Non-Asymptotic Generalization and Optimization Bounds for Stochastic Gauss-Newton in Deep Neural Networks

Semih Cayci
Department of Mathematics
RWTH Aachen University
cayci@mathc.rwth-aachen.de

Abstract

An important question in deep learning is how higher-order optimization methods affect generalization. In this work, we analyze a stochastic Gauss-Newton (SGN) method with Levenberg-Marquardt damping and mini-batch sampling for training overparameterized deep neural networks with smooth activations in a regression setting. Our theoretical contributions are twofold. First, we establish finite-time optimization bounds via a variable-metric analysis in parameter space, with explicit dependencies on the batch size, network width and depth. Second, we derive non-asymptotic generalization bounds for SGN using algorithmic stability in the overparameterized regime, characterizing the impact of curvature, batch size, and overparameterization on generalization performance. Our theoretical results identify a favorable generalization regime for SGN in which a larger minimum eigenvalue of the Gauss-Newton matrix along the optimization path, together with smaller batch sizes, yields tighter stability bounds.

1 INTRODUCTION

In overparameterized models, the training algorithm plays a central role in generalization by implicitly selecting a particular interpolating hypothesis that both fits the data and generalizes well (Vardi, 2023; Bartlett et al., 2021). A core theoretical and practical challenge in deep learning is to understand this phenomenon. Curvature-aware training algorithms, such as Gauss-Newton (GN) and its variants, have gained increasing attention in deep learning (Martens, 2010; LeCun et al., 2002; Botev et al., 2017; Martens and Grosse, 2015), reinforcement learning (Kakade, 2001), and scientific machine learning (Hao et al., 2024; Mckay et al., 2025; Müller and Zeinhofer, 2023) due to their ability to achieve robust and fast convergence even in ill-conditioned settings. Despite their empirical success, a concrete theoretical understanding of these second-order methods in deep learning, particularly in the non-asymptotic regime, remains largely elusive. In particular, a concrete theoretical characterization of both the global convergence and the generalization behavior of curvature-aware preconditioned optimization remains an important open problem in deep learning (Martens, 2020; Amari et al., 2020).

Motivated by these, we revisit the stochastic Gauss-Newton method, equipped with empirical Gauss-Newton preconditioner, Levenberg-Marquardt damping, mini-batch sampling and metric projection. Two main questions that motivate our work are as follows:

1. *Can we obtain explicit finite-time optimization bounds for SGN?*
2. *How does Gauss-Newton preconditioning impact the generalization performance of the resulting predictor?*

The use of data- and time-dependent, stochastic and curvature-aware preconditioners, which are correlated with the stochastic gradient, constitutes the main challenge in the analysis of SGN.

1.1 Main Contributions

In this work, we establish non-asymptotic optimization and generalization guarantees for SGN for deep and wide feedforward networks in the near-initialization regime. To the best of our knowledge, this is the

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

first theoretical study that establishes stability-based (path-dependent) generalization bounds for Gauss-Newton in deep learning. Our main contributions include the following:

• **Non-asymptotic optimization guarantees for SGN.** We prove that the stochastic Gauss-Newton method with Levenberg-Marquardt damping, mini-batch sampling and metric projection achieves an optimality gap

$$\mathcal{O}\left(\frac{1}{k}\left[\bar{r}_k \log(Bk) + \lambda + \lambda^{-1}\right] + \frac{1}{\sqrt{m}}\right)$$

relative to the best in-class predictor, where k is the number of iterations, B the batch size, m is the network width, λ is the Levenberg-Marquardt damping factor, and $\bar{r}_k \leq p$ is the rank of the average Jacobian covariance for a p-dimensional parameter space.

• **Robustness to ill-conditioning.** Our *optimization* bounds (see Section 4) do not require strict positive definiteness of the neural tangent kernel, indicating that SGN remains effective even under highly ill-conditioned kernels unlike existing analyses of GN methods. Moreover, the optimization guarantees avoid any polynomial dependence of the network size m on the sample size n.

• **Generalization bounds for SGN.** We establish non-asymptotic and *algorithm-dependent* generalization bounds for SGN via the concept of algorithmic stability in the sense of (Bousquet and Elisseeff, 2002), and show that SGN is stable under appropriately-chosen damping and learning rate choices, overparameterization and kernel non-degeneracy (see Theorem 2). Here, the damping factor λ governs a trade-off between optimization and generalization performance. We also identify benign cases, where persistence of excitation (i.e., well-conditioned preconditioners) imply a favorable generalization and optimization performance. Stability analysis of SGN is challenging due to its data- and path-dependent preconditioner. We address this by establishing approximate co-coercivity and metric perturbation bounds in the wide-network regime. Our analysis explicitly characterizes the impacts of Gauss-Newton preconditioning, damping factor, overparameterization, training duration and batch size on the generalization performance.

1.2 Related Works

Convergence of GN in Deep Learning. Gauss-Newton and its variants are classical tools for nonlinear least squares (Nocedal and Wright, 2006; Gratton et al., 2007; Bertsekas, 1996) and have inspired practical second-order routines in deep learning (Botev et al., 2017). However, a concrete theoretical understanding of GN methods in deep learning is still in a nascent stage, and there has been a surge of interest recently to analyze the convergence of Gauss-Newton variants in deep learning (Cayci, 2024; Arbel et al., 2023; Cai et al., 2019; Zhang et al., 2019), following the NTK analyses in (Chizat et al., 2019; Du et al., 2018). These works typically require (i) a lower bound on the NTK/Gram spectrum, and (ii) polynomial overparameterization $m = \Omega(n^q)$, $q \geq 2$ for convergence bounds. Our optimization results establish non-asymptotic bounds on the optimality gap for stochastic Gauss-Newton (i) without assuming NTK conditioning (i.e., data separation), and (ii) without massive overparameterization. On the technical side, to capture stochastic and path-dependent metric in SGN, we devise a new variable-metric analysis in parameter space for Gauss-Newton iterations, while existing works (Cai et al., 2019; Zhang et al., 2019; Arbel et al., 2023; Cayci, 2024) utilize a function space analysis.

Generalization of Gauss-Newton. The impact of curvature-aware preconditioning on generalization in deep learning has been a focal point of interest (Amari et al., 2020; Zhang et al., 2019; Arbel et al., 2023). The implicit bias of Gauss-Newton was studied by Arbel et al. (2023) empirically without any quantitative generalization bounds. The existing generalization bounds are either (i) algorithm-agnostic (e.g., bounds based on Rademacher complexity that solely depend on the hypothesis class rather than the training rule (Zhang et al., 2019)), or (ii) asymptotic comparisons that analyze the risk of the limiting solution via bias-variance decompositions (in ridgeless regression/RKHS) (Amari et al., 2020), and therefore do not yield finite-sample, finite-time bounds that track optimization trajectories in deep networks. Unlike the existing works, we derive generalization bounds for SGN that *track* the optimization path of SGN.

Algorithmic stability. Algorithmic stability framework is due to Bousquet and Elisseeff (2002), and has been improved and refined by Feldman and Vondrak (2019); Klochkov and Zhivotovskiy (2021). Most applications of the algorithmic stability framework in machine learning focus on first-order methods (e.g., SGD), where stability analyses are used for generalization bounds (Hardt et al., 2016; Lei and Ying, 2020; Bassily et al., 2020; Hellström et al., 2025). However, these results are not applicable for curvature-aware, preconditioned second-order methods, such as SGN, as the parameters are path-dependent (non-Markovian) and evolve under time- and data-dependent metrics, which we address in this work.

1.3 Notation

For $\boldsymbol{x} \in \mathbb{R}^d, \rho > 0$, $\mathcal{B}_2(\boldsymbol{x}, \rho) = \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \rho\}$. $d_H(\boldsymbol{s}, \boldsymbol{s'})$ denotes the Hamming distance between two vectors $\boldsymbol{s}, \boldsymbol{s'} \in \mathbb{R}^n$. For any $n \in \mathbb{Z}_+$, $[n] := \{1, 2, \ldots, n\}$. For any symmetric positive definite matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ and $w \in \mathbb{R}^n$, $\|\boldsymbol{w}\|_{\boldsymbol{M}} := \sqrt{w^\top \boldsymbol{M} w}$, and $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ denotes the minimum and maximum eigenvalue of $\boldsymbol{M}$, respectively.

2 PROBLEM SETUP

2.1 Supervised Learning Problem

Let $\mathscr{X} \subset \mathbb{R}^d$ be the input space, $\mathscr{Y} \subset \mathbb{R}$ be the output space, and $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$ be the sample space. Let $\mathcal{P}$ be the sample distribution over $\mathscr{Z}$, and let $\mu(\cdot) = \int_{\mathscr{Y}} \mathcal{P}(\cdot, dy)$ be the input distribution. Let $\boldsymbol{z} := (\boldsymbol{x}, y) \sim \mathcal{P}$ be an input-output pair. We assume $\mathcal{P}$ has a compact support such that

$$\|\boldsymbol{x}\|_2 \leq 1 \quad \text{and} \quad |y| \leq 1,$$

almost surely. Given a predictor $\varphi : \mathbb{R}^d \to \mathbb{R}$, the loss at $\boldsymbol{z} = (\boldsymbol{x}, y)$ is denoted as $\ell(\varphi(\boldsymbol{x}), \boldsymbol{z})$. The goal in supervised learning is to minimize the population risk

$$\mathcal{R}(\varphi) := \mathbb{E}_{\boldsymbol{z} \sim \mathcal{P}} \left[\ell(\varphi(\boldsymbol{x}), \boldsymbol{z})\right]. \tag{SL}$$

Assumption 1. *For any $\boldsymbol{z} \in \mathscr{Z}$, the loss function $y' \mapsto \ell(y', \boldsymbol{z})$ is twice-differentiable, ν-strongly convex for some $\nu > 0$, and (locally) β_K-Lipschitz within any compact $K \subset \mathbb{R}$ for some $\beta_K > 0$.*

For example, for nonlinear least-squares regression, the loss function $\ell(y', \boldsymbol{z}) = \frac{1}{2}(y - y')^2$ satisfies Assumption 1 with $\nu = 1$ and $\beta_K = 1 + \sup_{u \in K} |u|$.

Given a sequence of independent and identically distributed samples $\boldsymbol{z}_i := (\boldsymbol{x}_i, y_i) \overset{iid}{\sim} \mathcal{P}$, $i \in [n]$, we denote the training set by $\mathcal{S} := \{\boldsymbol{z}_i : i \in [n]\}$ and define the empirical risk as

$$\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi) := \frac{1}{n} \sum_{i=1}^{n} \ell(\varphi(\boldsymbol{x}_i), \boldsymbol{z}_i).$$

Given a parametric class $\mathcal{F}_\Theta := \{\varphi_\theta : \theta \in \Theta\}$ of predictors, the ERM problem is

$$\min_{\theta \in \Theta} \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_\theta). \tag{ERM}$$

For a learning algorithm A with output $A(\mathcal{S}) \in \mathcal{F}_\Theta$, the optimization error is defined by

$$\mathrm{opt}_{\mathcal{F}_\Theta}\big(A(\mathcal{S}); \mathcal{S}\big) := \widehat{\mathcal{R}}_{\mathcal{S}}(A(\mathcal{S})) - \inf_{\tilde{\varphi} \in \mathcal{F}_\Theta} \widehat{\mathcal{R}}_{\mathcal{S}}(\tilde{\varphi}).$$

A fundamental quantity in supervised learning is the generalization error (Hellström et al., 2025):

$$\mathrm{gen}(\varphi; \mathcal{S}) := \mathcal{R}(\varphi) - \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi), \quad \varphi \in \mathcal{F}_\Theta.$$

Remark 1 (Error decomposition)**.** *For a given class of predictors $\mathcal{H}$, let A be a learning algorithm and let $A(\mathcal{S}) \in \mathcal{H}$ denote the predictor returned by A on the training set $\mathcal{S}$. Then, the expected excess risk of $A(\mathcal{S})$ is bounded as follows:*

$$\begin{aligned} &\mathbb{E}[\mathcal{R}(A(\mathcal{S})) - \mathcal{R}(\varphi^\star)] \\ &\qquad \leq \mathbb{E}\Big[\mathrm{gen}\big(A(\mathcal{S}); \mathcal{S}\big)\Big] + \mathbb{E}\Big[\mathrm{opt}_{\mathcal{H}}\big(A(\mathcal{S}); \mathcal{S}\big)\Big], \end{aligned} \tag{1}$$

where $\varphi^\star \in \arg\min_{\varphi \in \mathcal{H}} \mathcal{R}(\varphi)$ is the population risk minimizer in $\mathcal{H}$.

An empirical risk minimizer with low generalization error effectively solves (SL) by attaining low population risk (Hellström et al., 2025). Given a learned predictor $A(\mathcal{S}) = \varphi_{\widehat{\theta}_{\mathcal{S}}}$, the fundamental challenge is to bound $\mathbb{E}_{\mathcal{S} \sim \mathcal{P}^n}[\mathrm{gen}(\varphi_{\widehat{\theta}_{\mathcal{S}}}; \mathcal{S})]$. Traditionally, uniform convergence bounds with Rademacher complexity (Bartlett and Mendelson, 2002) and VC dimension (Anthony and Bartlett, 2009) have been used to establish generalization bounds via $|\mathrm{gen}(A(\mathcal{S}); \mathcal{S})| \leq \sup_{\varphi \in \mathcal{H}} |\mathrm{gen}(\varphi; \mathcal{S})|$. However, these bounds are *algorithm-agnostic* (i.e., insensitive to the learning algorithm), and therefore fail to capture essential dynamics of overparameterization (Zhang et al., 2021).

2.2 Deep Neural Networks

We consider a feedforward deep neural network of depth $H \geq 1$ and width m with a smooth activation function $\sigma : \mathbb{R} \to \mathbb{R}$, which satisfies

$$\|\sigma\|_\infty \leq \sigma_0, \ \|\sigma'\|_\infty \leq \sigma_1, \ \text{and} \ \|\sigma''\|_\infty \leq \sigma_2.$$

Note that many widely-used activation functions, including tanh (with $\sigma_0 = 1, \sigma_1 = 2, \sigma_2 = 2$) and sigmoid function, satisfy this condition.

We consider the following model: let $c \in \mathbb{R}^m$, $W^{(1)} \in \mathbb{R}^{m \times d}$ and $W^{(h)} \in \mathbb{R}^{m \times m}$ for $h = 2, 3, \ldots, H$, and

$$\boldsymbol{W} := (W^{(1)}, \ldots, W^{(H)}).$$

Then, for an input $\boldsymbol{x}_j \in \mathbb{R}^d$, the neural network is defined recursively as

$$\begin{aligned} &\boldsymbol{x}_j^{(h)}(\boldsymbol{W}) = \frac{1}{\sqrt{m}} \cdot \vec{\sigma}\Big(W^{(h)} \boldsymbol{x}_j^{(h-1)}(\boldsymbol{W})\Big), \quad h \in [H], \\ &\varphi(\boldsymbol{x}_j; w) = c^\top \boldsymbol{x}_j^{(H)}(\boldsymbol{W}) \end{aligned}$$

where $\boldsymbol{x}_j^{(0)}(\boldsymbol{W}) = \boldsymbol{x}_j$, $w = \mathrm{vec}(\boldsymbol{W}, c)$ is the parameter vector, and $\vec{\sigma}(z) = [\sigma(z_1) \ \ldots \ \sigma(z_m)]^\top$. For each parameter w, we denote the corresponding predictor by $\varphi_w(\,\cdot\,) := \varphi(\,\cdot\,; w)$.

This parameterization, along with $\mathcal{O}(1/\sqrt{m})$ scaling, is standard in the neural tangent kernel literature

(Du et al., 2018; Chizat et al., 2019). We note that our analysis extends to general output scaling factors $\mathcal{O}(1/m^{\zeta})$ for any $\zeta > 0$ except Corollary 1 that necessitates $\zeta = 0.5$.

Let p be the number of parameters of the neural network, and for a compact and convex $\mathcal{C} \subset \mathbb{R}^p$, let

$$\overline{\mathrm{Lip}}_{\nabla\varphi,\mathcal{C}} := \sup_{w\in\mathcal{C}} \sup_{\boldsymbol{x}\in\mathcal{B}_2(0,1)} \|\nabla_w^2 \varphi(\boldsymbol{x}; w)\|_2,$$
$$\overline{\mathrm{Lip}}_{\varphi,\mathcal{C}} := \sup_{w\in\mathcal{C}} \sup_{\boldsymbol{x}\in\mathcal{B}_2(0,1)} \|\nabla_w \varphi(\boldsymbol{x}; w)\|_2.$$

The following elementary result will be useful in both the generalization and optimization bounds.

