_2 \leq 1$ we have

$$|\psi(\theta^\top \phi(x))| \leq L_{\sigma'} |\theta^\top \phi(x)| \leq L_{\sigma'} \|\theta\|_2 \|\phi(x)\|_2 \leq L_{\sigma'} B_x,$$

and $|z^\top \phi(x)| \leq B_x$ whenever $\|z\|_2 \leq 1$, then both $\frac{\psi(\theta^\top \phi(x))}{L_{\sigma'} B_x}$ and $\frac{z^\top \phi(x)}{B_x}$ lie in $[-1, 1]$. Define $H(s, t) = st$, then

$$\psi(\theta^\top \phi(x))\, z^\top \phi(x) = L_{\sigma'} B_x^2\, H\Big(\frac{\psi(\theta^\top \phi(x))}{L_{\sigma'} B_x},\ \frac{z^\top \phi(x)}{B_x} \Big),$$

and H is $\sqrt{2}$-Lipschitz on $[-1, 1]^2$ since for any $(s, t), (s', t') \in \mathbb{R}^2$,

$$\big| H(s,t) - H(s',t') \big| = |st - s't'| = \big| s(t - t') + t'(s - s') \big| \\ \leq |s|\, |t - t'| + |t'|\, |s - s'| \\ \leq |t - t'| + |s - s'| \\ \leq \sqrt{2}\, \sqrt{(s - s')^2 + (t - t')^2}.$$

Hence Lemma 37 together with Eq. (C.60) and Eq. (C.61) give

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\epsilon \sup_{\substack{\psi\in\Psi,\ \|\theta\|_2\le 1\\ \|z\|_2\le 1}} \sum_{i=1}^n \epsilon_i \psi(\theta^\top\phi(x_i)) z^\top \phi(x_i) \\
&\le \sqrt{2}\, L_{\sigma'} B_x^2 \cdot \frac{1}{n}\mathbb{E}_\epsilon \sup_{\substack{\psi\in\Psi,\ \|\theta\|_2\le 1\\ \|z\|_2\le 1}} \sum_{i=1}^n \Big(\epsilon_{i1}\frac{\psi(\theta^\top\phi(x_i))}{L_{\sigma'}B_x} + \epsilon_{i2}\frac{z^\top\phi(x_i)}{B_x}\Big) \\
&\le 2\Big(B_x\cdot\frac{1}{n}\mathbb{E}_\epsilon \sup_{\psi\in\Psi,\ \|\theta\|_2\le 1}\sum_{i=1}^n \epsilon_i\psi(\theta^\top\phi(x_i)) + L_{\sigma'}B_x\cdot\frac{1}{n}\mathbb{E}_\epsilon\sup_{\|z\|_2\le 1}\sum_{i=1}^n\epsilon_i z^\top\phi(x_i)\Big) \\
&\le \frac{cL_{\sigma'}B_x^2}{\sqrt{n}}\big(1+\log^{3/2}(2n)\big), \qquad \text{(C.62)}
\end{aligned}$$

where $\{\epsilon_{ik}\}$ are independent Rademacher variables. Finally, plugging Eq. (C.62) into Eq. (C.59), together with $\mathbb{E}\max_{i\in[n]}|g_i| \le \sqrt{2\log(2n)}$, yields

$$\begin{aligned}
\mathbb{E}_g \sup_{\psi\in\Psi,\ \|\theta\|_2\le 1}\Big\|\sum_{i=1}^n g_i\psi(\theta^\top\phi(x_i))\phi(x_i)\Big\|_2 &\le cL_{\sigma'}B_x^2\sqrt{n}\,\sqrt{\log(2n)}\big(1+\log^{3/2}(2n)\big) \\
&= cL_{\sigma'}B_x^2\sqrt{n}\,\Lambda_n. \qquad \text{(C.63)}
\end{aligned}$$

Plugging Eq. (C.63) into Eq. (C.58), we obtain

$$\mathbb{E}_g \sup_{\mathbf{W}\in\mathcal{B}_{R,r}}\Big\|\sum_{i=1}^n g_i\Big(a\odot\big(\sigma'(\mathbf{W}\phi(x_i))-\sigma'(0)\mathbf{1}\big)\Big)\phi(x_i)^\top\Big\|_* \le cL_{\sigma'}R\sqrt{r}\,\|a\|_2 B_x^2\sqrt{n}\,\Lambda_n. \qquad \text{(C.64)}$$

Combining (C.54), (C.55), and (C.64) completes the proof. □

Now we combine Lemma 39 and Lemma 40 to derive the localized Rademacher complexity of the gradient class induced by SNNs.

Lemma 41. *Assume that $\|\phi(x)\|_2 \le B_x$ for all x; $\|\sigma'(\cdot)\|_\infty \le L_\sigma$ and $\sigma'(\cdot)$ is $L_{\sigma'}$-Lipschitz; $t \mapsto \ell'(y,t)$ is L_ℓ-Lipschitz and $|\ell'(y,t)| \le G_\ell$. Let Assumption 3 holds. Let $\mathscr{L}_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell'\big(y_i, a^\top\sigma(\mathbf{W}\phi(x_i))\big)^2$ and $\widehat{\mathcal{W}}_\rho = \{\mathbf{W}\in\mathcal{W} : \mathscr{L}_S(\mathbf{W}) \le \rho\}$. Then there exist universal constants $c>0$ such that*