Lemma 1. *For any compact and convex $\mathcal{C} \subset \mathbb{R}^p$ with*

$$\sup_{w,w'\in\mathcal{C}} \|w - w'\|_2 =: r_{\mathcal{C}}, \tag{2}$$

let

$$\kappa_0 := \max_{h\in[H]} \frac{\|W_0^{(h)}\|_2}{\sqrt{m}} \text{ and } \zeta_0 := \|c_0\|_2.$$

Also, let $\kappa_{\mathcal{C}} = \kappa_0 + \frac{r_{\mathcal{C}}}{\sqrt{m}}$ and $\zeta_{\mathcal{C}} := \zeta_0 + r_{\mathcal{C}}$, and

$$L_{\mathcal{C}} := \max\{1, \sigma_1, \kappa_{\mathcal{C}}, \sigma_1\kappa_{\mathcal{C}}\}.$$

We have the following (local) Lipschitz continuity results in $\mathcal{C}$:

$$\overline{\mathrm{Lip}}_{\varphi,\mathcal{C}} \le \sigma_0 + \zeta_{\mathcal{C}}\sigma_0 \frac{HL_{\mathcal{C}}^H}{\sqrt{m}} =: \mathrm{Lip}_{\varphi,\mathcal{C}},$$

and

$$\overline{\mathrm{Lip}}_{\nabla\varphi,\mathcal{C}} \le \mathrm{Lip}_{\nabla\varphi,\mathcal{C}},$$

where

$$\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} := 2\sigma_0 \frac{HL_{\mathcal{C}}^H}{\sqrt{m}} + 8\zeta_{\mathcal{C}}(1 + \sigma_0 + \sigma_0^2\sigma_2) \frac{H^3 L_{\mathcal{C}}^{2H}}{\sqrt{m}}.$$

Hence, Lemma 1 implies that

$$\mathrm{Lip}_{\varphi,\mathcal{C}} = \mathcal{O}\Big(1 + \frac{HL_{\mathcal{C}}^H}{\sqrt{m}}\Big) \text{ and } \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} = \mathcal{O}\Big(\frac{H^3 L_{\mathcal{C}}^{2H}}{\sqrt{m}}\Big).$$

Under Assumption 1, we define $\mathrm{Lip}_{\ell,\mathcal{C}} := \beta_K$ where $K = [0, \sup_{w\in\mathcal{C},\boldsymbol{x}\in\mathcal{B}_2(0,1)} |\varphi(\boldsymbol{x}; w)|]$.

In the following section, we present SGN that we study, based on (Bertsekas, 1996; Martens, 2020).

3 STOCHASTIC GAUSS-NEWTON FOR ERM

Let $\mathcal{C} \subset \mathbb{R}^p$ be a closed and convex subset. Given a training set $\mathcal{S} = \{\boldsymbol{z}_i : i \in [n]\}$, for a fixed batch size $B \le n$, let $\{I_k : k \in \mathbb{N}\}$ be an i.i.d. sequence of random subsets of $[n]$, each drawn uniformly from $\{\mathcal{M} \subset [n] : |\mathcal{M}| = B\}$. Let

$$\boldsymbol{J}_k(w) := [\nabla_w^\top \varphi(\boldsymbol{x}_j; w)]_{j\in I_k} \in \mathbb{R}^{B\times p},\ k \in \mathbb{N}$$

be the Jacobian,

$$G_k(w) := [\ell'(\varphi(\boldsymbol{x}_j; w); \boldsymbol{z}_j)]_{j\in I_k} \in \mathbb{R}^B$$

be the gradient of the loss function with respect to the prediction, and

$$\Phi_k(w) := \sum_{j\in I_k} \ell(\varphi(\boldsymbol{x}_j; w); \boldsymbol{z}_j)$$

be the loss evaluated on I_k. Then, the stochastic gradient on I_k is

$$\Psi_k(w) := \nabla\Phi_k(w) = \boldsymbol{J}_k^\top(w) G_k(w).$$

Given $\lambda, \alpha > 0$, we consider the following preconditioner, which is an incremental version of the Gauss-Newton preconditioner under stochastic subsampling (Bertsekas, 1996; Martens, 2020):

$$\boldsymbol{H}_k = \alpha \sum_{t=0}^{k} \boldsymbol{J}_t^\top(w_t) \boldsymbol{J}_t(w_t) + \lambda \boldsymbol{I}.$$

For an initial parameter $w_0 \in \mathcal{C}$, for any $k = 0, 1, \ldots$, the (stochastic) Gauss-Newton update is

$$\begin{aligned} u_k &= w_k - \eta \boldsymbol{H}_k^{-1} \Psi_k(w_k), \\ w_{k+1} &= \pi_{\mathcal{C}}^{\boldsymbol{H}_k}(u_k), \end{aligned} \tag{3}$$

where $\pi_{\mathcal{C}}^{\boldsymbol{H}_k}(w) = \arg\min_{u\in\mathcal{C}} \|u - w\|_{\boldsymbol{H}_k}^2$. SGN returns the predictor $\varphi_{\widehat{w}_k} \in \mathcal{F}_{\mathcal{C}}$ after $k \ge 1$ iterations, where

$$\mathcal{F}_{\mathcal{C}} := \{\varphi_w : w \in \mathcal{C}\},$$

and $\widehat{w}_k := \frac{1}{k}\sum_{t=0}^{k-1} w_t$ is the Polyak-Ruppert average of the SGN iterates.

In view of the error decomposition in (1), we study for SGN both the optimization error $\mathrm{opt}_{\mathcal{F}_{\mathcal{C}}}(\varphi_{\widehat{w}_k}; \mathcal{S})$ and the generalization error $\mathrm{gen}(\varphi_{\widehat{w}_k}; \mathcal{S})$.

Remark 2. The Gauss-Newton method can be interpreted in multiple ways, in both Euclidean and Riemannian geometries (Nocedal and Wright, 2006; Cayci, 2024). Deviating from the classical analysis approach that interprets Gauss-Newton as an approximate Newton method (see (Nocedal and Wright, 2006; Bertsekas, 1996)), we adopt a variable-metric approach building on Duchi (2018); Hazan et al. (2007):

$$w_{k+1} = \arg\min_{w\in\mathcal{C}} \left\{ \Psi_k^\top(w_k)(w - w_k) + \frac{1}{\eta}\mathcal{D}_k(w, w_k) \right\},$$

where the Bregman divergence is

$$\mathcal{D}_k(w, w_k) = \frac{1}{2}(w - w_k)^\top \boldsymbol{H}_k (w - w_k),\ k \in \mathbb{N}.$$

Note that $\mathcal{D}_k$ is akin to the (generalized) Mahalanobis distance in mirror descent, yet it is more complicated as it is (i) parameter- and path-dependent, and (ii) stochastic to capture curvature information. Note that $\boldsymbol{H}_k$ is incrementally updated using the Gram matrices $\boldsymbol{J}_k^\top(w_k)\boldsymbol{J}_k(w_k)$ throughout the optimization path, reminiscent of the adaptive gradient methods such as AdaGrad (Duchi et al., 2011) and Shampoo (Gupta et al., 2018; Morwani et al., 2025); however, SGN uses an approximation $\boldsymbol{J}_k^\top(w_k)\boldsymbol{J}_k(w_k)$ of the Hessian $\nabla_w^2\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{w_k})$ instead of gradient outer products, leading to significantly different dynamics.

Remark 3 (Beyond quadratic loss)**.** The stochastic Gauss-Newton preconditioner $\boldsymbol{H}_k$ is derived for the squared loss $\ell(f,\boldsymbol{z}) = \frac{1}{2}(f-y)^2$. Our analysis in this paper applies to a broader class of loss functions that satisfy Assumption 1, which includes the squared loss ($\nu = 1$). Extending the theory to more general preconditioners that incorporate $\nabla_f^2\ell(f,\boldsymbol{z})\big|_{f=\varphi(\boldsymbol{x};w)}$ is an interesting direction for future work.

4 OPTIMIZATION BOUNDS FOR SGN: A NON-ASYMPTOTIC ANALYSIS

In this section, we establish finite-time and finite-width optimization bounds for SGN applied to (ERM) with deep networks. In particular, we prove (i) global near-optimality over $\mathcal{C}$, and (ii) global optimality in the neural tangent kernel regime for sufficiently wide neural networks.

4.1 Finite-Time Optimization Bounds for SGN

Recall that $\mathcal{C} \subset \mathbb{R}^p$ is a compact and convex parameter set[1] with diameter $r_{\mathcal{C}}$. Let $\mathcal{F}_{\mathcal{C}} = \{\varphi_w : w \in \mathcal{C}\}$ denote the hypothesis class induced by the parameter set $\mathcal{C}$. We study the optimization error

$$\mathrm{opt}_{\mathcal{F}_{\mathcal{C}}}(\varphi_{\widehat{w}_k};\mathcal{S}) := \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{\widehat{w}_k}) - \inf_{\tilde{\varphi}\in\mathcal{F}_{\mathcal{C}}} \widehat{\mathcal{R}}_{\mathcal{S}}(\tilde{\varphi})$$

of the predictor induced by the Polyak-Ruppert average of the SGN iterates.

Theorem 1 (Optimization Guarantees for SGN)**.** *Let*

$$V_k := \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{w_k}) - \inf_{w\in\mathcal{C}} \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_w)$$

be the optimality gap at iteration k of SGN, and define

$$\gamma := \frac{\lambda}{\alpha\cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2 \cdot B} \quad \textit{and} \quad \xi := \frac{\eta}{\alpha}. \tag{4}$$

Then, with any choice of (η,α,λ) such that $\gamma > 0$ and $\xi \geq \frac{2}{\nu}$, SGN satisfies

$$\begin{aligned}\frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}[V_t] &\leq \frac{\mathrm{Lip}_{\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^2(\gamma+3)}{2k\xi} \\ &+ \frac{\xi}{2k}\mathrm{Lip}_{\ell,\mathcal{C}}^2\left(\mathbb{E}\log\frac{\det \boldsymbol{H}_k}{\det \boldsymbol{H}_0} + \frac{1}{\gamma+1}\right) \\ &+ \frac{1}{4\xi}r_{\mathcal{C}}^4\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 + \frac{1}{2}\mathrm{Lip}_{\ell,\mathcal{C}}r_{\mathcal{C}}^2\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} =: \kappa_{k,m}.\end{aligned}$$

Let $\widehat{w}_k := k^{-1}\sum_{t=0}^{k-1} w_t$ be the Polyak-Ruppert average of the SGN iterates. Then,

$$\mathbb{E}\big[\mathrm{opt}_{\mathcal{F}_{\mathcal{C}}}(\varphi_{\widehat{w}_k};\mathcal{S})\big] \leq \kappa_{k,m} + \mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}r_{\mathcal{C}}^2$$

for any $k \in \mathbb{N}$.

Remark 4. Theorem 1 and Lemma 1 imply

$$\frac{1}{k}\sum_{t<k}\mathbb{E}[V_t] \lesssim \frac{1}{k}\left(\frac{\gamma}{\xi} + \frac{\xi}{\gamma} + \xi\bar{r}_k\log(Bk)\right) + \frac{1}{m\xi} + \frac{1}{\sqrt{m}},$$

where $\bar{r}_k \leq p$ is specified in Proposition 1, and the last two terms come from $\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} = \mathcal{O}(m^{-1/2})$ by Lemma 1, which decay with overparameterization. Similar width-dependent terms that vanish in the overparameterized regime also appear in prior work (Cayci and Eryilmaz, 2025; Ji and Telgarsky, 2020). In contrast to Cai et al. (2019); Zhang et al. (2019), Theorem 1 does *not* require positive definiteness of the kernel matrix $\mathbb{E}[\boldsymbol{J}_0(w_0)\boldsymbol{J}_0^\top(w_0)]$. Moreover, the bound does not depend on the smoothness constant of ℓ, indicating robustness of SGN in regimes with ill-conditioned kernels and losses with large smoothness constants.

The proof of Theorem 1 is provided in Appendix A. On the technical side, our approach deviates significantly from the existing analyses (Cayci, 2024; Arbel et al., 2023; Zhang et al., 2019): in order to capture both the stochasticity and the incremental nature of the preconditioner along the optimization path, we analyze the SGN method in parameter space via a time-varying Lyapunov function, $\mathscr{L}_k(w) := \|w-\bar{w}\|_{\boldsymbol{H}_k}^2$, bringing analytical tools from adaptive gradient methods (Duchi, 2018; Hazan et al., 2007), instead of function space analyses (Arbel et al., 2023; Zhang et al., 2019; Cayci, 2024). Crucially, the Jacobian-based preconditioner $\boldsymbol{H}_k$ necessitates a novel analysis beyond standard adaptive methods.

In the following, we characterize the $\log\det$ term.

Proposition 1. *For any compact and convex $\mathcal{C} \subset \mathbb{R}^p$,*

$$\log\frac{\det \boldsymbol{H}_k}{\det \boldsymbol{H}_0} \leq p\log\left(1 + \frac{\alpha}{\lambda}(k+1)B\mathrm{Lip}_{\varphi,\mathcal{C}}^2\right),\ k\in\mathbb{N}.$$

Furthermore, let $\bar{\Sigma}_k := \frac{1}{k+1}\sum_{t=0}^{k}\mathbb{E}[\Sigma(w_t)]$, where

$$\Sigma(w) := \frac{1}{n}\sum_{j=1}^{n}\nabla\varphi(\boldsymbol{x}_j;w)\nabla^\top\varphi(\boldsymbol{x}_j;w).$$

[1] Since $\mathcal{C}$ is a compact subset of $\mathbb{R}^p$ and $w \mapsto \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_w)$ is a continuous function, Weierstrass' extreme value theorem (Theorem 4.4.2 in (Abbott, 2015)) implies that $\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_w)$ attains its minimum over $\mathcal{C}$.

Then, for $\bar{r}_k := \operatorname{rank}(\bar{\Sigma}_k) \le p$, *we have*

$$\mathbb{E} \log \frac{\det \boldsymbol{H}_k}{\det \boldsymbol{H}_0} \le \bar{r}_k \log \Big(1 + \frac{\alpha}{\lambda} B(k+1) \operatorname{tr}(\bar{\Sigma}_k)\Big).$$

4.2 Global Near-Optimality in the Neural Tangent Kernel Regime

Theorem 1 indicates that SGN achieves near-optimality within the parameter set $\mathcal{C}$. In this subsection, we show that for sufficiently large $\mathcal{C}$, SGN can attain a globally optimal predictor for $\widehat{\mathcal{R}}_{\mathcal{S}}$ up to an arbitrarily small error within a rich class of predictors.

First, we define a rich function class $\mathcal{F}_{ntk}$ that corresponds to the infinite-width limit of shallow and wide neural networks (Cayci and Eryilmaz, 2025).

NTK Function Class. Let

$$\mathcal{G} := \Big\{ (v_c, v_W) : \mathbb{R}^{d+1} \to \mathbb{R} \times \mathbb{R}^d : \sup_{\omega \in \mathbb{R}^{d+1}} |v_c(\omega)| \le \bar{v}_c, \sup_{\omega \in \mathbb{R}^{d+1}} \|v_W(\omega)\|_2 \le \bar{v}_W \Big\}.$$

Then, for $\boldsymbol{v} := (v_c, v_W)$, we define

$$\mathcal{F}_{ntk} := \{ \boldsymbol{x} \mapsto \mathbb{E}_{\omega \sim \mathcal{N}(0, \boldsymbol{I}_{d+1})} \langle \boldsymbol{v}(\omega), \phi(\boldsymbol{x}; \omega) \rangle : \boldsymbol{v} \in \mathcal{G} \},$$

where

$$\phi(\boldsymbol{x}; \omega) = (\sigma(\langle W, \boldsymbol{x} \rangle),\ c\boldsymbol{x}\sigma'(\langle W, \boldsymbol{x} \rangle))$$

is the random feature for $\omega = (c, W) \in \mathbb{R}^{d+1}$. The completion of the function class induced by transportation mappings $\boldsymbol{v} : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ satisfying $\mathbb{E}\|\boldsymbol{v}(\omega)\|_2^2 < \infty$ is the reproducing kernel Hilbert space (RKHS) of the NTK for a single hidden-layer network (Cayci and Eryilmaz, 2025; Ji and Telgarsky, 2020), which is dense in the space of continuous functions on a compact set (Ji et al., 2020).