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\epsilon \sup_{\mathbf{W}\in\widehat{\mathcal{W}}_\rho}\Big\|\sum_{i=1}^n \epsilon_i\nabla_{\mathbf{W}} f(\mathbf{W};z_i)\Big\|_* &\le \frac{12\sqrt{\rho}}{\sqrt{n}} + \frac{c\sqrt{\rho}}{\sqrt{n}}L_\sigma^2\|a\|_2^2 B_x^2\gamma R\sqrt{r}\,\Lambda_n\left[\log\Big(\frac{G_\ell L_\sigma\|a\|_2 B_x\sqrt{n}}{3\sqrt{\rho}}\Big)\right]_+ \\
&\quad + \frac{c\sqrt{\rho}}{\sqrt{n}}\|a\|_2 B_x\big(L_\sigma + L_{\sigma'}R\sqrt{r}\,B_x\Lambda_n\big)\left[\log\Big(\frac{G_\ell L_\sigma\|a\|_2 B_x\sqrt{n}}{3\sqrt{\rho}}\Big)\right]_+, \qquad \text{(C.65)}
\end{aligned}$$

where $[t]_+ := \max\{t,0\}$.

Proof of Lemma 41. If $a=0$, then

$$\nabla_{\mathbf{W}} f(\mathbf{W};z) = 0,$$

and the assertion is immediate. Henceforth assume $a \ne 0$.

Denote $s_{\mathbf{W}}(x) = a^\top\sigma(\mathbf{W}\phi(x))$. By the chain rule,

$$\nabla_{\mathbf{W}} f(\mathbf{W};z) = \ell'\big(y, s_{\mathbf{W}}(x)\big)\big(a\odot\sigma'(\mathbf{W}\phi(x))\big)\phi(x)^\top. \qquad \text{(C.66)}$$

For $\mathbf{W}\in\mathcal{W}$ and $\|\mathbf{C}\|_{\mathrm{op}} \le 1$, define

$$\begin{aligned}
h_{\mathbf{W},\mathbf{C}}(z) &:= \big\langle \nabla_{\mathbf{W}} f(\mathbf{W};z), \mathbf{C}\big\rangle \\
&= \ell'\big(y, s_{\mathbf{W}}(x)\big)\big(a\odot\sigma'(\mathbf{W}\phi(x))\big)^\top\mathbf{C}\phi(x). \qquad \text{(C.67)}
\end{aligned}$$

The dual representation of the nuclear norm gives

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\mathbf{W}\in\widehat{\mathcal{W}}_\rho} \Big\| \sum_{i=1}^n \epsilon_i \nabla_{\mathbf{W}} f(\mathbf{W}; z_i)\Big\|_* = \frac{1}{n}\mathbb{E}_\epsilon \sup_{\substack{\mathbf{W}\in\widehat{\mathcal{W}}_\rho \\ \|\mathbf{C}\|_{\mathrm{op}}\le 1}} \sum_{i=1}^n \epsilon_i h_{\mathbf{W},\mathbf{C}}(z_i). \tag{C.68}$$

Step 1: Metric comparison.

Define

$$h^{(1)}_{\mathbf{W}}(z) := \ell'\big(y, a^\top \sigma(\mathbf{W}\phi(x))\big),$$
$$h^{(2)}_{\mathbf{W},\mathbf{C}}(x) := \big(a \odot \sigma'(\mathbf{W}\phi(x))\big)^\top \mathbf{C}\phi(x).$$

Then

$$h_{\mathbf{W},\mathbf{C}} = h^{(1)}_{\mathbf{W}} h^{(2)}_{\mathbf{W},\mathbf{C}}.$$

Since $\|\sigma'(\cdot)\|_\infty \le L_\sigma$, we know

$$\begin{aligned} |h^{(2)}_{\mathbf{W},\mathbf{C}}(x_i)| &\le \big\|a \odot \sigma'(\mathbf{W}\phi(x_i))\big\|_2 \|\mathbf{C}\phi(x_i)\|_2 \\ &\le L_\sigma \|a\|_2 \|\mathbf{C}\|_{\mathrm{op}} \|\phi(x_i)\|_2 \le L_\sigma \|a\|_2 B_x. \end{aligned} \tag{C.69}$$

For any $\mathbf{W}, \mathbf{W}' \in \widehat{\mathcal{W}}_\rho$ and $\|\mathbf{C}\|_{\mathrm{op}}, \|\mathbf{C}'\|_{\mathrm{op}} \le 1$,

$$h_{\mathbf{W},\mathbf{C}} - h_{\mathbf{W}',\mathbf{C}'} = \big(h^{(1)}_{\mathbf{W}} - h^{(1)}_{\mathbf{W}'}\big) h^{(2)}_{\mathbf{W},\mathbf{C}} + h^{(1)}_{\mathbf{W}'} \big(h^{(2)}_{\mathbf{W},\mathbf{C}} - h^{(2)}_{\mathbf{W}',\mathbf{C}'}\big).$$

Using $(u+v)^2 \le 2u^2 + 2v^2$, (C.69) and the definition of $\mathscr{L}_S(\mathbf{W})$, we obtain

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \big(h_{\mathbf{W},\mathbf{C}}(z_i) - h_{\mathbf{W}',\mathbf{C}'}(z_i)\big)^2 \\ &\le \frac{2}{n}\sum_{i=1}^n \big(h^{(1)}_{\mathbf{W}}(z_i)\big)^2 \big(h^{(2)}_{\mathbf{W},\mathbf{C}}(x_i) - h^{(2)}_{\mathbf{W}',\mathbf{C}'}(x_i)\big)^2 + \frac{2}{n}\sum_{i=1}^n \big(h^{(2)}_{\mathbf{W}',\mathbf{C}'}(x_i)\big)^2 \big(h^{(1)}_{\mathbf{W}}(z_i) - h^{(1)}_{\mathbf{W}'}(z_i)\big)^2 \\ &\le 2\mathscr{L}_S(\mathbf{W}) \max_{i\in[n]} \big(h^{(2)}_{\mathbf{W},\mathbf{C}}(x_i) - h^{(2)}_{\mathbf{W}',\mathbf{C}'}(x_i)\big)^2 + \frac{2L_\sigma^2 \|a\|_2^2 B_x^2}{n} \sum_{i=1}^n \big(h^{(1)}_{\mathbf{W}}(z_i) - h^{(1)}_{\mathbf{W}'}(z_i)\big)^2. \end{aligned} \tag{C.70}$$

Plugging this estimate into the second term of (C.70) and using $\mathscr{L}_S(\mathbf{W}) \le \rho$, $\mathscr{L}_S(\mathbf{W}') \le \rho$ yields

$$\begin{aligned} d_S(h_{\mathbf{W},\mathbf{C}}, h_{\mathbf{W}',\mathbf{C}'}) \le \sqrt{2\rho} \max_{i\in[n]} |h^{(2)}_{\mathbf{W},\mathbf{C}}(x_i) - h^{(2)}_{\mathbf{W}',\mathbf{C}'}(x_i)| \\ + 2L_\sigma \|a\|_2 B_x \gamma \sqrt{\rho} \max_{i\in[n]} |s_{\mathbf{W}}(x_i) - s_{\mathbf{W}'}(x_i)|, \end{aligned} \tag{C.71}$$

where

$$d_S(g, g') = \Big(\frac{1}{n}\sum_{i=1}^n (g(z_i) - g'(z_i))^2\Big)^{1/2}.$$

Step 2: Covering-number decomposition.

Let

$$\mathcal{H} = \{h_{\mathbf{W},\mathbf{C}} : \mathbf{W} \in \widehat{\mathcal{W}}_\rho,\ \|\mathbf{C}\|_{\mathrm{op}} \le 1\},$$

and define

$$\mathcal{H}_1 = \{h^{(1)}_{\mathbf{W}} : \mathbf{W} \in \widehat{\mathcal{W}}_\rho\}, \qquad \mathcal{H}_{2,\sigma} = \{h^{(2)}_{\mathbf{W},\mathbf{C}} : \mathbf{W} \in \widehat{\mathcal{W}}_\rho,\ \|\mathbf{C}\|_{\mathrm{op}} \le 1\}.$$

It follows from (C.71) that

$$\log \mathcal{N}(\epsilon, \mathcal{H}, d_S) \le \log \mathcal{N}\Big(\frac{\epsilon}{2\sqrt{2\rho}\, L_\sigma \|a\|_2 B_x \gamma}, \mathcal{S}, d_{S_x,\infty}\Big) + \log \mathcal{N}\Big(\frac{\epsilon}{2\sqrt{2\rho}}, \mathcal{H}_{2,\sigma}, d_{S_x,\infty}\Big), \tag{C.72}$$

where $\mathcal{S} = \{x \mapsto s_{\mathbf{W}}(x) : \mathbf{W} \in \mathcal{W}\}$ and $d_{S_x,\infty}$ denotes the empirical ℓ_∞ metric on $S_x = \{x_1, \dots, x_n\}$, given by $d_{S_x,\infty}(x_i, x_j) = \|x_i - x_j\|_\infty$.

Step 3: Covering $\mathcal{S}$.

Define

$$n\mathfrak{R}_S(\mathcal{S}) := \mathbb{E}_g\Big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n g_i a^\top \sigma(\mathbf{W}\phi(x_i))\Big].$$

By Sudakov minoration inequality (Lemma 26), for any $\eta > 0$,

$$\log^{1/2} \mathcal{N}(\eta, \mathcal{S}, d_{S_x,\infty}) \le \frac{1}{\sqrt{n}\,\eta} n\mathfrak{R}_S(\mathcal{S}). \tag{C.73}$$

Then taking $\eta = \frac{\epsilon}{2\sqrt{2\rho}\,L_\sigma \|a\|_2 B_x \gamma}$ in Eq. (C.73) gives

$$\log^{1/2} \mathcal{N}\Big(\frac{\epsilon}{2\sqrt{2\rho}\,L_\sigma\|a\|_2 B_x\gamma}, \mathcal{S}, d_{S_x,\infty}\Big) \le \frac{2\sqrt{2\rho}\,L_\sigma\|a\|_2 B_x\gamma}{\sqrt{n}\,\epsilon} n\mathfrak{R}_S(\mathcal{S}). \tag{C.74}$$

Step 5: Covering $\mathcal{H}_{2,\sigma}$.