The following result for single hidden-layer network shows that $f^\star \in \mathcal{F}_{ntk}$ can be learned by SGN with high probability over the random initialization.

Corollary 1 (Near-optimality in $\mathcal{F}_{ntk}$)**.** *Assume that there exists* $f^\star \in \mathcal{F}_{ntk}$ *such that* $\widehat{\mathcal{R}}_{\mathcal{S}}(f^\star) = 0$. *Consider the random initialization* $w_0 = [u_0^{(i)}]_{i \in [m]}$, *where*

$$\begin{aligned} &(c_0^{(i)}, W_0^{(i)}) \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_{d+1}), \\ &c_0^{(i+m/2)} = -c_0^{(i)} \quad \text{and} \quad W_0^{(i+m/2)} = W_0^{(i)} \end{aligned} \tag{5}$$

for $i = 1, 2, \ldots, m/2$. *Let*

$$\rho_{\boldsymbol{v}} := \sqrt{\bar{v}_c^2 + \bar{v}_W^2},$$

and $\mathcal{C} := \mathcal{B}_2(w_0, \rho_{\boldsymbol{v}}) \subset \mathbb{R}^{m(d+1)}$. *Then, for any* $\delta \in (0,1)$, *with probability at least* $1 - \delta$ *over the random initialization, SGN with* $\xi = \eta/\alpha \ge 2/\nu$ *and* $\lambda = \gamma \alpha \operatorname{Lip}_{\varphi,\mathcal{C}}^2 B$ *satisfies*

$$\mathbb{E}_0 \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{\widehat{w}_k}) \lesssim C_{n,\delta,\boldsymbol{v}} \Big(\frac{\bar{r}_k \log k}{k} + \frac{1}{\sqrt{m}} \Big)$$

for any $k \ge 1$, *where* $\widehat{w}_k$ *is the Polyak-Ruppert average of the SGN iterates,* $C_{n,\delta,\boldsymbol{v}} = \operatorname{poly}\big(\log(n/\delta), \rho_{\boldsymbol{v}}\big)$, *and* $\mathbb{E}_0[\cdot] := \mathbb{E}[\cdot | w_0]$.

The proof of Corollary 1 with exact constants can be found in Appendix A.

5 GENERALIZATION BOUNDS FOR SGN: ALGORITHMIC STABILITY

In this section, we establish finite-sample and finite-width generalization bounds for deep neural networks trained by SGN from the perspective of *algorithmic stability* (Bousquet and Elisseeff, 2002; Hardt et al., 2016). The intuition behind the algorithmic stability is that learning algorithms with a weak dependency on the specific training data (e.g., nearly-insensitive to an arbitrary change in one sample point in the training set) generalize well to test data (Hellström et al., 2025).

We first provide a concise mathematical description of the algorithmic stability concept, and then present our generalization bounds.

5.1 Algorithmic Stability and Generalization Bounds

We first present an overview of algorithmic stability (Hardt et al., 2016; Bousquet and Elisseeff, 2002), which will constitute the basis of our generalization bounds.

For a given training set $\boldsymbol{s} \in \mathscr{Z}^n$, let $A(\boldsymbol{s})$ be the output of a randomized learning algorithm. Let $\mathcal{S} = \{\boldsymbol{z}_i\}_{i=1}^n, \mathcal{S}' = \{\boldsymbol{z}_i'\}_{i=1}^n \sim \mathcal{P}^n$ be independent training sets, and

$$\mathcal{S}^{(i)} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{i-1}, \boldsymbol{z}_i', \boldsymbol{z}_{i+1}, \ldots, \boldsymbol{z}_n),\ i \in [n].$$

If, for all $i \in [n]$,

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'} \sup_{\boldsymbol{z} \in \mathscr{Z}} \mathbb{E}_A \big[|\ell(A(\mathcal{S}); \boldsymbol{z}) - \ell(A(\mathcal{S}^{(i)}); \boldsymbol{z})| \big| \mathcal{S}, \mathcal{S}' \big] \le \epsilon_i,$$

then it follows that

$$|\mathbb{E}_{A,\mathcal{S}} [\operatorname{gen}(A(\mathcal{S}); \mathcal{S})]| \le \frac{1}{n} \sum_{i=1}^n \epsilon_i \tag{6}$$

by the proof of Theorem 2.2 in Hardt et al. (2016). In SGN, the randomness of the learning algorithm stems from the subsampling process $\{I_k : k \in \mathbb{N}\}$ since $A(\boldsymbol{s}) = \varphi_{\widehat{w}_k}$ is the predictor returned by SGN.

5.2 Algorithmic Stability of SGN

Let $\boldsymbol{s}, \boldsymbol{s}' \in \mathscr{Z}^n$ be two training sets such that $d_H(\boldsymbol{s}, \boldsymbol{s}') = 1$, and let $j^\star \in [n]$ denote the sample index where $\boldsymbol{s}$ and $\boldsymbol{s}'$ differ, i.e., $\boldsymbol{z}_{j^\star} \neq \boldsymbol{z}'_{j^\star}$. In the following, $\{(w_k, \widehat{w}_k, \Phi_k, \boldsymbol{J}_k, G_k, \Psi_k, \boldsymbol{H}_k) : k \in \mathbb{N}\}$ and $\{(w'_k, \widehat{w}'_k, \Phi'_k, \boldsymbol{J}'_k, G'_k, \Psi'_k, \boldsymbol{H}'_k) : k \in \mathbb{N}\}$ denote the parameters and related mappings in SGN trained on $\boldsymbol{s}$ and $\boldsymbol{s}'$, respectively, using the same trajectory of the subsampling process $\{I_k : k \in \mathbb{N}\}$ and from the same initial condition $w_0 = w'_0$. The following yields a stability bound on the average iterate, which is the output of SGN.

Lemma 2 (Stability with midpoint metric)**.** *For any $k \in \mathbb{N}$, let $\Delta_k := w_k - w'_k$, and define the midpoint metric*

$$\bar{\boldsymbol{H}}_k := \frac{1}{2}(\boldsymbol{H}_k + \boldsymbol{H}'_k). \tag{7}$$

For any $\boldsymbol{s}, \boldsymbol{s}' \in \mathscr{Z}^n$ such that $d_H(\boldsymbol{s}, \boldsymbol{s}') = 1$, we have

$$\sup_{\boldsymbol{z} \in \mathscr{Z}} \mathbb{E}_{(I_t)_{t \le k}} |\ell(\varphi(\boldsymbol{x}; \widehat{w}_k); \boldsymbol{z}) - \ell(\varphi(\boldsymbol{x}; \widehat{w}'_k); \boldsymbol{z})| \le \frac{\mathrm{Lip}_{\ell,\mathcal{C}} \mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}} \frac{1}{k} \sum_{t=1}^{k} \mathbb{E}\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}},$$

where $\widehat{w}_k := \frac{1}{k}\sum_{t<k} w_t$ is the Polyak-Ruppert average of the SGN iterates.

Throughout the rest of this section, we establish finite-sample bounds on $\mathbb{E}\|w_k - w'_k\|_{\bar{\boldsymbol{H}}_{k-1}}$, which implies the stability of SGN together with Lemma 2.

We make the following assumption for the generalization bounds. Let $\boldsymbol{J}(w) := [\nabla^\top \varphi(\boldsymbol{x}_j; w)]_{j \in [n]} \in \mathbb{R}^{n \times p}$.

Assumption 2. *There exists $\mu_0 > 0$ such that, for both $\boldsymbol{s}, \boldsymbol{s}'$, $\boldsymbol{J}(w)\boldsymbol{J}^\top(w) \succeq \mu_0^2 \boldsymbol{I}$ for all $w \in \mathcal{C}$.*

Assumption 2 is a natural non-degeneracy condition in the lazy training literature. Under the standard assumption $\lambda_{\min}\big(\boldsymbol{J}(w_0)\boldsymbol{J}^\top(w_0)\big) =: \lambda_0^2 > 0$, which holds with high probability at random initialization for non-collinear data points and sufficiently wide networks, local Jacobian perturbation arguments (see, e.g., Chizat et al. (2019), Cayci (2024) and Du et al. (2018)) imply

$$\lambda_{\min}\big(\boldsymbol{J}(w)\boldsymbol{J}^\top(w)\big) \ge \lambda_0^2/4$$

uniformly over $\|w - w_0\|_2 \le r_{\mathcal{C}}$, provided $m \gtrsim r_{\mathcal{C}}^2 n/\lambda_0^2$.

Next, we establish the stability of SGN.

Theorem 2 (Stability of SGN)**.** *Let*

$$\varepsilon := B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} \quad \text{and} \quad \Lambda := B\mathrm{Lip}_{\varphi,\mathcal{C}}^2 + 2\varepsilon,$$

and $\{\lambda_t : t \in \mathbb{N}\}$[2] be such that $\lambda_{\min}(\boldsymbol{H}_t), \lambda_{\min}(\boldsymbol{H}'_t) \ge \lambda_t$ for all $t \ge 0$. Under Assumption 2, with the choices that satisfy

$$\frac{\eta}{\lambda} \le \frac{1}{2\Lambda} \quad \text{and} \quad \frac{\alpha}{\eta} \le \frac{\mu_0^2 \nu^2}{8B(\Lambda + \varepsilon)}, \tag{8}$$

we obtain, for each $k \ge 1$,

$$\mathbb{E}\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}} \lesssim \underbrace{k\sqrt{\eta B \mathrm{Lip}_{\nabla\varphi,\mathcal{C}}} + k\sqrt{\alpha} B^{\frac{3}{2}} \mathrm{Lip}_{\nabla\varphi,\mathcal{C}}}_{\text{approximate non-expansivity}} + \underbrace{\alpha B^2 \Big(\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{1}{n}\Big) \sum_{t=0}^{k-1} \frac{t+1}{\lambda_t^{\frac{1}{2}}}}_{\text{metric/preconditioner mismatch}} + \underbrace{\frac{\eta B^2}{n} \sum_{t=0}^{k-1} \lambda_t^{-\frac{1}{2}}}_{\text{gradient mismatch}}$$

A proof sketch for Theorem 2 is provided in Section 5.3, where each term above is explained. The complete proof can be found in Appendix B.

We have the following result since $\inf_{t \ge 0} \lambda_t \ge \lambda$ always holds.

Corollary 2 (Worst-case stability)**.** *Given $\lambda > 0$, let*

$$M := \max\left\{2/\nu, 8B\left(\Lambda + \varepsilon\right)/(\nu^2 \mu_0^2)\right\},$$

and choose $\alpha = \frac{1}{Bk}$ and $\eta = \xi\alpha$ for $\xi \in [M, Bk\lambda/(2\Lambda)]$. Then, for any $k \ge 1$,

$$\mathbb{E}\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}} \lesssim \sqrt{\xi k \mathrm{Lip}_{\nabla\varphi,\mathcal{C}}} + \sqrt{k} B \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{Bk}{\lambda^{\frac{1}{2}}}\Big(\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{1}{n}\Big) + \frac{B\xi}{n\lambda^{\frac{1}{2}}}.$$

Corollary 2 provides a finite-sample bound on $\mathbb{E}\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}}$. Substituting this bound into Lemma 2 and then invoking the stability-to-generalization argument in (6) yields the following expected generalization bound for SGN.

Corollary 3. *For $\mu > 0$, define the following event:*

$$E(\mathcal{S}) := \{\lambda_{\min}(\boldsymbol{J}(w)\boldsymbol{J}^\top(w)) \ge \mu,\ w \in \mathcal{C}\}.$$

Let $\mathcal{E}_i := E(\mathcal{S}) \cap E(\mathcal{S}^{(i)})$ for $i \in [n]$. Then,

$$\big|\mathbb{E}[\mathrm{gen}(\varphi_{\widehat{w}_k}; \mathcal{S})]\big| \le \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}} \frac{1}{nk} \sum_{i=1}^{n}\sum_{t=1}^{k} \mathbb{E}[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}} \mathbf{1}_{\mathcal{E}_i}] + \delta\Upsilon_{\mathcal{C}},$$

where $\delta := \max_{i \in [n]} \mathbb{P}(\mathcal{E}_i^c)$ and $\Upsilon_{\mathcal{C}} := \mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}} r_{\mathcal{C}}$.

For each $i \in [n]$, on $\mathcal{E}_i$, Theorem 2 and Corollary 2 can be used to bound $\frac{1}{k}\sum_{t=1}^{k} \mathbb{E}[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}} | \mathcal{S}, \mathcal{S}']$. Note that for shallow neural networks ($H = 1$), under the scaling and Lipschitz concentration assumptions on $\mathcal{P}$ in Assumptions 2.1 and 2.2 of Bombari et al. (2022), one can show $\delta \lesssim ne^{-c\log^2 m} + C\, e^{-c\log^2 n}$

[2] Since $\lambda_{\min}(\boldsymbol{H}_k), \lambda_{\min}(\boldsymbol{H}'_k) \ge \lambda > 0$ for all $k \in \mathbb{N}$, the existence of such a sequence is guaranteed.

and $\mu = \Omega(1)$ for sufficiently large m, building on Bombari et al. (2022); Cayci (2024). The data regularity assumptions are satisfied by a broad class of standard input distributions, including Gaussian measures, uniform laws on the sphere and on the hypercube, and, more generally, distributions obeying a dimension-independent log-Sobolev inequality (Bombari et al., 2022). A similar argument extends to deep neural networks ($H > 1$). In particular, lower bounds on $\lambda_{\min}(\boldsymbol{J}(w_0)\boldsymbol{J}^\top(w_0))$ for deep networks are available under standard data assumptions (Bombari et al., 2022; Karhadkar et al., 2024), and extending this lower bound uniformly to $w \in \mathcal{C}$ can be achieved by local Jacobian perturbation arguments in (Cayci, 2024; Du et al., 2019).

Remark 5 (Dual impact of Levenberg-Marquardt damping)**.** Note that (η, λ, α) in Corollary 2 satisfy the conditions of Theorems 1 and 2 for constant $\lambda > 0$ and yield

$$\mathbb{E}\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}} \lesssim \frac{\sqrt{\xi k}}{m^{1/4}} + \frac{Bk}{\lambda^{1/2}}\Big(\frac{1}{n} + \frac{1}{\sqrt{m}}\Big) + \frac{\sqrt{k}B}{\sqrt{m}} + \frac{B\xi}{n\lambda^{1/2}}.$$

In this case, a larger damping factor $\lambda > 0$ implies better conditioning for $\boldsymbol{H}_k$ and improves the generalization bounds for SGN. On the other hand, the optimization error increases with λ since $\gamma = \frac{\lambda}{\alpha \operatorname{Lip}^2_{\varphi,\mathcal{C}} B} \asymp \lambda k$ in Theorem 1. Thus, larger λ can hinder optimization with a non-vanishing error term of order $\mathcal{O}(\lambda)$, revealing a fundamental trade-off governed by λ. This illustrates the regularization effect of damping in SGN.

Remark 6 (Benefit of overparameterization)**.** Theorems 1 and 2 indicate that both optimization and stability bounds for SGN improve with m, demonstrating the benefits of overparameterization. The bound in Corollary 2 reveals a trade-off between training duration k and overparameterization: as k increases, the stability bound in Corollary 2 and thus the generalization bound in Corollary 3 deteriorate, but this effect can be compensated by taking the neural network width m sufficiently large.

Remark 7 (Benefit of cumulative preconditioning.)**.** An important factor that helps generalization is the cumulative Gauss-Newton term in the preconditioner. In benign cases where $\alpha \sum_{t=0}^{k} \boldsymbol{J}_t^\top(w_t)\boldsymbol{J}_t(w_t)$ is well-conditioned, favorable optimization and stability bounds can be simultaneously achieved.

The following assumption is analogous to persistence of excitation conditions in adaptive control and recursive least squares (Lai and Wei, 1982; Ioannou and Sun, 1996; Anderson and Taylor, 1979).

Condition 1 (Persistence of excitation)**.** *There exist constants $C, q > 0$, an integer $k_0 \in \mathbb{N}$, and an event $\mathcal{E} \in \sigma(I_t : t \in \mathbb{N})$ such that, on $\mathcal{E} \cap \mathcal{E}_i$,*

$$\lambda_{\min}\Big(\sum_{s=0}^{t} \boldsymbol{J}_s^\top(w_s)\boldsymbol{J}_s(w_s)\Big) \geq CB(t+1)^q, \qquad t \geq k_0,$$

for $i \in [n]$.

We have $\lambda_{\min}(\boldsymbol{H}_t) \geq \alpha CB(t+1)^q + \lambda$ for all $t \in \{k_0, \ldots, k\}$ on $\mathcal{E} \cap \mathcal{E}_i$. Using this in Theorem 2, we obtain the following stability bound.

Proposition 2 (Stability under PE)**.** *Under Condition 1 for $i \in [n]$,*

$$\mathbb{E}[\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}} \mathbf{1}_{\mathcal{E}\cap\mathcal{E}_i}] \lesssim \sqrt{\alpha} B^{\frac{3}{2}}\Big(\frac{1}{n} + \operatorname{Lip}_{\nabla\varphi,\mathcal{C}}\Big)k^{2-\frac{q}{2}} + k\Big(\sqrt{\alpha}B^{\frac{3}{2}}\operatorname{Lip}_{\nabla\varphi,\mathcal{C}} + \sqrt{\eta B \operatorname{Lip}_{\nabla\varphi,\mathcal{C}}}\Big) + \frac{\eta B^{3/2}}{n\sqrt{\alpha}}k.$$

Let $M := \max\Big\{\frac{2}{\nu}, \frac{8B(\Lambda+\varepsilon)}{\nu^2\mu_0^2}\Big\}$. Then, $\alpha = \frac{1}{Bk}$, $\eta = \xi\alpha$ and $\lambda = \frac{\gamma \operatorname{Lip}^2_{\varphi,\mathcal{C}}}{k}$ for $\xi \geq M$ and $\gamma \geq \frac{2\Lambda\xi}{B \operatorname{Lip}^2_{\varphi,\mathcal{C}}}$ satisfy all conditions of Theorems 1 and 2, and yield

$$\mathbb{E}[\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}} \mathbf{1}_{\mathcal{E}\cap\mathcal{E}_i}] \lesssim \sqrt{k}\Big(\frac{B}{\sqrt{m}} + \frac{\sqrt{B}}{m^{\frac{1}{4}}}\Big) + k^{\frac{3-q}{2}}B\Big(\frac{1}{n} + \frac{1}{\sqrt{m}}\Big) + \frac{B\xi}{n}.$$

Notably, under the persistence of excitation condition, SGN achieves $\widetilde{\mathcal{O}}(1/k)$ rate (Theorem 1 with the ξ and γ choices specified in Proposition 2) and $\mathcal{O}\Big(k^{\frac{3-q}{2}}(\frac{1}{n} + \frac{1}{\sqrt{m}}) + \sqrt{k}(\frac{1}{\sqrt{m}} + m^{-\frac{1}{4}}) + \frac{1}{n}\Big)$ stability. Particularly, $q = 1$ yields a stability bound of $\mathcal{O}\left(k(n^{-1} + m^{-1/2}) + \sqrt{k}(m^{-1/2} + m^{-1/4}) + n^{-1}\right)$.

5.3 Stability Analysis of SGN

In this subsection, we provide a proof sketch for Theorem 2, which is highly illuminating to showcase the impact of (i) curvature, (ii) preconditioner mismatch, and (iii) metric mismatch on the stability of SGN.

Proof of Theorem 2. Let $\pi_k := \pi_{\mathcal{C}}^{\boldsymbol{H}_k}$, $\pi'_k := \pi_{\mathcal{C}}^{\boldsymbol{H}'_k}$ and $\bar{\pi}_k := \pi_{\mathcal{C}}^{\bar{\boldsymbol{H}}_k}$ be the projection operators, and define

$$u_k = w_k - \eta \boldsymbol{H}_k^{-1}\Psi_k(w_k)$$
$$u'_k = w'_k - \eta[\boldsymbol{H}'_k]^{-1}\Psi'_k(w'_k).$$

Then, we have the following error decomposition:

$$\|\Delta_{k+1}\|_{\bar{\boldsymbol{H}}_k} \leq \|\underbrace{\bar{\pi}_k(u_k - u'_k)}_{(A_k)}\|_{\bar{\boldsymbol{H}}_k} + \|\underbrace{(\pi_k - \bar{\pi}_k)u_k}_{(B_k)}\|_{\bar{\boldsymbol{H}}_k} + \|\underbrace{(\bar{\pi}_k - \pi'_k)u'_k}_{(B'_k)}\|_{\bar{\boldsymbol{H}}_k}.$$

In this inequality, (B_k) and (B'_k) are error terms due to the metric (or projection) mismatch. We show that the critical term (A_k) will yield $\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}}$ plus controllable error terms with n and m.

Bounding $(A_{k,1})$. $\mathcal{C}$ is compact and convex, and $\bar{\boldsymbol{H}}_k$ is positive definite, thus $\bar{\pi}_k$ is non-expansive (Nesterov, 2018; Brézis, 2011):

$$\|\bar{\pi}_k(u_k - u'_k)\|_{\bar{\boldsymbol{H}}_k} \leq \|u_k - u'_k\|_{\bar{\boldsymbol{H}}_k}.$$

Let $T_k(w) := w - \eta \bar{\boldsymbol{H}}_k^{-1}\Psi_k(w)$. Then, we further decompose (A_k) as follows:

$$\begin{aligned}\|u_k - u'_k\|_{\bar{\boldsymbol{H}}_k} \leq \|&\underbrace{T_k(w_k) - T_k(w'_k)}_{(A_{k,1})}\|_{\bar{\boldsymbol{H}}_k} \\ &+ \eta\|\underbrace{\bar{\boldsymbol{H}}_k^{-1}(\Psi_k(w'_k) - \Psi'_k(w'_k))}_{(A_{k,2})}\|_{\bar{\boldsymbol{H}}_k} \\ &+ \eta\|\underbrace{(\bar{\boldsymbol{H}}_k^{-1} - \boldsymbol{H}_k^{-1})\Psi_k(w_k)}_{(A_{k,3})}\|_{\bar{\boldsymbol{H}}_k} \\ &+ \eta\|\underbrace{\big((\boldsymbol{H}'_k)^{-1} - \bar{\boldsymbol{H}}_k^{-1}\big)\Psi'_k(w'_k)}_{(A'_{k,3})}\|_{\bar{\boldsymbol{H}}_k}.\end{aligned}$$

In this decomposition, $(A_{k,3})$ and $(A'_{k,3})$ correspond to the *preconditioner mismatch* terms.

Lemma 3 (Approximate non-expansivity of T_k)**.** *For any $u, v \in \mathcal{C}$, if (η, λ) satisfy (8), then we have the (approximate) non-expansivity*

$$\begin{aligned}\|T_k(u) - T_k(v)\|_{\bar{\boldsymbol{H}}_k} \leq \|u - v\|_{\bar{\boldsymbol{H}}_{k-1}} + 2r_{\mathcal{C}}\sqrt{\eta\varepsilon} \\ + 2\sqrt{2\alpha}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} B r_{\mathcal{C}}\Big(\frac{r_{\mathcal{C}}}{2} + \frac{\sqrt{B}\mathrm{Lip}_{\ell,\mathcal{C}}}{\mu_0\nu}\Big)\end{aligned}$$

almost surely.

A key part to prove Lemma 3 is to show an approximate co-coercivity in the overparameterized regime.

Bounding $(A_{k,2})$. This term corresponds to *gradient mismatch.* Since $\{j^\star \notin I_k\} \subset \{\Psi_k(\cdot) = \Psi'_k(\cdot)\}$, we obtain

$$\begin{aligned}\|\bar{\boldsymbol{H}}_k^{-1}(\Psi_k(w'_k) - \Psi'_k(w'_k))\|_{\bar{\boldsymbol{H}}_k} \\ \leq 2\sup_{w\in\mathcal{C}}\|\Psi_k(w)\|_{\bar{\boldsymbol{H}}_k^{-1}}\mathbf{1}_{\{j^\star\in I_k\}} \\ \leq \frac{2B}{\lambda_{\min}^{1/2}(\bar{\boldsymbol{H}}_k)}\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\mathbf{1}_{\{j^\star\in I_k\}}. \quad (9)\end{aligned}$$

Bounding $(A_{k,3})$ and $(A'_{k,3})$. To bound these error terms that stem from the preconditioner mismatch, we use

$$\begin{aligned}\|(\bar{\boldsymbol{H}}_k^{-1} - \boldsymbol{H}_k^{-1})\Psi_k(w_k)\|_{\bar{\boldsymbol{H}}_k} \\ \leq \|\bar{\boldsymbol{H}}_k^{-1/2}(\bar{\boldsymbol{H}}_k - \boldsymbol{H}_k)\boldsymbol{H}_k^{-1}\|_2\|\Psi_k(w_k)\|_2. \quad (10)\end{aligned}$$

We have

$$\lambda_{\max}(\bar{\boldsymbol{H}}_k) \leq \alpha B(k+1)\mathrm{Lip}_{\varphi,\mathcal{C}}^2 + \lambda,$$

$\sup_{w\in\mathcal{C}}\|\Psi_k(w)\|_2 \leq B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}$, and

$$\|\boldsymbol{H}'_k - \bar{\boldsymbol{H}}_k\|_2 = \|\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k\|_2 = \frac{1}{2}\|\boldsymbol{H}_k - \boldsymbol{H}'_k\|_2.$$

The following lemma upper bounds the error term $\mathbb{E}\|\boldsymbol{H}_k - \boldsymbol{H}'_k\|_2$.

Lemma 4 (Stability of $\boldsymbol{H}_k$)**.** *For any $k \in \mathbb{N}$,*

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{H}_k - \boldsymbol{H}'_k\|_2] \leq 2B\alpha(k+1)r_{\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} \\ + \alpha\mathrm{Lip}_{\varphi,\mathcal{C}}^2\frac{(k+1)B}{n}.\end{aligned}$$

Bounding (B_k) and (B'_k). Finally, we bound the error due to metric (i.e., projection) mismatch.

Lemma 5 (Metric mismatch)**.** *For any $k \in \mathbb{N}$, we obtain*

$$\begin{aligned}&\|(\pi_k - \bar{\pi}_k)u_k\|_{\bar{\boldsymbol{H}}_k} \\ &\quad\leq \|\bar{\boldsymbol{H}}_k^{-1/2}(\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k)\|_2\sup_{w\in\mathcal{C}}\|w - u_k\|_2 \\ &\quad\leq \frac{\|\boldsymbol{H}_k - \boldsymbol{H}'_k\|_2}{2\lambda_{\min}^{1/2}(\bar{\boldsymbol{H}}_k)}\sup_{w\in\mathcal{C}}\|w - u_k\|_2 \quad (11)\end{aligned}$$

Also, for any $w \in \mathcal{C}$, $\|w - u_k\|_2 \leq r_{\mathcal{C}} + \frac{\eta B}{\lambda}\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}$.

The proof of Theorem 2 follows from substituting the inequalities to establish an upper bound for $\mathbb{E}\|\Delta_{k+1}\|_{\bar{\boldsymbol{H}}_k} - \mathbb{E}\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}}$, telescoping sum over k, and finally noting that $\Delta_0 = 0$ since $w_0 = w'_0$. □

6 CONCLUSIONS

In this work, we analyzed SGN for deep neural networks in the near-initialization regime. We established optimization and (algorithm-dependent) generalization bounds for SGN, with explicit dependencies on key factors such as damping, overparameterization, batch size, training duration, and the spectrum of the preconditioner. Our analysis demonstrates the robustness of SGN in regimes with ill-conditioned kernels and large loss curvature in deep learning. To the best of our knowledge, this is the first work to establish stability-based generalization bounds for Gauss-Newton in deep learning. Interesting directions for future research include extension of the analysis framework developed in this paper to more general preconditioned (e.g., adaptive gradient or Hessian-based) methods in deep learning, and also extension to other neural architectures.

References

Abbott, S. (2015). *Understanding Analysis*. Springer, 2nd edition.

Amari, S.-i., Ba, J., Grosse, R., Li, X., Nitanda, A., Suzuki, T., Wu, D., and Xu, J. (2020). When does preconditioning help or hurt generalization? *arXiv preprint arXiv:2006.10732*.

Anderson, T. and Taylor, J. B. (1979). Strong consistency of least squares estimates in dynamic models. *The annals of Statistics*, 7(3):484–489.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Arbel, M., Menegaux, R., and Wolinski, P. (2023). Rethinking gauss-newton for learning over-parameterized models. *Advances in neural information processing systems*, 36:33379–33402.

Baillon, J.-B. and Haddad, G. (1977). Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. *Israel Journal of Mathematics*, 26(2):137–150.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. (2020). Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391.

Beck, A. (2017). *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Bertsekas, D. P. (1996). Incremental least squares methods and the extended kalman filter. *SIAM Journal on Optimization*, 6(3):807–822.

Bombari, S., Amani, M. H., and Mondelli, M. (2022). Memorization and optimization in deep neural networks with minimum over-parameterization. *Advances in Neural Information Processing Systems*, 35:7628–7640.

Botev, A., Ritter, H., and Barber, D. (2017). Practical gauss-newton optimisation for deep learning. In *International Conference on Machine Learning*, pages 557–565. PMLR.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.

Brézis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Cai, T., Gao, R., Hou, J., Chen, S., Wang, D., He, D., Zhang, Z., and Wang, L. (2019). Gram-gauss-newton method: Learning overparameterized neural networks for regression problems. *arXiv preprint arXiv:1905.11675*.

Cayci, S. (2024). A riemannian optimization perspective of the gauss-newton method for feedforward neural networks. *arXiv preprint arXiv:2412.14031*.

Cayci, S. and Eryilmaz, A. (2025). Convergence of gradient descent for recurrent neural networks: A nonasymptotic analysis. *SIAM Journal on Mathematics of Data Science*, 7(2):826–854.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Duchi, J. C. (2018). Introductory lectures on stochastic optimization. In Mahoney, M. W., Duchi, J. C., and Gilbert, A. C., editors, *The Mathematics of Data*, volume 25 of *IAS/Park City Mathematics Series*, pages 99–185. American Mathematical Society.

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

Gratton, S., Lawless, A. S., and Nichols, N. K. (2007). Approximate gauss–newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132.

Gupta, V., Koren, T., and Singer, Y. (2018). Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR.

Hamm, T. (2022). A derivation of feedforward neural network gradients using fr\'echet calculus. *arXiv preprint arXiv:2209.13234*.

Hao, W., Hong, Q., and Jin, X. (2024). Gauss newton method for solving variational problems of pdes with neural network discretizaitons. *Journal of scientific computing*, 100(1):17.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192.

Hellström, F., Durisi, G., Guedj, B., Raginsky, M., et al. (2025). Generalization bounds: Perspectives from information theory and pac-bayes. *Foundations and Trends® in Machine Learning*, 18(1):1–223.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Ioannou, P. A. and Sun, J. (1996). *Robust adaptive control*, volume 1. PTR Prentice-Hall Upper Saddle River, NJ.

Ji, Z. and Telgarsky, M. (2020). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*.

Ji, Z., Telgarsky, M., and Xian, R. (2020). Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*.

Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

Karhadkar, K., Murray, M., and Montúfar, G. (2024). Bounds for the smallest eigenvalue of the ntk for arbitrary spherical data of arbitrary dimension. *Advances in Neural Information Processing Systems*, 37:138197–138249.

Klochkov, Y. and Zhivotovskiy, N. (2021). Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, pages 154–166.

Landford, O. and Robinson, D. W. (1967). Mean entropy of states in quantu, statistical mechanics. Technical report, CM-P00057584.

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (2002). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer.

Lei, Y. and Ying, Y. (2020). Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR.

Martens, J. (2010). Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 735–742.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76.

Martens, J. and Grosse, R. (2015). Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR.

Mckay, M. B., Kaur, A., Greif, C., and Wetton, B. (2025). Near-optimal sketchy natural gradients for physics-informed neural networks. In *Forty-second International Conference on Machine Learning*.