Since the zero function belongs to $\mathcal{H}_{2,\sigma}$ by taking $\mathbf{C} = 0$, Sudakov minoration inequality (Lemma 26) gives

$$\log^{1/2} \mathcal{N}(\eta, \mathcal{H}_{2,\sigma}, d_{S_x,\infty}) \le \frac{K_2(S)}{\sqrt{n}\,\eta}.$$

Taking $\eta = \frac{\epsilon}{2\sqrt{2\rho}}$ and yields

$$\log^{1/2} \mathcal{N}\Big(\frac{\epsilon}{2\sqrt{2\rho}}, \mathcal{H}_{2,\sigma}, d_{S_x,\infty}\Big) \le \frac{2\sqrt{2\rho}}{\sqrt{n}\epsilon} K_2(S). \tag{C.75}$$

Using (C.72), $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$, (C.74), and (C.75), we obtain

$$\log^{1/2} \mathcal{N}(\epsilon, \mathcal{H}, d_S) \le \frac{2\sqrt{2\rho}\,L_\sigma\|a\|_2 B_x\gamma}{\sqrt{n}\,\epsilon} n\mathfrak{R}_S(\mathcal{S}) + \frac{2\sqrt{2\rho}}{\sqrt{n}\epsilon} K_2(S). \tag{C.76}$$

Step 6: Dudley's entropy integral. By (C.69) and the localization constraint $\mathscr{L}_S(\mathbf{W}) \le \rho$, $\frac{1}{n}\sum_{i=1}^n h_{\mathbf{W},\mathbf{C}}(z_i)^2 \le L_\sigma^2\|a\|_2^2 B_x^2 \rho$. In particular, $\max_i |h_{\mathbf{W},\mathbf{C}}(z_i)| \le G_\ell L_\sigma \|a\|_2 B_x$. Applying Dudley's entropy integral with $\alpha = 3\sqrt{\rho}/\sqrt{n}$ gives

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\substack{\mathbf{W}\in\widehat{\mathcal{W}}_\rho \\ \|\mathbf{C}\|_{\mathrm{op}}\le 1}} \sum_{i=1}^n \epsilon_i h_{\mathbf{W},\mathbf{C}}(z_i) \le \frac{12\sqrt{\rho}}{\sqrt{n}} + \frac{12}{\sqrt{n}} \int_{3\sqrt{\rho}/\sqrt{n}}^{G_\ell L_\sigma\|a\|_2 B_x} \log^{1/2} \mathcal{N}(\epsilon, \mathcal{H}, d_S)\, d\epsilon,$$

where the integral is understood as zero if its upper endpoint is smaller than its lower endpoint. Substituting (C.76) and using

$$\int_{3\sqrt{\rho}/\sqrt{n}}^{G_\ell L_\sigma\|a\|_2 B_x} \frac{d\epsilon}{\epsilon} = \Big[\log\Big(\frac{G_\ell L_\sigma \|a\|_2 B_x \sqrt{n}}{3\sqrt{\rho}}\Big)\Big]_+$$

yields

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{\mathbf{W}\in\widehat{\mathcal{W}}_\rho} \Big\|\sum_{i=1}^n \epsilon_i \nabla_{\mathbf{W}} f(\mathbf{W}; z_i)\Big\|_* \le \frac{12\sqrt{\rho}}{\sqrt{n}} + 24\sqrt{2\rho} L_\sigma \|a\|_2 B_x \gamma \mathfrak{R}_S(\mathcal{S}) \Big[\log\Big(\frac{G_\ell L_\sigma\|a\|_2 B_x\sqrt{n}}{3\sqrt{\rho}}\Big)\Big]_+$$
$$+ \frac{24\sqrt{2\rho}}{\sqrt{n}} K_2(S) \Big[\log\Big(\frac{G_\ell L_\sigma\|a\|_2 B_x\sqrt{n}}{3\sqrt{\rho}}\Big)\Big]_+. \tag{C.77}$$

Plugging Eq. (C.50) and Eq. (C.53) into Eq. (C.77) completes the proof. □

We recall the following uniform comparison between

$$\mathscr{L}_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell'\big(y_i, a^\top\sigma(\mathbf{W}\phi(x_i))\big)^2 \quad \text{and} \quad \mathscr{L}(\mathbf{W}) = \mathbb{E}_Z[\ell'(Y, a^\top\sigma(\mathbf{W}\phi(x)))^2.$$

Lemma 42 (Authors (2026)). *Let Assumption 3 hold. Assume that $|\ell'(y,t)| \le G_\ell$ for all $y \in \mathcal{Y}$ and $t \in \mathbb{R}$. Let $\mathcal{S} = \{x \mapsto s_{\mathbf{W}}(x) : \mathbf{W} \in \mathcal{W}\}$, $\mathfrak{R}_n(\mathcal{S}) := \sup_{S_x:|S_x|=n} \mathfrak{R}_{S_x}(\mathcal{S})$. Assume that $\gamma\mathfrak{R}_n(\mathcal{S}) \lesssim 1$. Then, for every $x > 0$, with probability at least $1 - 2\exp(-x)$, the following inequality holds simultaneously for all $\mathbf{W} \in \mathcal{W}$:*

$$\max\{\mathscr{L}_S(\mathbf{W}) - 2\mathscr{L}(\mathbf{W}), \mathscr{L}(\mathbf{W}) - 2\mathscr{L}_S(\mathbf{W})\} \lesssim \gamma^2 G_\ell^2 R_n(\mathcal{S})^2 \log^4(G_\ell b_s \gamma n) + \frac{G_\ell^2\big(x + \log\log n\big)}{n},$$

where $b_s := \sup_{\mathbf{W}\in\mathcal{W},\, x\in\mathcal{X}} |s_{\mathbf{W}}(x)|$.

The following lemma gives the localized uniform convergence for shallow networks.