Morwani, D., Shapira, I., Vyas, N., eran malach, Kakade, S. M., and Janson, L. (2025). A new perspective on shampoo's preconditioner. In *The Thirteenth International Conference on Learning Representations*.

Müller, J. and Zeinhofer, M. (2023). Achieving high accuracy with pinns via energy natural gradient descent. In *International Conference on Machine Learning*, pages 25471–25485. PMLR.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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

Zhang, G., Martens, J., and Grosse, R. B. (2019). Fast convergence of natural gradient descent for overparameterized neural networks. *Advances in Neural Information Processing Systems*, 32.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] **Justification.** Mathematical descriptions of the problem, model, and algorithms were provided in Sections 2 and 3. Assumptions and conditions are clearly stated.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] **Justification.** We establish theoretical analyses of Stochastic Gauss-Newton algorithm, which provide finite-time, finite-sample and finite-width optimization and generalization bounds. For main theoretical results that yield iteration, sample and network complexity, see Theorems 1 and 2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] **Justification.** Assumptions and conditions are clearly stated. For instance, see Assumptions 1-2 and Condition 1.
   (b) Complete proofs of all theoretical results. [Yes] **Justification.** Complete proofs can be found in the supplementary material. A proof sketch for our generalization bound is provided in Section 5.3.
   (c) Clear explanations of any assumptions. [Yes] **Justification.** Each assumption and condition is followed by explanations.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

A PROOFS FOR SECTION 4

The following elementary results, which follow from Lipschitz smoothness (see Lemma 3.4 in Bubeck (2015)), will be useful in the proofs.

Lemma 6. *For any $\boldsymbol{x} \in \mathcal{B}_2(0,1)$ and $w, w' \in \mathcal{C}$, we have*

$$|\varphi(\boldsymbol{x}; w') - \varphi(\boldsymbol{x}; w) - \nabla^\top \varphi(\boldsymbol{x}; w)\big(w' - w\big)| \leq \frac{1}{2}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2.$$

For any $k \in \mathbb{N}$, let

$$\boldsymbol{f}_k(w) := [\varphi(\boldsymbol{x}_j; w)]_{j \in I_k} \in \mathbb{R}^B \tag{12}$$

be output vector of the neural network with parameter $w \in \mathbb{R}^p$ evaluated on I_k, and

$$\boldsymbol{\epsilon}_k := \boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k) - \boldsymbol{J}_k(w_k)\big(\bar{w} - w_k\big).$$

Then, we have $\|\boldsymbol{\epsilon}_k\|_2 \leq \frac{\sqrt{B}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2}{2}$ a.s.

In the following, we present a proof of Lemma 1 based on a standard application of Fréchet calculus and repeated chain-rule arguments. A first-order derivation of feedforward network gradients in this framework was established by Hamm (2022). Our argument builds on this approach, and extends it to second-order derivatives.

Proof of Lemma 1. Given an input $\boldsymbol{x} \in \mathscr{X}$, let $x^{(0)}(\boldsymbol{x}; w) := \boldsymbol{x}$, $x^{(h)}(\boldsymbol{x}; w) := f_h\big(W^{(h)}, x^{(h-1)}(\boldsymbol{x}; w)\big)$ for $h \in [H]$, and $\varphi(\boldsymbol{x}; w) := c^\top x^{(H)}(\boldsymbol{x}; w)$, with

$$f_h(W, y) := \frac{1}{\sqrt{m}}\vec{\sigma}(Wy), \qquad w = (W^{(1)}, \ldots, W^{(H)}, c).$$

Let $a := Wy$ and $D := \mathrm{diag}(\sigma'(a_1), \ldots, \sigma'(a_m))$. Then, we have the Fréchet derivatives of f_h as follows:

$$D_W f_h(W, y)[\Delta] = \frac{1}{\sqrt{m}} D\Delta y, \qquad D_y f_h(W, y)[\eta] = \frac{1}{\sqrt{m}} DW\eta,$$

where Δ is the perturbation of W and η is the perturbation of y (Hamm, 2022). We also have

$$\begin{aligned}
D_W^2 f_h(W, y)[\Delta_1, \Delta_2] &= \frac{1}{\sqrt{m}}\big(\sigma''(a) \odot (\Delta_1 y) \odot (\Delta_2 y)\big), \\
D_y^2 f_h(W, y)[\eta_1, \eta_2] &= \frac{1}{\sqrt{m}}\big(\sigma''(a) \odot (W\eta_1) \odot (W\eta_2)\big), \\
D_{W,y}^2 f_h(W, y)[\Delta, \eta] &= \frac{1}{\sqrt{m}}\big(\sigma''(a) \odot (\Delta y) \odot (W\eta)\big) + \frac{1}{\sqrt{m}} D\Delta\eta,
\end{aligned}$$

where $\odot$ denotes Hadamard product. By using Cauchy-Schwarz inequality and $\|\sigma\|_\infty \leq \sigma_0$, $\|\sigma'\|_\infty \leq \sigma_1$, $\|\sigma''\|_\infty \leq \sigma_2$, we have

$$\|f_h(W, y)\|_2 \leq \sigma_0, \ \|D_W f_h(W, y)\|_{\mathrm{op}} \leq \frac{\sigma_1\|y\|_2}{\sqrt{m}}, \text{ and } \|D_y f_h(W, y)\|_{\mathrm{op}} \leq \frac{\sigma_1\|W\|_2}{\sqrt{m}}.$$

Similarly, we have

$$\|D_W^2 f_h(W, y)\|_{\mathrm{op}} \leq \frac{\sigma_2\|y\|_2^2}{\sqrt{m}}, \ \|D_y^2 f_h(W, y)\|_{\mathrm{op}} \leq \frac{\sigma_2\|W\|_2^2}{\sqrt{m}}, \text{ and } \|D_{W,y}^2 f_h(W, y)\|_{\mathrm{op}} \leq \frac{\sigma_2\|W\|_2\|y\|_2 + \sigma_1}{\sqrt{m}}.$$

For any parameter $w \in \mathcal{C}$, we have $\frac{\|W^{(h)}\|_2}{\sqrt{m}} \le \kappa_{\mathcal{C}}$, $\|c\|_2 \le \zeta_{\mathcal{C}}$, and $\|x^{(h)}(\boldsymbol{x}; w)\|_2 \le \sigma_0$ for any $\|\boldsymbol{x}\|_2 \le 1$. Let the parameter moduli of Lipschitz continuity and Lipschitz smoothness of the h-th layer be defined as

$$A_h := \sup_{\substack{w \in \mathcal{C},\ \|\boldsymbol{x}\|_2 \le 1 \\ \|u\|_2 = 1}} \|D_w x^{(h)}(\boldsymbol{x}; w)[u]\|_2, \qquad B_h := \sup_{\substack{w \in \mathcal{C},\ \|\boldsymbol{x}\|_2 \le 1 \\ \|u\|_2 = \|v\|_2 = 1}} \|D_w^2 x^{(h)}(\boldsymbol{x}; w)[u, v]\|_2.$$

We have $A_0 = 0$ and $B_0 = 0$. Using the chain rule, we obtain the following recursive formula for A_h:

$$A_h \le \frac{\sigma_0 \sigma_1}{\sqrt{m}} + \sigma_1 \kappa_{\mathcal{C}} A_{h-1} \le \frac{\sigma_0 L_{\mathcal{C}}}{\sqrt{m}} + L_{\mathcal{C}} A_{h-1},$$

hence $A_h \le \sigma_0 \frac{h L_{\mathcal{C}}^h}{\sqrt{m}}$. Similarly, by using the second-order chain rule, we obtain

$$\begin{aligned} D_w^2 x^{(h)}(\boldsymbol{x}; w)[u, v] = {} & D_W^2 f_h\big(W^{(h)}, x^{(h-1)}(\boldsymbol{x}; w)\big)[u_W^{(h)}, v_W^{(h)}] \\ & + D_{W,y}^2 f_h\big(W^{(h)}, x^{(h-1)}(\boldsymbol{x}; w)\big)\big[u_W^{(h)}, D_w x^{(h-1)}(\boldsymbol{x}; w)[v]\big] \\ & + D_{W,y}^2 f_h\big(W^{(h)}, x^{(h-1)}(\boldsymbol{x}; w)\big)\big[v_W^{(h)}, D_w x^{(h-1)}(\boldsymbol{x}; w)[u]\big] \\ & + D_y^2 f_h\big(W^{(h)}, x^{(h-1)}(\boldsymbol{x}; w)\big)\big[D_w x^{(h-1)}(\boldsymbol{x}; w)[u], D_w x^{(h-1)}(\boldsymbol{x}; w)[v]\big] \\ & + D_y f_h\big(W^{(h)}, x^{(h-1)}(\boldsymbol{x}; w)\big)\big[D_w^2 x^{(h-1)}(\boldsymbol{x}; w)[u, v]\big]. \end{aligned}$$

Using the bounds above and $\|u_W^{(h)}\|_F, \|v_W^{(h)}\|_F \le 1$, we obtain

$$\begin{aligned} B_h &\le \frac{\sigma_2 \sigma_0^2}{\sqrt{m}} + 2 \cdot \frac{\sigma_2 \|W^{(h)}\|_2 \sigma_0 + \sigma_1}{\sqrt{m}} A_{h-1} + \frac{\sigma_2 \|W^{(h)}\|_2^2}{\sqrt{m}} A_{h-1}^2 + \frac{\sigma_1 \|W^{(h)}\|_2}{\sqrt{m}} B_{h-1} \\ &\le L_{\mathcal{C}} B_{h-1} + 8(1 + \sigma_0 + \sigma_0^2 \sigma_2) \frac{h^2 L_{\mathcal{C}}^{2h}}{\sqrt{m}}, \end{aligned}$$

which implies

$$B_h \le 8(1 + \sigma_0 + \sigma_0^2 \sigma_2) \frac{h^3 L_{\mathcal{C}}^{2h}}{\sqrt{m}}.$$

Finally, since $\varphi(\boldsymbol{x}; w) = c^\top x^{(H)}(\boldsymbol{x}; w)$, the directional derivative of φ along $u = (u_W, u_c)$ is

$$D_w \varphi(\boldsymbol{x}; w)[u] = u_c^\top x^{(H)}(\boldsymbol{x}; w) + c^\top D_w x^{(H)}(\boldsymbol{x}; w)[u].$$

Hence,

$$|D_w \varphi(\boldsymbol{x}; w)[u]| \le \|u_c\|_2 \|x^{(H)}(\boldsymbol{x}; w)\|_2 + \|c\|_2 \|D_w x^{(H)}(\boldsymbol{x}; w)[u]\|_2 \le \sigma_0 + \zeta_{\mathcal{C}} A_H.$$

Moreover, differentiating once more in the direction $v = (v_W, v_c)$, we obtain

$$D_w^2 \varphi(\boldsymbol{x}; w)[u, v] = u_c^\top D_w x^{(H)}(\boldsymbol{x}; w)[v] + v_c^\top D_w x^{(H)}(\boldsymbol{x}; w)[u] + c^\top D_w^2 x^{(H)}(\boldsymbol{x}; w)[u, v].$$

Therefore,

$$|D_w^2 \varphi(\boldsymbol{x}; w)[u, v]| \le \|u_c\|_2 A_H + \|v_c\|_2 A_H + \|c\|_2 B_H \le 2A_H + \zeta_{\mathcal{C}} B_H.$$

Substituting the bounds for A_H and B_H concludes the proof. □

Proof of Theorem 1. The key idea behind our analysis of this variant of stochastic Gauss-Newton is to interpret it as a combination of a variable-metric method akin to adaptive gradient methods (Duchi, 2018), which have path-dependent and time-variant Bregman divergences. In essence, we exploit this insight in quantifying the impact of time-variance of the Bregman divergence ((a_k) below) and bounding the gradient norm ($(b_{k,2})$ below). The most significant component of the Lyapunov drift is the negative drift term ($(b_{k,1})$ below), which we handle by using the specific properties of SGN.

First, let

$$\mathscr{L}_k(w) := \|w - \bar{w}\|_{\boldsymbol{H}_k}^2, \quad w \in \mathcal{C}$$

be the Lyapunov function. We apply the following decomposition (Duchi, 2018):

$$\begin{aligned}\mathscr{L}_{k+1}(w_{k+1}) &= \mathscr{L}_k(w_{k+1}) + \|w_{k+1}-\bar{w}\|^2_{\boldsymbol{H}_{k+1}-\boldsymbol{H}_k}\\ &\le \|u_k-\bar{w}\|^2_{\boldsymbol{H}_k} + \|w_{k+1}-\bar{w}\|^2_{\boldsymbol{H}_{k+1}-\boldsymbol{H}_k}\\ &= \mathscr{L}_k(w_k) + 2\eta \underbrace{(\bar{w}-w_k)^\top \Psi_k(w_k)}_{=:(b_{k,1})} + \eta^2 \underbrace{\|\Psi_k(w_k)\|^2_{\boldsymbol{H}_k^{-1}}}_{=:(b_{k,2})} + \underbrace{\|w_{k+1}-\bar{w}\|^2_{\boldsymbol{H}_{k+1}-\boldsymbol{H}_k}}_{=:(a_k)}.\end{aligned}$$

Consider the filtration $\mathscr{F}_k := \sigma(I_0, I_1, \dots, I_k),\ k \in \mathbb{N}$, which is the history of the optimization path up to time k. As elementary properties, notice that w_k is $\mathscr{F}_{k-1}$-measurable, and

$$\mathbb{E}[\Phi_k(w_k) \mid \mathscr{F}_{k-1}] = B \cdot \widehat{\mathcal{R}}_{\mathcal{S}}(w_k) \Rightarrow \mathbb{E}[\Phi_k(w_k) - \Phi_k(\bar{w}) \mid \mathscr{F}_{k-1}] = B\Big(\widehat{\mathcal{R}}_{\mathcal{S}}(w_k) - \widehat{\mathcal{R}}_{\mathcal{S}}(\bar{w})\Big)$$

for each k.

Bounding $(b_{k,1})$. Recall that $\Psi_k(w) = \boldsymbol{J}_k^\top(w)\boldsymbol{G}_k(w)$ and $\boldsymbol{\epsilon}_k = \boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k) - \boldsymbol{J}_k(w_k)\big(\bar{w} - w_k\big)$. Then,

$$\begin{aligned}(b_{k,1}) &= \boldsymbol{G}_k^\top(w_k)\boldsymbol{J}_k(w_k)\big(\bar{w}-w_k\big)\\ &= \boldsymbol{G}_k^\top(w_k)\Big(\boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k) - \boldsymbol{\epsilon}_k\Big)\\ &\le \boldsymbol{G}_k^\top(w_k)\big(\boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k)\big) + \frac{B}{2}\cdot \mathrm{Lip}_{\ell,\mathcal{C}} \cdot \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} \cdot r_{\mathcal{C}}^2,\end{aligned}$$

since $\|\boldsymbol{G}_k(w)\|_2 \le \sqrt{B}\mathrm{Lip}_{\ell,\mathcal{C}}$ and $\|\boldsymbol{\epsilon}_k\|_2 \le \frac{\sqrt{B}}{2}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2$ by Lemma 6. Using the ν-strong convexity of ℓ,

$$\boldsymbol{G}_k^\top(w_k)\big(\boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k)\big) \le \Phi_k(\bar{w}) - \Phi_k(w_k) - \frac{\nu}{2}\|\boldsymbol{f}_k(w_k) - \boldsymbol{f}_k(\bar{w})\|_2^2.$$

Therefore, we have

$$\mathbb{E}[(b_{k,1}) \mid \mathscr{F}_{k-1}] \le -B \cdot V_k - \frac{\nu}{2}\mathbb{E}[\|\boldsymbol{f}_k(w_k) - \boldsymbol{f}_k(\bar{w})\|_2^2 \mid \mathscr{F}_{k-1}] + \frac{B}{2}\cdot \mathrm{Lip}_{\ell,\mathcal{C}} \cdot \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} \cdot r_{\mathcal{C}}^2.$$

By the law of iterated expectation, we obtain

$$\mathbb{E}[(b_{k,1})] \le -B \cdot \mathbb{E}[V_k] - \frac{\nu}{2}\mathbb{E}[\|\boldsymbol{f}_k(w_k) - \boldsymbol{f}_k(\bar{w})\|_2^2] + \frac{B}{2}\cdot \mathrm{Lip}_{\ell,\mathcal{C}} \cdot \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} \cdot r_{\mathcal{C}}^2. \tag{13}$$

Bounding $(b_{k,2})$. This part corresponds to the squared norm of the preconditioned gradient update with respect to the $\boldsymbol{H}_k$-norm. The initial term is bounded by using Sherman-Morrison-Woodbury matrix identity, and the subsequent terms are bounded by exploiting trace inequalities (building on Hazan et al. (2007)) and the incremental nature of $\boldsymbol{H}_k$.