Theorem 43 (Localized uniform convergence for shallow networks). *Let $s_{\mathbf{W}}(x) = a^\top \sigma(\mathbf{W}\phi(x))$, $\mathscr{L}(\mathbf{W}) = \mathbb{E}_Z[\ell'(Y, s_{\mathbf{W}}(X))^2$, and define $\mathcal{W}_\rho = \{\mathbf{W} \in \mathcal{W} : \mathscr{L}(\mathbf{W}) \le \rho\}$. Assume that $\|\phi(x)\|_2 \le B_x$, $\|\sigma'(\cdot)\|_\infty \le L_\sigma$, and $|\ell'(y,t)| \le G_\ell$. Then, for every $\rho \ge 0$ and $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sup_{\mathbf{W}\in\mathcal{W}_\rho} \Big\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\Big\|_* \le 4\mathbb{E}_\epsilon\left[\sup_{\mathbf{W}\in\mathcal{W}_\rho} \frac{1}{n}\Big\|\sum_{i=1}^n \epsilon_i \nabla_{\mathbf{W}} f(\mathbf{W}; z_i)\Big\|_*\right] + L_\sigma \|a\|_2 B_x \Big(\frac{2\rho\log(2/\delta)}{n}\Big)^{1/2} + \frac{8G_\ell L_\sigma \|a\|_2 B_x \log(2/\delta)}{3n}.$$

Proof. The proof follows similarly from the proof of Theorem 1 in Authors (2026) with the ℓ_2 norm replaced by the nuclear norm. Let

$$\mathbb{B}_\sigma = \{\mathbf{C} : \|\mathbf{C}\|_\sigma \le 1\}, \qquad \mathcal{H}_\rho = \{z \mapsto \langle \nabla_{\mathbf{W}} f(\mathbf{W}; z), \mathbf{C}\rangle : (\mathbf{W}, \mathbf{C}) \in \mathcal{W}_\rho \times \mathbb{B}_\sigma\}.$$

Differentiating $f(\mathbf{W}; z) = \ell(y, s_{\mathbf{W}}(x))$ with respect to $\mathbf{W}$ gives

$$\nabla_{\mathbf{W}} f(\mathbf{W}; z) = \ell'\big(y, s_{\mathbf{W}}(x)\big) D_{\mathbf{W}}(x) a \phi(x)^\top, \qquad D_{\mathbf{W}}(x) := \mathrm{Diag}\big(\sigma'(\mathbf{W}\phi(x))\big). \tag{C.78}$$

For every $(\mathbf{W}, \mathbf{C}) \in \mathcal{W}_\rho \times \mathbb{B}_\sigma$, nuclear-operator norm duality yields

$$\begin{aligned}|\langle \nabla_{\mathbf{W}} f(\mathbf{W}; z), \mathbf{C}\rangle| &\le \|\nabla_{\mathbf{W}} f(\mathbf{W}; z)\|_* \\ &\le |\ell'(y, s_{\mathbf{W}}(x))| \|D_{\mathbf{W}}(x) a\|_2 \|\phi(x)\|_2 \le G_\ell L_\sigma \|a\|_2 B_x.\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathrm{Var}_Z[\langle \nabla_{\mathbf{W}} f(\mathbf{W}; Z), \mathbf{C}\rangle] &\le \mathbb{E}_Z[\langle \nabla_{\mathbf{W}} f(\mathbf{W}; Z), \mathbf{C}\rangle^2] \\ &\le L_\sigma^2 \|a\|_2^2 B_x^2 \mathbb{E}_Z[\ell'\big(Y, s_{\mathbf{W}}(X)\big)^2] \\ &\le L_\sigma^2 \|a\|_2^2 B_x^2 \rho.\end{aligned}$$

Applying Lemma 29 to $\mathcal{H}_\rho$, with $b = G_\ell L_\sigma \|a\|_2 B_x$, $r = L_\sigma^2 \|a\|_2^2 B_x^2 \rho$ shows that, with probability at least $1 - \delta$,

$$\begin{aligned}&\sup_{(\mathbf{W},\mathbf{C})\in\mathcal{W}_\rho\times\mathbb{B}_\sigma} \left[\mathbb{E}_Z \langle \nabla_{\mathbf{W}} f(\mathbf{W}; Z), \mathbf{C}\rangle - \frac{1}{n}\sum_{i=1}^n \langle \nabla_{\mathbf{W}} f(\mathbf{W}; z_i), \mathbf{C}\rangle\right] \\ &\le 4\mathfrak{R}_S(\mathcal{H}_\rho) + L_\sigma \|a\|_2 B_x \left(\frac{2\rho\log(2/\delta)}{n}\right)^{1/2} + \frac{8G_\ell L_\sigma \|a\|_2 B_x \log(2/\delta)}{3n}.\end{aligned} \tag{C.79}$$

By nuclear–operator norm duality, the left-hand side of (C.79) equals

$$\sup_{\mathbf{W}\in\mathcal{W}_\rho} \|\nabla F(\mathbf{W}) - \nabla F_S(\mathbf{W})\|_*.$$

Then,

$$\begin{aligned}\mathfrak{R}_S(\mathcal{H}_\rho) &= \mathbb{E}_\epsilon\left[\sup_{(\mathbf{W},\mathbf{C})\in\mathcal{W}_\rho\times\mathbb{B}_\sigma} \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle \nabla_{\mathbf{W}} f(\mathbf{W}; z_i), \mathbf{C}\rangle\right] \\ &= \mathbb{E}_\epsilon\left[\sup_{\mathbf{W}\in\mathcal{W}_\rho} \frac{1}{n}\Big\|\sum_{i=1}^n \epsilon_i \nabla_{\mathbf{W}} f(\mathbf{W}; z_i)\Big\|_*\right].\end{aligned}$$

Combining with Eq. (C.79) completes the proof. □

Now we give the proof of the uniform convergence for SNNs.

Proof of Theorem 11. By Lemma 39,

$$\mathfrak{R}_n(\mathcal{S}) \lesssim \frac{L_\sigma R\sqrt{r}\,\|a\|_2 B_x \Lambda_n}{\sqrt{n}}.$$

Fix $\rho \in [0, G_\ell^2]$. By Lemma 42, with probability at least $1-\delta$, every $\mathbf{W}$ satisfying $\mathscr{L}(\mathbf{W}) \le \rho$ also satisfies

$$\mathscr{L}_S(\mathbf{W}) \le 2\rho + c\frac{\gamma^2 G_\ell^2 L_\sigma^2 R^2 r \|a\|_2^2 B_x^2 \Lambda_n^2 \log^4(G_\ell b_s \gamma n)}{n} + c\frac{G_\ell^2\big(\log(2/\delta) + \log\log n\big)}{n}. \quad \text{(C.80)}$$

Then applying Lemma 41 followed by Theorem 43, yields

$$\sup_{\mathbf{W}:\,\mathscr{L}(\mathbf{W})\le\rho} \|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* = \widetilde{\mathcal{O}}\left(\frac{1 + L_\sigma\|a\|_2 B_x + R\sqrt{r}\,\|a\|_2 B_x^2\left(L_{\sigma'} + \gamma L_\sigma^2\|a\|_2\right)}{\sqrt{n}} \times \left[\sqrt{\rho} + \frac{G_\ell\left(1 + \gamma L_\sigma R\sqrt{r}\,\|a\|_2 B_x\right)\log(1/\delta)}{\sqrt{n}}\right]\right). \quad \text{(C.81)}$$

The right-hand side of (C.81) is non-decreasing in ρ.

We apply Lemma 34 with $\widetilde{V}(\mathbf{W}) = \mathscr{L}(\mathbf{W})$, $\widetilde{R} = G_\ell^2$, and $s_0 = G_\ell^2/n$. After replacing the confidence level by $\delta/\log_2(2n)$, the peeling argument shows that, with probability at least $1-\delta$, the bound in (C.81) holds simultaneously for every $\mathbf{W} \in \mathcal{W}$,

$$\|\nabla F_S(\mathbf{W}) - \nabla F(\mathbf{W})\|_* = \widetilde{\mathcal{O}}\left(\frac{1 + L_\sigma\|a\|_2 B_x + R\sqrt{r}\,\|a\|_2 B_x^2\left(L_{\sigma'} + \gamma L_\sigma^2\|a\|_2\right)}{\sqrt{n}} \times \left[\sqrt{\mathscr{L}(\mathbf{W})} + \frac{G_\ell\left(1 + \gamma L_\sigma R\sqrt{r}\,\|a\|_2 B_x\right)\log(1/\delta)}{\sqrt{n}}\right]\right), \quad \text{(C.82)}$$

where $\sqrt{\rho}$ is replaced, up to a universal constant, by $\sqrt{\mathscr{L}(\mathbf{W})} + G_\ell/\sqrt{n}$.

Finally, Lemma 42, together with Lemma 39, gives simultaneously for all $\mathbf{W} \in \mathcal{W}$,

$$\sqrt{\mathscr{L}(\mathbf{W})} = \widetilde{\mathcal{O}}\left(\sqrt{\mathscr{L}_S(\mathbf{W})} + \frac{G_\ell\left(1 + \gamma L_\sigma R\sqrt{r}\,\|a\|_2 B_x \Lambda_n\right)\log(2/\delta)}{\sqrt{n}}\right),$$

where logarithmic factors in n and $G_\ell b_s \gamma$ are absorbed into $\widetilde{\mathcal{O}}$. Substituting this inequality into the uniform peeled bound completes the proof. □

We now apply Theorem 11 to the Muon iterates, which gives the dimension-independent high-probability bound on the generalization gap of Muon for SNNs.

Proof of Corollary 12. By Lemma 2, with the constant learning rate $\eta_t = \eta$, the operator norm of the iterates $\mathbf{W}_t$ for any $t \le T$ is bounded by

$$\|\mathbf{W}_t\|_\sigma \le \|\mathbf{W}_0\|_\sigma + \sum_{k=0}^{t-1}\eta \le \|\mathbf{W}_0\|_\sigma + T\eta.$$

This implies that for all $0 \le t \le T$, the iterates $\mathbf{W}_t$ reside within a parameter ball $\mathcal{B}_{R,r}$ with radius $R = \|\mathbf{W}_0\|_\sigma + T\eta$.
By Theorem 11, the uniform generalization gap bound holds simultaneously for all $\mathbf{W} \in \mathcal{B}_{R,r}$ with probability at least $1-\delta$. Substituting $R_T = \|\mathbf{W}_0\|_\sigma + T\eta$ into the bound from Theorem 11 directly completes the proof. □

We now specialize the optimization errors established in Theorem 4 and Corollary 5 to the SNNs (4.8), which gives the following optimization error.