First, note that

$$\begin{aligned}\|\boldsymbol{J}_k^\top(w_k)\boldsymbol{G}_k(w_k)\|^2_{\boldsymbol{H}_k^{-1}} &= \boldsymbol{G}_k^\top(w_k)\boldsymbol{J}_k(w_k)\boldsymbol{H}_k^{-1}\boldsymbol{J}_k^\top(w_k)\boldsymbol{G}_k(w_k)\\ &\le B \cdot \mathrm{Lip}^2_{\ell,\mathcal{C}} \cdot \mathrm{tr}\Big(\boldsymbol{J}_k^\top(w_k)\boldsymbol{J}_k(w_k)\boldsymbol{H}_k^{-1}\Big),\end{aligned}$$

where the last inequality follows from the circular shift invariance of tr. By Klein's inequality (Lemma 1 in Landford and Robinson (1967), also see Hazan et al. (2007)), we have

$$\mathrm{tr}\log \boldsymbol{H}_{k+1} - \alpha \cdot \mathrm{tr}\Big(\boldsymbol{J}_{k+1}^\top(w_{k+1})\boldsymbol{J}_{k+1}(w_{k+1})\boldsymbol{H}_{k+1}^{-1})\Big) - \mathrm{tr}\log \boldsymbol{H}_k \ge 0.$$

Hence, we have

$$\begin{aligned}\sum_{k=0}^{t-1}(b_{k,2}) &\le B \cdot \mathrm{Lip}^2_{\ell,\mathcal{C}}\left(\|\boldsymbol{J}_0(w_0)\boldsymbol{H}_0^{-1}\boldsymbol{J}_0^\top(w_0)\|_2 + \sum_{k=1}^{t-1}\mathrm{tr}\Big(\boldsymbol{J}_k^\top(w_k)\boldsymbol{J}_k(w_k)\boldsymbol{H}_k^{-1}\Big)\right)\\ &\le B \cdot \mathrm{Lip}^2_{\ell,\mathcal{C}}\Big(\|\boldsymbol{J}_0(w_0)\boldsymbol{H}_0^{-1}\boldsymbol{J}_0^\top(w_0)\|_2 + \frac{1}{\alpha}\big(\mathrm{tr}\log \boldsymbol{H}_t - \mathrm{tr}\log \boldsymbol{H}_0\big)\Big).\end{aligned}$$

In order to bound $\|\boldsymbol{J}_0(w_0)\boldsymbol{H}_0^{-1}\boldsymbol{J}_0^\top(w_0)\|_2$, we use the Sherman-Morrison-Woodbury matrix identity (Horn and Johnson, 2012; Cayci, 2024):

$$\boldsymbol{J}_0(w_0)\boldsymbol{H}_0^{-1}\boldsymbol{J}_0^\top(w_0) = \frac{1}{\lambda}\Big(\boldsymbol{K}_0 - \frac{\alpha}{\lambda}\boldsymbol{K}_0[\boldsymbol{I} + \alpha\lambda^{-1}\boldsymbol{K}_0]^{-1}\boldsymbol{K}_0\Big),$$

where $\boldsymbol{K}_0 := \boldsymbol{J}_0(w_0)\boldsymbol{J}_0^\top(w_0)$. If $(\mu^2, u) \in \mathbb{R}^+ \times \mathbb{R}^B$ is an eigenpair of $\boldsymbol{K}_0$, then $\Big(\frac{\mu^2}{\lambda+\alpha\mu^2}, u\Big)$ is an eigenpair of $\boldsymbol{J}_0(w_0)\boldsymbol{H}_0^{-1}\boldsymbol{J}_0^\top(w_0)$. Note that $s \mapsto \frac{s}{\lambda+\alpha s}$ is an increasing function, thus

$$\begin{aligned}\|\boldsymbol{J}_0(w_0)\boldsymbol{H}_0^{-1}\boldsymbol{J}_0^\top(w_0)\|_2 &\leq \frac{\|\boldsymbol{K}_0\|_2}{\lambda + \alpha\|\boldsymbol{K}_0\|_2} \\ &\leq \frac{B\cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2}{\lambda + \alpha\cdot B\cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2}\end{aligned}$$

where the second line follows from

$$\sup_{w\in\mathcal{C}} \|\boldsymbol{J}_k(w)\|_2 \leq \sqrt{B}\mathrm{Lip}_{\varphi,\mathcal{C}} \quad a.s. \text{ for all } k \in \mathbb{N}.$$

Also, note that

$$\mathrm{tr}\log \boldsymbol{H}_t = \log\det \boldsymbol{H}_t,\ \ t \in \mathbb{N}.$$

Putting everything together, we obtain:

$$\sum_{k=0}^{t-1}(b_{k,2}) \leq B\cdot \mathrm{Lip}_{\ell,\mathcal{C}}^2\Big(\frac{B\cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2}{\lambda + \alpha\cdot B\cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2} + \frac{1}{\alpha}\cdot\log\frac{\det \boldsymbol{H}_t}{\det \boldsymbol{H}_0}\Big).$$

For any $\gamma > 0$, let $\lambda := \gamma\alpha B\mathrm{Lip}_{\varphi,\mathcal{C}}^2$. Substituting this into the above inequality and taking expectation yields

$$\sum_{k=0}^{t-1}\mathbb{E}\left[(b_{k,2})\right] \leq \frac{B\mathrm{Lip}_{\ell,\mathcal{C}}^2}{\alpha}\Big(\frac{1}{1+\gamma} + \mathbb{E}\log\frac{\det \boldsymbol{H}_t}{\det \boldsymbol{H}_0}\Big). \tag{14}$$

Bounding (a_k). Note that the Bregman divergence in SGN updates are time-variant, and this error term quantifies the impact of this metric change at k-th iteration. First, note that

$$\|w_{k+1} - \bar{w}\|_{\boldsymbol{H}_{k+1}-\boldsymbol{H}_k}^2 = \alpha\|\boldsymbol{J}_{k+1}(w_{k+1})(w_{k+1} - \bar{w})\|^2.$$

By using Lemma 6 and $(x+y)^2 \leq 2(x^2+y^2)$,

$$\begin{aligned}\|\boldsymbol{J}_{k+1}(w_{k+1})(w_{k+1} - \bar{w})\|^2 &= \|\boldsymbol{\epsilon}_{k+1} - \boldsymbol{f}_{k+1}(\bar{w}) + \boldsymbol{f}_{k+1}(w_{k+1})\|_2^2 \\ &\leq \frac{B\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^4}{2} + 2\cdot\|\boldsymbol{f}_{k+1}(\bar{w}) - \boldsymbol{f}_{k+1}(w_{k+1})\|_2^2.\end{aligned}$$

Hence, we obtain

$$\sum_{k=0}^{t-1}\mathbb{E}[(a_k)] \leq \frac{\alpha t B\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^4}{2} + 2\alpha\sum_{k=1}^{t}\mathbb{E}\|\boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k)\|_2^2. \tag{15}$$

Drift analysis. Note that

$$\begin{aligned}\mathbb{E}[\mathscr{L}_t(w_t)] - \mathbb{E}[\mathscr{L}_0(w_0)] &= \sum_{k=0}^{t-1}\mathbb{E}[\mathscr{L}_{k+1}(w_{k+1}) - \mathscr{L}_k(w_k)] \\ &\leq 2\eta\sum_{k=0}^{t-1}\mathbb{E}[(b_{k,1})] + \eta^2\sum_{k=0}^{t-1}\mathbb{E}[(b_{k,2})] + \sum_{k=0}^{t-1}\mathbb{E}[(a_k)].\end{aligned}$$

Substituting (13),(14) and (15) into the above inequality, we obtain

$$\begin{aligned}\mathbb{E}[\mathscr{L}_t(w_t)] - \mathbb{E}[\mathscr{L}_0(w_0)] \leq -2\eta B \sum_{k=0}^{t-1} \mathbb{E}[V_k] + (2\alpha - \nu\eta) \sum_{k=1}^{t-1} \mathbb{E}\|\boldsymbol{f}_k(\bar{w}) - \boldsymbol{f}_k(w_k)\|_2^2 \\ + 2\alpha \mathbb{E}\|\boldsymbol{f}_t(w_t) - \boldsymbol{f}_t(\bar{w})\|_2^2 + \eta t B \mathrm{Lip}_{\ell,\mathcal{C}} \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2 \\ + \frac{\eta^2 B \mathrm{Lip}_{\ell,\mathcal{C}}^2}{\alpha} \Big(\frac{1}{1+\gamma} + \mathbb{E} \log \frac{\det \boldsymbol{H}_t}{\det \boldsymbol{H}_0} \Big) + \frac{\alpha t B \mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^4}{2}.\end{aligned}$$

By rearranging terms and noting that $\mathscr{L}_t(\cdot) > 0$ and

$$\|\boldsymbol{f}_t(\bar{w}) - \boldsymbol{f}_t(w_t)\|_2^2 = \sum_{j \in I_t} |\varphi(\boldsymbol{x}_j; w_t) - \varphi(\boldsymbol{x}_j; \bar{w})|^2 \leq B \mathrm{Lip}_{\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^2,$$

we obtain

$$\begin{aligned}\frac{1}{t} \sum_{k=0}^{t-1} \mathbb{E}[V_k] \leq \frac{\mathbb{E}\mathscr{L}_0(w_0)}{2\eta B t} + \frac{\alpha}{\eta t} \mathrm{Lip}_{\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^2 + \frac{\eta}{\alpha} \frac{\mathrm{Lip}_{\ell,\mathcal{C}}^2}{2t} \Big(\frac{1}{1+\gamma} + \mathbb{E} \log \frac{\det \boldsymbol{H}_t}{\det \boldsymbol{H}_0} \Big) \\ + \frac{1}{2} \mathrm{Lip}_{\ell,\mathcal{C}} r_{\mathcal{C}}^2 \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{1}{4} \cdot \frac{\alpha}{\eta} \cdot r_{\mathcal{C}}^4 \cdot \mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2.\end{aligned}$$

We have

$$\begin{aligned}\mathscr{L}_0(w_0) = \|w_0 - \bar{w}\|_{\boldsymbol{H}_0}^2 &\leq r_{\mathcal{C}}^2 \Big(\alpha B \mathrm{Lip}_{\varphi,\mathcal{C}}^2 + \lambda \Big) \\ &= \alpha \cdot r_{\mathcal{C}}^2 \cdot (\gamma + 1) \cdot B \cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2\end{aligned}$$

almost surely. Using these and the definition $\xi := \frac{\eta}{\alpha}$, we finally get

$$\begin{aligned}\frac{1}{t} \sum_{k=0}^{t-1} \mathbb{E}[V_k] \leq \frac{\gamma + 3}{2\xi t} \cdot \mathrm{Lip}_{\varphi,\mathcal{C}}^2 \cdot r_{\mathcal{C}}^2 + \xi \frac{\mathrm{Lip}_{\ell,\mathcal{C}}^2}{2t} \Big(\frac{1}{1+\gamma} + \mathbb{E} \log \frac{\det \boldsymbol{H}_t}{\det \boldsymbol{H}_0} \Big) \\ + \frac{1}{2} \mathrm{Lip}_{\ell,\mathcal{C}} r_{\mathcal{C}}^2 \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{1}{4\xi} \cdot r_{\mathcal{C}}^4 \cdot \mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2.\end{aligned} \tag{16}$$

This concludes the proof of the first part.

In order to prove the average-iterate convergence result, we first define the following linearization:

$$\varphi_{\mathrm{lin}}(\boldsymbol{x}; w) = \varphi(\boldsymbol{x}; w_0) + \nabla^\top \varphi(\boldsymbol{x}; w_0)\big(w - w_0\big), \quad \boldsymbol{x} \in \mathcal{B}_2(0,1), w \in \mathcal{C},$$

and use the fact that

$$\sup_{\substack{\boldsymbol{x} \in \mathcal{B}_2(0,1) \\ w \in \mathcal{C}}} |\varphi(\boldsymbol{x}; w) - \varphi_{\mathrm{lin}}(\boldsymbol{x}; w)| \leq \frac{1}{2} \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2. \tag{17}$$

Then, for any $w \in \mathcal{C}$,

$$\begin{aligned}\widehat{\mathcal{R}}_{\mathcal{S}}(w) &= \frac{1}{n} \sum_{j=1}^n \ell(\varphi_{\mathrm{lin}}(\boldsymbol{x}_j; w); \boldsymbol{z}_j) + \frac{1}{n} \sum_{j=1}^n \Big(\ell(\varphi(\boldsymbol{x}_j; w); \boldsymbol{z}_j) - \ell(\varphi_{\mathrm{lin}}(\boldsymbol{x}_j; w); \boldsymbol{z}_j) \Big) \\ &\leq \frac{1}{n} \sum_{j=1}^n \ell(\varphi_{\mathrm{lin}}(\boldsymbol{x}_j; w); \boldsymbol{z}_j) + \frac{1}{2} \mathrm{Lip}_{\ell,\mathcal{C}} \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2,\end{aligned}$$

and similarly

$$\frac{1}{n} \sum_{j=1}^n \ell(\varphi_{\mathrm{lin}}(\boldsymbol{x}_j; w); \boldsymbol{z}_j) \leq \widehat{\mathcal{R}}_{\mathcal{S}}(w) + \frac{1}{2} \mathrm{Lip}_{\ell,\mathcal{C}} \mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2.$$

Now, note that $w \mapsto \varphi_{\text{lin}}(\boldsymbol{x}; w)$ is an affine function, and $y' \mapsto \ell(y'; \boldsymbol{z})$ is (strongly) convex, which implies that $w \mapsto \ell(\varphi_{\text{lin}}(\boldsymbol{x}_j; w); \boldsymbol{z}_j)$ is convex for each $j \in [n]$. Thus, Jensen's inequality implies

$$\frac{1}{n}\sum_{j\in[n]} \ell(\varphi_{\text{lin}}(\boldsymbol{x}_j; \widehat{w}_k); \boldsymbol{z}_j) \le \frac{1}{k}\sum_{t=0}^{k-1}\frac{1}{n}\sum_{j\in[n]} \ell(\varphi_{\text{lin}}(\boldsymbol{x}_j; w_t); \boldsymbol{z}_j) \tag{18}$$

$$\le \frac{1}{k}\sum_{t=0}^{k-1} \widehat{\mathcal{R}}_{\mathcal{S}}(w_t) + \frac{1}{2}\text{Lip}_{\ell,\mathcal{C}}\text{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2. \tag{19}$$

Finally, we obtain

$$\widehat{\mathcal{R}}_{\mathcal{S}}(\widehat{w}_k) \le \frac{1}{n}\sum_{j\in[n]} \ell(\varphi_{\text{lin}}(\boldsymbol{x}_j; \widehat{w}_k); \boldsymbol{z}_j) + \frac{1}{2}\text{Lip}_{\ell,\mathcal{C}}\text{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2 \le \frac{1}{k}\sum_{t=0}^{k-1} \widehat{\mathcal{R}}_{\mathcal{S}}(w_t) + \text{Lip}_{\ell,\mathcal{C}}\text{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2.$$

Taking expectation yields

$$\begin{aligned}\mathbb{E}[\widehat{\mathcal{R}}_{\mathcal{S}}(\widehat{w}_k)] - \widehat{\mathcal{R}}_{\mathcal{S}}(\bar{w}) &\le \underbrace{\frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}[V_t]}_{(\spadesuit)} + \text{Lip}_{\ell,\mathcal{C}}\text{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2 \\ &\le \frac{\gamma+3}{2\xi k}\cdot \text{Lip}_{\varphi,\mathcal{C}}^2 \cdot r_{\mathcal{C}}^2 + \xi\frac{\text{Lip}_{\ell,\mathcal{C}}^2}{2k}\Big(\frac{1}{1+\gamma} + \mathbb{E}\log\frac{\det \boldsymbol{H}_k}{\det \boldsymbol{H}_0}\Big) \\ &\quad + \frac{3}{2}\text{Lip}_{\ell,\mathcal{C}} r_{\mathcal{C}}^2 \text{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{1}{4\xi}\cdot r_{\mathcal{C}}^4 \cdot \text{Lip}_{\nabla\varphi,\mathcal{C}}^2,\end{aligned} \tag{20}$$

which concludes the proof. The bounds presented above have sharper universal constants compared to the main paper. □