Theorem 44 (High-probability optimization error of Muon for SNNs)**.** *Suppose f takes the form in Eq. (4.8). Let assumptions and definitions in Theorem 11 hold. Then Theorem 4 and Corollary 5 hold with $L_* = B_x^2\big(L_\ell L_\sigma^2\|a\|_2^2 + G_\ell L_{\sigma'}\|a\|_2\big)$, $G_F = 2G_\ell L_\sigma\|a\|_2 B_x$, $\sigma = G_\ell L_\sigma\|a\|_2 B_x$, $\Gamma_\delta = G_\ell L_\sigma\|a\|_2 B_x\big(\frac{1}{\sqrt{B}} + \frac{2}{3}\big)\log(2/\delta)$ and $q = \min\{d_1, d_2\}$.*

Proof of Theorem 44. We calculate separately all model-dependent parameters appearing in Theorem 4.

Gradient formula and uniform gradient bound. For $f(\mathbf{W}; z) = \ell\big(y, a^\top\sigma(\mathbf{W}\phi(x))\big)$, the chain rule gives

$$\nabla f(\mathbf{W}; z) = \ell'\big(y, s_{\mathbf{W}}(x)\big)\,(a \odot \sigma'(\mathbf{W}\phi(x)))\,\phi(x)^\top. \tag{C.83}$$

Since $t \mapsto \ell(y,t)$ is G_ℓ-Lipschitz, $|\ell'(y,t)| \le G_\ell$. Moreover, $\|\sigma'(\cdot)\|_\infty \le L_\sigma$ implies $\|a \odot \sigma'(\mathbf{W}\phi(x))\|_2 \le L_\sigma\|a\|_2$. Together with $\|\phi(x)\|_2 \le B_x$, this yields

$$\|\nabla f(\mathbf{W}; z)\|_{\mathrm{F}} = |\ell'(y, s_{\mathbf{W}}(x))|\|a \odot \sigma'(\mathbf{W}\phi(x))\|_2\|\phi(x)\|_2 \le G_\ell L_\sigma\|a\|_2 B_x. \tag{C.84}$$

Computation of L_*. For any $\mathbf{W}, \mathbf{W}'$ We split the difference of sample gradients into two terms by adding and subtracting $\ell'\big(y, a^\top\sigma(\mathbf{W}'\phi(x))\big)\big(a \odot \sigma'(\mathbf{W}\phi(x))\big)\phi(x)^\top$:

$$\begin{aligned}
&\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z) \\
&= \big(\ell'\big(y, a^\top\sigma(\mathbf{W}\phi(x))\big) - \ell'\big(y, a^\top\sigma(\mathbf{W}'\phi(x))\big)\big)\big(a \odot \sigma'(\mathbf{W}\phi(x))\big)\phi(x)^\top \\
&\quad + \ell'\big(y, a^\top\sigma(\mathbf{W}'\phi(x))\big)\big(a \odot \big(\sigma'(\mathbf{W}\phi(x)) - \sigma'(\mathbf{W}'\phi(x))\big)\big)\phi(x)^\top.
\end{aligned}$$

Hence, by the triangle inequality and $\|uv^\top\|_* = \|u\|_2\|v\|_2$,

$$\begin{aligned}
\|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z)\|_* \le{}& \big|\ell'\big(y, a^\top\sigma(\mathbf{W}\phi(x))\big) - \ell'\big(y, a^\top\sigma(\mathbf{W}'\phi(x))\big)\big|\|a \odot \sigma'(\mathbf{W}\phi(x))\|_2\|\phi(x)\|_2 \\
&+ \big|\ell'\big(y, a^\top\sigma(\mathbf{W}'\phi(x))\big)\big|\|a \odot \big(\sigma'(\mathbf{W}\phi(x)) - \sigma'(\mathbf{W}'\phi(x))\big)\|_2\|\phi(x)\|_2.
\end{aligned} \tag{C.85}$$

Since $t \mapsto \ell(y,t)$ is L_ℓ-smooth, $\|\sigma'(\cdot)\|_\infty \le L_\sigma$, $\sigma(\cdot)$ is L_σ-Lipschitz, and $\|\phi(x)\|_2 \le B_x$,

$$\begin{aligned}
\mathrm{I} &\le L_\ell\big|a^\top\big(\sigma(\mathbf{W}\phi(x)) - \sigma(\mathbf{W}'\phi(x))\big)\big|\|a \odot \sigma'(\mathbf{W}\phi(x))\|_2\|\phi(x)\|_2 \\
&\le L_\ell\|a\|_2\|\sigma(\mathbf{W}\phi(x)) - \sigma(\mathbf{W}'\phi(x))\|_2 \cdot L_\sigma\|a\|_2 \cdot B_x \\
&\le L_\ell\|a\|_2 \cdot L_\sigma\|(\mathbf{W} - \mathbf{W}')\phi(x)\|_2 \cdot L_\sigma\|a\|_2 \cdot B_x \\
&\le L_\ell\|a\|_2 \cdot L_\sigma B_x\|\mathbf{W} - \mathbf{W}'\|_{\mathrm{op}} \cdot L_\sigma\|a\|_2 \cdot B_x \\
&= L_\ell L_\sigma^2\|a\|_2^2 B_x^2\|\mathbf{W} - \mathbf{W}'\|_{\mathrm{op}}.
\end{aligned} \tag{C.86}$$