Proof of Proposition 1. Let

$$\boldsymbol{S}_k := \sum_{t=0}^{k} \boldsymbol{J}_t^\top(w_t)\boldsymbol{J}_t(w_t), \quad k \in \mathbb{N}.$$

Then, we have

$$\log\frac{\det \boldsymbol{H}_k}{\det \boldsymbol{H}_0} = \log\det\Big(\boldsymbol{I} + \frac{\alpha}{\lambda}\boldsymbol{S}_k\Big).$$

Denote the non-zero eigenvalues of $\boldsymbol{S}_k$ by $\mu_i,\ i = 1, 2, \ldots, r_k'$ where $r_k' = \text{rank}(\boldsymbol{S}_k) \le p$. Then,

$$\begin{aligned}\log\det\big(\boldsymbol{I} + \frac{\alpha}{\lambda}\boldsymbol{S}_k\big) = \sum_{i=1}^{r_k'}\log(1+\frac{\alpha}{\lambda}\mu_i) &= r_k' \cdot \frac{1}{r_k'}\sum_{i=1}^{r_k'}\log(1+\frac{\alpha}{\lambda}\mu_i) \\ &\le r_k'\log\Big(1 + \frac{\alpha}{\lambda r_k'}\sum_{i=1}^{r_k'}\mu_i\Big),\end{aligned}$$

where the inequality stems from Jensen's inequality and the concavity of $x \mapsto \log(1+x)$. By using $\mu_i \le \text{tr}(\boldsymbol{S}_k)$ and $r_k' \le p$, we obtain

$$\log\det\Big(\boldsymbol{I} + \frac{\alpha}{\lambda}\boldsymbol{S}_k\Big) \le r_k'\log\Big(1 + \frac{\alpha}{\lambda}\text{tr}(\boldsymbol{S}_k)\Big) \le p\log\Big(1 + \frac{\alpha}{\lambda}\text{tr}(\boldsymbol{S}_k)\Big).$$

Finally, we have

$$\text{tr}(\boldsymbol{S}_k) = \sum_{t=0}^{k}\text{tr}\left(\boldsymbol{J}_t^\top(w_t)\boldsymbol{J}_t(w_t)\right) = \sum_{t=0}^{k}\|\boldsymbol{J}_t(w_t)\|_F^2 \le (k+1)B\,\text{Lip}_{\varphi,\mathcal{C}}^2,$$

which concludes the proof of the first part. For the second part, we use concavity of $\boldsymbol{A} \mapsto \log\det(\boldsymbol{I} + \boldsymbol{A})$ on the cone of symmetric positive semidefinite matrices to obtain

$$\mathbb{E}\log\det\Big(\boldsymbol{I} + \frac{\alpha}{\lambda}\boldsymbol{S}_k\Big) \le \log\det\Big(\boldsymbol{I} + \frac{\alpha}{\lambda}\mathbb{E}[\boldsymbol{S}_k]\Big).$$

We have $\mathbb{E}[\boldsymbol{S}_k] = \sum_{t=0}^{k} \mathbb{E}\big[\boldsymbol{J}_t^\top(w_t)\boldsymbol{J}_t(w_t)\big] = B(k+1)\bar{\Sigma}_k$, thus

$$\mathbb{E}\log\frac{\det \boldsymbol{H}_k}{\det \boldsymbol{H}_0} \leq \log\det\Big(\boldsymbol{I} + \frac{\alpha}{\lambda}B(k+1)\bar{\Sigma}_k\Big).$$

Let $\bar{\mu}_i, i = 1, \ldots, \bar{r}_k$, be the nonzero eigenvalues of $\bar{\Sigma}_k$. Then, we obtain

$$\log\det\Big(\boldsymbol{I} + \frac{\alpha}{\lambda}B(k+1)\bar{\Sigma}_k\Big) = \sum_{i=1}^{\bar{r}_k}\log\Big(1 + \frac{\alpha}{\lambda}B(k+1)\bar{\mu}_i\Big).$$

Since $\bar{\mu}_i \leq \operatorname{tr}(\bar{\Sigma}_k)$, we obtain

$$\sum_{i=1}^{\bar{r}_k}\log\Big(1 + \frac{\alpha}{\lambda}B(k+1)\bar{\mu}_i\Big) \leq \bar{r}_k\log\Big(1 + \frac{\alpha}{\lambda}B(k+1)\operatorname{tr}(\bar{\Sigma}_k)\Big).$$

□

Proof of Corollary 1. The idea here is to construct a parameter set $\mathcal{C}$ such that there exists $w_f^\star \in \mathcal{C}$ that approximates $\{f^\star(\boldsymbol{x}_j) : j \in [n]\}$ well. To that end, first let

$$c_i^\star := c_0^{(i)} + \frac{1}{\sqrt{m}}v_c(u_0^{(i)}) \quad \text{and} \quad W_i^\star := W_0^{(i)} + \frac{1}{\sqrt{m}}v_W(u_0^{(i)}), \tag{21}$$

where (v_c, v_W) is the transportation mapping. Set $w_f^\star := [(c_i^\star, W_i^\star)]_{i\in[m]}$. Note that

$$\begin{aligned}\|w_f^\star - w_0\|_2^2 &= \sum_{i=1}^{m}(\underbrace{|c_i^\star - c_0^{(i)}|^2}_{\leq \frac{1}{m}\bar{v}_c^2} + \underbrace{\|W_i^\star - W_0^{(i)}\|_2^2}_{\leq \frac{1}{m}\bar{v}_W^2}) \\ &\leq \|(\bar{v}_c, \bar{v}_W)\|_2^2,\end{aligned}$$

hence we always have

$$w_f^\star \in \mathcal{B}_2(w_0, \|(\bar{v}_c, \bar{v}_W)\|_2) \subset \mathcal{C}. \tag{22}$$

Next, we show that $\max_{i\in[n]}|f^\star(\boldsymbol{x}_i) - \varphi(\boldsymbol{x}_i; w_f^\star)| \lesssim \sqrt{\frac{\log(2n/\delta)}{m}}$ with high probability over the random initialization to bound $\widehat{\mathcal{R}}_{\mathcal{S}}(w_f^\star)$. Fix $\boldsymbol{x} \in \mathcal{B}_2(0,1)$ first. Then,

$$\varphi_{\text{lin}}(\boldsymbol{x}; w_f^\star) := \underbrace{\varphi(\boldsymbol{x}; w_0)}_{=0} + \nabla^\top\varphi(\boldsymbol{x}; w_0)(w_f^\star - w_0) \tag{23}$$

$$= \frac{1}{m}\sum_{i\in[m]}\Big(\sigma(\langle W_0^{(i)}, \boldsymbol{x}\rangle)v_c(u_0^{(i)}) + c_0^{(i)}v_W^\top(u_0^{(i)})\boldsymbol{x}\sigma'(\langle W_0^{(i)}, \boldsymbol{x}\rangle)\Big) \xrightarrow[m\to\infty]{a.s.} f^\star(\boldsymbol{x}), \tag{24}$$

by the Strong Law of Large Numbers. Now, for each $i \in [m]$,

$$|\sigma(\langle W_0^{(i)}, \boldsymbol{x}\rangle)v_c(u_0^{(i)})| \leq \sigma_0\bar{v}_c,$$

and

$$|c_0^{(i)}v_W^\top(u_0^{(i)})\boldsymbol{x}\sigma'(\langle W_0^{(i)}, \boldsymbol{x}\rangle)| \leq \bar{v}_W\sigma_1|c_0^{(i)}|.$$

For the sub-Gaussian ψ_2-norm (see Definition 2.5.6 in Vershynin (2018)), these two inequalities imply

$$\|\sigma(\langle W_0^{(i)}, \boldsymbol{x}\rangle)v_c(u_0^{(i)})\|_{\psi_2} \leq \frac{\sigma_0\bar{v}_c}{\sqrt{\log 2}},$$
$$\|c_0^{(i)}v_W^\top(u_0^{(i)})\boldsymbol{x}\sigma'(\langle W_0^{(i)}, \boldsymbol{x}\rangle)\|_{\psi_2} \leq C\bar{v}_W\sigma_1$$

for an absolute constant (Exercise 2.5.8 in Vershynin (2018)). Let

$$X_i := \sigma(\langle W_0^{(i)}, \boldsymbol{x}\rangle)v_c(u_0^{(i)}) + c_0^{(i)}v_W^\top(u_0^{(i)})\boldsymbol{x}\sigma'(\langle W_0^{(i)}, \boldsymbol{x}\rangle), \quad i \in [m]. \tag{25}$$

Then, since $\|\cdot\|_{\psi_2}$ is a norm on sub-Gaussian random variables, we obtain

$$\|X_i - \mathbb{E}[X_i]\|_{\psi_2} \le 2\max\{\sigma_0/\sqrt{\log 2}, C\sigma_1\}\|(\bar{v}_c, \bar{v}_W)\|_2 =: K, \quad i \in [m].$$

Thus, by Generalized Hoeffding's inequality for ψ_2-bounded random variables (Theorem 2.6.3 in Vershynin (2018)), we obtain

$$\max\Big\{\Big|\frac{1}{m/2}\sum_{i=1}^{m/2} X_i - f^\star(\boldsymbol{x})\Big|, \Big|\frac{1}{m/2}\sum_{i=m/2+1}^{m} X_i - f^\star(\boldsymbol{x})\Big|\Big\} \le cK\sqrt{\frac{2\log(2/\delta)}{m}},$$

with probability at least $1-\delta$ for some absolute constant $c > 0$. Note that $u_0^{(i)}$ and $u_0^{(i+m/2)}$ are correlated for each $i \in [m/2]$ due to symmetric initialization, and the bound above is divided into two halves to handle this. Using triangle inequality, we obtain

$$\Big|\underbrace{\frac{1}{m}\sum_{i=1}^{m} X_i}_{=\varphi_{\mathrm{lin}}(\boldsymbol{x};w_f^\star)} - f^\star(\boldsymbol{x})\Big| \le 2cK\sqrt{\frac{2\log(2/\delta)}{m}}. \tag{26}$$

with probability at least $1-\delta$. Recall that the bound above holds for $\boldsymbol{x} \in \mathcal{B}_2(0,1)$, thus holds for $\boldsymbol{x}_j,\ j \in [n]$ individually. By using union bound, we obtain

$$\max_{j\in[n]}\Big|\varphi_{\mathrm{lin}}(\boldsymbol{x}_j; w_f^\star) - f^\star(\boldsymbol{x}_j)\Big| \le 2cK\sqrt{\frac{2\log(2n/\delta)}{m}}. \tag{27}$$

Finally, (17) implies that

$$\max_{j\in[n]}\Big|\varphi(\boldsymbol{x}_j; w_f^\star) - f^\star(\boldsymbol{x}_j)\Big| \le 2cK\sqrt{\frac{2\log(2n/\delta)}{m}} + \frac{1}{2}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2 \tag{28}$$

with probability at least $1-\delta$ over the random initialization. Hence,

$$\begin{aligned}
\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{w_f^\star}) &= \frac{1}{n}\sum_{j=1}^{n} \ell(\varphi(\boldsymbol{x}_j; w_f^\star); \boldsymbol{z}_j) \\
&= \frac{1}{n}\sum_{j=1}^{n} \ell(\varphi(\boldsymbol{x}_j; w_f^\star); \boldsymbol{z}_j) - \frac{1}{n}\sum_{j=1}^{n} \ell(f^\star(\boldsymbol{x}_j); \boldsymbol{z}_j) \\
&\le \frac{\mathrm{Lip}_{\ell,\mathcal{C}}}{n}\sum_{j=1}^{n} |\varphi(\boldsymbol{x}_j; w_f^\star) - f^\star(\boldsymbol{x}_j)| \\
&\le \mathrm{Lip}_{\ell,\mathcal{C}}\Big(2cK\sqrt{\frac{2\log(2n/\delta)}{m}} + \frac{1}{2}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} r_{\mathcal{C}}^2\Big).
\end{aligned} \tag{29}$$

Hence, $w_f^\star \in \mathcal{C}$ and $\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{w_f^\star}) \le c'$ with probability at least $1-\delta$ over the random initialization. Since

$$\bar{w} \in \arg\min_{w\in\mathcal{C}} \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_w),$$

we have $\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{\bar{w}}) \le \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{w_f^\star})$ and therefore

$$\mathbb{E}_0[\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{\widehat{w}_k})] \le \underbrace{\mathbb{E}_0[\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{\widehat{w}_k})] - \widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{\bar{w}})}_{(20)} + \underbrace{\widehat{\mathcal{R}}_{\mathcal{S}}(\varphi_{w_f^\star})}_{(29)}. \tag{30}$$

Substituting (20) and (29) into the above inequality concludes the proof.

□

B PROOFS FOR SECTION 5

Proof of Lemma 2. Consider $\widehat{w}_k, \widehat{w}'_k \in \mathcal{C}$. Then,

$$\begin{aligned}|\ell(\varphi(\boldsymbol{x};\widehat{w}_k);\boldsymbol{z}) - \ell(\varphi(\boldsymbol{x};\widehat{w}'_k);\boldsymbol{z}))| &\leq \mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\|\widehat{w}_k - \widehat{w}'_k\|_2 \\ &\leq \mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\frac{1}{k}\sum_{t=1}^{k}\|w_t - w'_t\|_2 \\ &\leq \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}}\frac{1}{k}\sum_{t=1}^{k}\|w_t - w'_t\|_{\bar{\boldsymbol{H}}_{t-1}}\end{aligned}$$

where the second line follows from the triangle inequality and convexity of the norm, and the third line follows from the fact that $\bar{\boldsymbol{H}}_k \succeq \lambda \boldsymbol{I}$. RHS of the above inequality is independent of $\boldsymbol{z} \in \mathscr{Z}$, thus

$$\sup_{\boldsymbol{z}\in\mathscr{Z}} |\ell(\varphi(\boldsymbol{x};\widehat{w}_k);\boldsymbol{z}) - \ell(\varphi(\boldsymbol{x};\widehat{w}'_k);\boldsymbol{z})| \leq \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}}\frac{1}{k}\sum_{t=1}^{k}\|w_t - w'_t\|_{\bar{\boldsymbol{H}}_{t-1}}.$$

Taking expectation over the optimization path concludes the proof. □

Proof of Theorem 2. Let $\pi_k := \pi_{\mathcal{C}}^{\boldsymbol{H}_k}$, $\pi'_k := \pi_{\mathcal{C}}^{\boldsymbol{H}'_k}$ and $\bar{\pi}_k := \pi_{\mathcal{C}}^{\bar{\boldsymbol{H}}_k}$ be the projection operators, and define

$$\begin{aligned}u_k &= w_k - \eta \boldsymbol{H}_k^{-1}\Psi_k(w_k) \\ u'_k &= w'_k - \eta[\boldsymbol{H}'_k]^{-1}\Psi'_k(w'_k).\end{aligned}$$

Then, we have the following error decomposition:

$$\|\Delta_{k+1}\|_{\bar{\boldsymbol{H}}_k} \leq \|\underbrace{\bar{\pi}_k u_k - \bar{\pi}_k u'_k}_{(A_k)}\|_{\bar{\boldsymbol{H}}_k} + \|\underbrace{\pi_k u_k - \bar{\pi}_k u_k}_{(B_k)}\|_{\bar{\boldsymbol{H}}_k} + \|\underbrace{\bar{\pi}_k u'_k - \pi'_k u'_k}_{(B'_k)}\|_{\bar{\boldsymbol{H}}_k}.$$

In this inequality, (B_k) and (B'_k) are error terms due to the metric (or projection) mismatch. We show that the critical term (A_k) will yield $\|\Delta_k\|_{\bar{\boldsymbol{H}}_{k-1}}$ plus controllable error terms with n and m.