Since $|\ell'(y,\cdot)| \le G_\ell$, $\|a \odot v\|_2 \le \|a\|_2\|v\|_\infty$, $\sigma'(\cdot)$ is $L_{\sigma'}$-Lipschitz, and $\|\phi(x)\|_2 \le B_x$,

$$\begin{aligned}
\mathrm{II} &\le G_\ell\|a\|_2\|\sigma'(\mathbf{W}\phi(x)) - \sigma'(\mathbf{W}'\phi(x))\|_\infty\|\phi(x)\|_2 \\
&\le G_\ell\|a\|_2 \cdot L_{\sigma'}\|(\mathbf{W} - \mathbf{W}')\phi(x)\|_\infty \cdot B_x \\
&\le G_\ell\|a\|_2 \cdot L_{\sigma'}\|(\mathbf{W} - \mathbf{W}')\phi(x)\|_2 \cdot B_x \\
&\le G_\ell\|a\|_2 \cdot L_{\sigma'}B_x\|\mathbf{W} - \mathbf{W}'\|_{\mathrm{op}} \cdot B_x \\
&= G_\ell L_{\sigma'}\|a\|_2 B_x^2\|\mathbf{W} - \mathbf{W}'\|_{\mathrm{op}}.
\end{aligned} \tag{C.87}$$

Plugging Eq. (C.86) and Eq. (C.87) into Eq. (C.85) yields

$$\|\nabla f(\mathbf{W}; z) - \nabla f(\mathbf{W}'; z)\|_* \le B_x^2\big(L_\ell L_\sigma^2\|a\|_2^2 + G_\ell L_{\sigma'}\|a\|_2\big)\|\mathbf{W} - \mathbf{W}'\|_{\mathrm{op}}.$$

Then, we can take $L_* = L_{\mathrm{snn}}$, where $L_{\mathrm{snn}} = B_x^2\big(L_\ell L_\sigma^2\|a\|_2^2 + G_\ell L_{\sigma'}\|a\|_2\big)$.

Computation of b_*. Since $\sigma(0) = 0$, we have $s_{\mathbf{0}}(x) = 0$. Moreover, $\nabla f(\mathbf{0}; z)$ is an outer product and hence has rank at most one. Its nuclear norm therefore equals its Frobenius norm. Using $|\ell'(y,0)| \le G_\ell$ and $\|\sigma'(\mathbf{0})\|_\infty \le L_\sigma$, we obtain

$$b_* = \sup_z \|\nabla f(\mathbf{0}; z)\|_* = \sup_{(x,y)} |\ell'(y,0)|\|a \odot \sigma'(\mathbf{0})\|_2\|\phi(x)\|_2 \le G_\ell L_\sigma\|a\|_2 B_x.$$

Computation of σ. Since

$$\begin{aligned}\mathbb{E}_z\left[\|\nabla f(\mathbf{W};z)-\nabla F_S(\mathbf{W})\|_{\mathrm{F}}^2\right] &= \mathbb{E}_z\|\nabla f(\mathbf{W};z)\|_{\mathrm{F}}^2-\|\nabla F_S(\mathbf{W})\|_{\mathrm{F}}^2\\ &\le \mathbb{E}_z\|\nabla f(\mathbf{W};z)\|_{\mathrm{F}}^2 \le G_\ell^2 L_\sigma^2\|a\|_2^2 B_x^2,\end{aligned}$$

the variance parameter can be taken as $\sigma = G_\ell L_\sigma \|a\|_2 B_x$.
Computation of G_F. By Eq. (C.84), $\|\nabla F_S(\mathbf{W})\|_{\mathrm{F}} \le G_\ell L_\sigma \|a\|_2 B_x$, then by the triangle inequality,

$$\|\nabla f(\mathbf{W};z)-\nabla F_S(\mathbf{W})\|_{\mathrm{F}} \le \|\nabla f(\mathbf{W};z)\|_{\mathrm{F}} + \|\nabla F_S(\mathbf{W})\|_{\mathrm{F}} \le 2G_\ell L_\sigma \|a\|_2 B_x.$$

Thus, we can take $G_F = 2G_\ell L_\sigma \|a\|_2 B_x$.
Computation of q. Every sample gradient in Eq. (C.84) is an outer product and therefore has rank at most one. Unlike in the linear-network case, however, its left factor depends on both the parameter and the sample:

$$\begin{aligned}\nabla f(\mathbf{W};z) &= \mathbf{u}(\mathbf{W};z)\phi(x)^\top,\\ \mathbf{u}(\mathbf{W};z) &= \ell'\big(y, s_{\mathbf{W}}(x)\big)\left(a \odot \sigma'(\mathbf{W}\phi(x))\right).\end{aligned}$$

Consequently, different sample gradients need not share a common left factor. Their averages, the mini-batch gradients, and the momentum matrices do not have the rank one structure. Then we must take $q = \min\{d_1, d_2\}$.

Substitution into the general result. We plug in all the model-dependent parameters

$$\begin{aligned}&L_* = L_{\mathrm{sh}} = B_x^2\left(L_\ell L_\sigma^2\|a\|_2^2 + G_\ell L_{\sigma'}\|a\|_2\right), \quad b_* \le G_\ell L_\sigma \|a\|_2 B_x,\\ &G_F = 2G_\ell L_\sigma \|a\|_2 B_x, \qquad\qquad \sigma = G_\ell L_\sigma \|a\|_2 B_x, \qquad q = \min\{d_1, d_2\}.\end{aligned}$$

into Theorem 4 and Corollary 5 complete the proof. □