$\mathcal{C}$ is compact and convex, and $\bar{\boldsymbol{H}}_k$ is positive definite, thus $\bar{\pi}_k$ is non-expansive (Nesterov, 2018; Brézis, 2011):

$$\|\bar{\pi}_k u_k - \bar{\pi}_k u'_k\|_{\bar{\boldsymbol{H}}_k} \leq \|u_k - u'_k\|_{\bar{\boldsymbol{H}}_k}.$$

Let $T_k(w) := w - \eta\bar{\boldsymbol{H}}_k^{-1}\Psi_k(w)$. Then, we further decompose (A_k) as follows:

$$\begin{aligned}\|u_k - u'_k\|_{\bar{\boldsymbol{H}}_k} \leq\ & \|\underbrace{T_k(w_k) - T_k(w'_k)}_{(A_{k,1})}\|_{\bar{\boldsymbol{H}}_k} + \eta\|\underbrace{\bar{\boldsymbol{H}}_k^{-1}(\Psi_k(w'_k) - \Psi'_k(w'_k))}_{(A_{k,2})}\|_{\bar{\boldsymbol{H}}_k} \\ &+ \eta\|\underbrace{(\bar{\boldsymbol{H}}_k^{-1} - \boldsymbol{H}_k^{-1})\Psi_k(w_k)}_{(A_{k,3})}\|_{\bar{\boldsymbol{H}}_k} + \eta\|\underbrace{\big((\boldsymbol{H}'_k)^{-1} - \bar{\boldsymbol{H}}_k^{-1}\big)\Psi'_k(w'_k)}_{(A'_{k,3})}\|_{\bar{\boldsymbol{H}}_k}.\end{aligned}$$

In this decomposition, $(A_{k,3})$ and $(A'_{k,3})$ correspond to the *preconditioner mismatch* terms.

Bounding $(A_{k,1})$. The first part is to establish an approximate co-coercivity result in the sense of Baillon and Haddad (1977).

Lemma 7 (Approximate co-coercivity)**.** *Let $\varepsilon = B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}$ and $\Lambda = 2\varepsilon + B\mathrm{Lip}^2_{\varphi,\mathcal{C}}$. Then, for any $u, v \in \mathcal{C}$, we have*

$$\Big(\Psi_k(u) - \Psi_k(v)\Big)^\top(u - v) \geq \frac{1}{2\Lambda}\|\Psi_k(u) - \Psi_k(v)\|_2^2 - 2\varepsilon\|u - v\|_2^2. \tag{31}$$

Proof of Lemma 7. The proof builds on Baillon and Haddad (1977); see (Beck, 2017, Theorem 5.8). First, note the following decomposition of the Hessian:

$$\nabla^2 \Phi_k(w) = \boldsymbol{J}_k^\top(w)\boldsymbol{J}_k(w) + \sum_{i \in I_k} \ell'(\varphi(\boldsymbol{x}_i; w); \boldsymbol{z}_i)\nabla^2 \varphi(\boldsymbol{x}_i; w). \tag{32}$$

Since $\boldsymbol{J}_k^\top(w)\boldsymbol{J}_k(w) \succeq 0$, we have

$$-\varepsilon := -B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} \leq \lambda_{\min}(\nabla^2\Phi_k(w)) \leq \lambda_{\max}(\nabla^2\Phi_k(w)) \leq B\mathrm{Lip}_{\varphi,\mathcal{C}}^2 + \varepsilon$$

by Weyl's inequality (Horn and Johnson, 2012). Let

$$\Phi_k^\varepsilon(w) = \Phi_k(w) + \frac{1}{2}\varepsilon\|w\|_2^2.$$

Then, the above inequality implies

$$\nabla^2\Phi_k^\varepsilon(w) \succeq 0,$$

and we also have

$$\nabla\Phi_k^\varepsilon(w) = \nabla\Phi_k(w) + \varepsilon w.$$

For any $u, v \in \mathcal{C}$,

$$\nabla\Phi_k^\varepsilon(u) - \nabla\Phi_k^\varepsilon(v) = \int_0^1 \nabla^2\Phi_k^\varepsilon((1-s)u + sv)ds\Big(u - v\Big) = M(u-v).$$

Then,

$$\begin{aligned}
\|\nabla\Phi_k^\varepsilon(u) - \nabla\Phi_k^\varepsilon(v)\|_2^2 &= (u-v)^\top M^2(u-v) \\
&\leq \Lambda(u-v)^\top M(u-v) \\
&= \Lambda(u-v)^\top\Big(\nabla\Phi_k^\varepsilon(u) - \nabla\Phi_k^\varepsilon(v)\Big) \\
&= \Lambda(u-v)^\top(\nabla\Phi_k(u) - \nabla\Phi_k(v)) + \varepsilon\Lambda\|u-v\|_2^2.
\end{aligned}$$

Also,

$$\|\nabla\Phi_k(u) - \nabla\Phi_k(v)\|_2^2 \leq 2\|\nabla\Phi_k^\varepsilon(u) - \nabla\Phi_k^\varepsilon(v)\|_2^2 + 2\varepsilon^2\|u-v\|_2^2.$$

Hence,

$$\frac{1}{2}\|\nabla\Phi_k(u) - \nabla\Phi_k(v)\|_2^2 - \varepsilon^2\|u-v\|_2^2 \leq \Lambda(u-v)^\top(\nabla\Phi_k(u) - \nabla\Phi_k(v)) + \varepsilon\Lambda\|u-v\|_2^2,$$

and therefore,

$$\|\nabla\Phi_k(u) - \nabla\Phi_k(v)\|_2^2 \leq 2\Lambda(u-v)^\top(\nabla\Phi_k(u) - \nabla\Phi_k(v)) + 2\varepsilon(\varepsilon+\Lambda)\|u-v\|_2^2,$$

and equivalently

$$\begin{aligned}
(u-v)^\top(\nabla\Phi_k(u) - \nabla\Phi_k(v)) &\geq \frac{1}{2\Lambda}\|\nabla\Phi_k(u) - \nabla\Phi_k(v)\|_2^2 - \varepsilon\frac{\varepsilon+\Lambda}{\Lambda}\|u-v\|_2^2 \\
&\geq \frac{1}{2\Lambda}\|\nabla\Phi_k(u) - \nabla\Phi_k(v)\|_2^2 - 2\varepsilon\|u-v\|_2^2
\end{aligned}$$

since $\Lambda \geq \varepsilon$. □

Using Lemma 7, we obtain a non-expansivity result for T_k.

Lemma 8 (Approximate non-expansivity of T_k)**.** *For any $u, v \in \mathcal{C}$, if (η, λ) satisfy the conditions in the paper, then we have the (approximate) non-expansivity*

$$\|T_k(u) - T_k(v)\|_{\bar{\boldsymbol{H}}_k} \leq \|u-v\|_{\bar{\boldsymbol{H}}_{k-1}} + 2r_{\mathcal{C}}\sqrt{\eta\varepsilon} + 2\sqrt{2\alpha}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}Br_{\mathcal{C}}\Big(\frac{r_{\mathcal{C}}}{2} + \frac{\sqrt{B}\mathrm{Lip}_{\ell,\mathcal{C}}}{\mu_0\nu}\Big)$$

almost surely.

Proof of Lemma 8. First we make the following decomposition:

$$\|T_k(u)-T_k(v)\|_{\bar{\boldsymbol{H}}_k}^2 = \|u-v\|_{\bar{\boldsymbol{H}}_k}^2 - 2\eta(u-v)^\top\Big(\Psi_k(u)-\Psi_k(v)\Big) + \eta^2\|\Psi_k(u)-\Psi_k(v)\|_{\bar{\boldsymbol{H}}_k^{-1}}^2.$$

Using Lemma 7, we obtain

$$\|T_k(u)-T_k(v)\|_{\bar{\boldsymbol{H}}_k}^2 \le \|u-v\|_{\bar{\boldsymbol{H}}_k}^2 - \Big(\frac{\eta}{\Lambda}-\frac{\eta^2}{\lambda}\Big)\|\Psi_k(u)-\Psi_k(v)\|_2^2 + 4\eta\varepsilon\|u-v\|_2^2. \tag{33}$$

By choosing $\eta \le \frac{\lambda}{2\Lambda}$, the above implies

$$\|T_k(u)-T_k(v)\|_{\bar{\boldsymbol{H}}_k}^2 \le \|u-v\|_{\bar{\boldsymbol{H}}_k}^2 - \frac{\eta}{2\Lambda}\|\Psi_k(u)-\Psi_k(v)\|_2^2 + 4\eta\varepsilon\|u-v\|_2^2. \tag{34}$$

Note that

$$\begin{aligned}\|u-v\|_{\bar{\boldsymbol{H}}_k}^2 &= \|u-v\|_{\bar{\boldsymbol{H}}_{k-1}}^2 + \frac{\alpha}{2}\Big(\|\boldsymbol{J}_k(w_k)(u-v)\|_2^2 + \|\boldsymbol{J}_k'(w_k')(u-v)\|_2^2\Big)\\ &\le \|u-v\|_{\bar{\boldsymbol{H}}_{k-1}}^2 + 2\alpha\|\boldsymbol{J}_k(u)(u-v)\|_2^2 + 2\alpha B\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^4\\ &\le \|u-v\|_{\bar{\boldsymbol{H}}_{k-1}}^2 + 2\alpha\|\boldsymbol{J}_k(u)(u-v)\|_1^2 + 2\alpha B\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 r_{\mathcal{C}}^4\end{aligned} \tag{35}$$

Now,

$$\begin{aligned}\|\Psi_k(u)-\Psi_k(v)\|_2 &\ge \|\boldsymbol{J}_k^\top(u)\Big(\boldsymbol{G}_k(u)-\boldsymbol{G}_k(v)\Big)\|_2 - \|\Big(\boldsymbol{J}_k(u)-\boldsymbol{J}_k(v)\Big)^\top\boldsymbol{G}_k(v)\|_2\\ &\ge \|\boldsymbol{J}_k^\top(u)\Big(\boldsymbol{G}_k(u)-\boldsymbol{G}_k(v)\Big)\|_2 - B\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}\mathrm{Lip}_{\ell,\mathcal{C}} r_{\mathcal{C}}.\end{aligned}$$

Using Assumption 2, we obtain

$$\|\boldsymbol{J}_k^\top(u)\Big(\boldsymbol{G}_k(u)-\boldsymbol{G}_k(v)\Big)\|_2 \ge \mu_0\|\boldsymbol{G}_k(u)-\boldsymbol{G}_k(v)\|_2.$$

We also have

$$\|\boldsymbol{G}_k(u)-\boldsymbol{G}_k(v)\|_2 \ge \frac{\nu}{\sqrt{B}}\sum_{j\in I_k}|\varphi(\boldsymbol{x}_j;u)-\varphi(\boldsymbol{x}_j;v)|.$$

From Lipschitz smoothness,

$$\|\boldsymbol{G}_k(u)-\boldsymbol{G}_k(v)\|_2 \ge \frac{\nu}{\sqrt{B}}\|\boldsymbol{J}_k(u)(u-v)\|_1 - \frac{\nu\sqrt{B}}{2}r_{\mathcal{C}}^2\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}.$$

Hence,

$$\|\Psi_k(u)-\Psi_k(v)\|_2 \ge \mu_0\Big(\frac{\nu}{\sqrt{B}}\|\boldsymbol{J}_k(u)(u-v)\|_1 - \frac{\nu\sqrt{B}}{2}r_{\mathcal{C}}^2\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}\Big) - B\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}\mathrm{Lip}_{\ell,\mathcal{C}}r_{\mathcal{C}},$$

which implies that

$$\|\boldsymbol{J}_k(u)(v-u)\|_1^2 \le \frac{2B}{\mu_0^2\nu^2}\|\Psi_k(u)-\Psi_k(v)\|_2^2 + 4\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 B^2 r_{\mathcal{C}}^2\Big(\frac{r_{\mathcal{C}}^2}{4}+\frac{B\mathrm{Lip}_{\ell,\mathcal{C}}^2}{\nu^2\mu_0^2}\Big). \tag{36}$$

By substituting (35) and (36) into (34), we obtain

$$\begin{aligned}\|T_k(u)-T_k(v)\|_{\bar{\boldsymbol{H}}_k}^2 \le \|u-v\|_{\bar{\boldsymbol{H}}_{k-1}}^2 &- \Big(\frac{\eta}{2\Lambda}-\frac{4\alpha B}{\mu_0^2\nu^2}\Big)\|\Psi_k(u)-\Psi_k(v)\|_2^2\\ &+ 8\alpha\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}^2 B^2 r_{\mathcal{C}}^2\Big(\frac{r_{\mathcal{C}}^2}{4}+\frac{B\mathrm{Lip}_{\ell,\mathcal{C}}^2}{\nu^2\mu_0^2}\Big) + 4\eta\varepsilon r_{\mathcal{C}}^2\end{aligned}$$

The choice of η/α cancels the second term, thereby concluding the proof. □

Bounding $(A_{k,2})$. This term corresponds to *gradient mismatch.* Since $\{j^\star \notin I_k\} \subset \{\Psi_k(\cdot) = \Psi_k'(\cdot)\}$, we obtain

$$\begin{aligned}
&\|\bar{\boldsymbol{H}}_k^{-1}(\Psi_k(w_k') - \Psi_k'(w_k'))\|_{\bar{\boldsymbol{H}}_k} \\
&\qquad \le 2\sup_{w\in\mathcal{C}} \|\Psi_k(w)\|_{\bar{\boldsymbol{H}}_k^{-1}} \mathbf{1}_{\{j^\star \in I_k\}} \\
&\qquad \le \frac{2B}{\lambda_{\min}^{1/2}(\bar{\boldsymbol{H}}_k)} \mathrm{Lip}_{\ell,\mathcal{C}} \mathrm{Lip}_{\varphi,\mathcal{C}} \mathbf{1}_{\{j^\star \in I_k\}}.
\end{aligned} \tag{37}$$

Here, note that $\mathbb{E}[\mathbf{1}_{\{j^\star \in I_k\}}] = \mathcal{P}(j^\star \in I_k) = \frac{B}{n}$ for any $k \in \mathbb{N}$. In the worst-case scenario,

$$\lambda_{\min}(\bar{\boldsymbol{H}}_k) \ge \lambda,$$

which implies that

$$\mathbb{E}\|\bar{\boldsymbol{H}}_k^{-1}(\Psi_k(w_k') - \Psi_k'(w_k'))\|_{\bar{\boldsymbol{H}}_k} \le \frac{2B^2}{\lambda^{1/2} n} \mathrm{Lip}_{\ell,\mathcal{C}} \mathrm{Lip}_{\varphi,\mathcal{C}}.$$

Bounding $(A_{k,3})$ and $(A_{k,3}')$. To bound these error terms that stem from the preconditioner mismatch, we use

$$\|(\bar{\boldsymbol{H}}_k^{-1} - \boldsymbol{H}_k^{-1})\Psi_k(w_k)\|_{\bar{\boldsymbol{H}}_k} \le \|\bar{\boldsymbol{H}}_k^{-1/2}(\bar{\boldsymbol{H}}_k - \boldsymbol{H}_k)\boldsymbol{H}_k^{-1}\|_2 \|\Psi_k(w_k)\|_2. \tag{38}$$

We have

$$\begin{aligned}
\lambda_{\max}(\bar{\boldsymbol{H}}_k) &\le \alpha B(k+1)\mathrm{Lip}_{\varphi,\mathcal{C}}^2 + \lambda, \\
\sup_{w\in\mathcal{C}} \|\Psi_k(w)\|_2 &\le B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}
\end{aligned}$$

and

$$\|\boldsymbol{H}_k' - \bar{\boldsymbol{H}}_k\|_2 = \|\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k\|_2 = \frac{1}{2}\|\boldsymbol{H}_k - \boldsymbol{H}_k'\|_2.$$

The following lemma upper bounds the error term $\mathbb{E}\|\boldsymbol{H}_k - \boldsymbol{H}_k'\|_2$.

Lemma 9 (Stability of $\boldsymbol{H}_k$)**.** *For any $k \in \mathbb{N}$,*

$$\mathbb{E}[\|\boldsymbol{H}_k - \boldsymbol{H}_k'\|_2] \le 2B\alpha(k+1)r_{\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \alpha\mathrm{Lip}_{\varphi,\mathcal{C}}^2\frac{(k+1)B}{n}.$$

Proof of Lemma 9. By using the uniform $\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}$-Lipschitz continuity of $\nabla\varphi$, we have

$$\|\boldsymbol{H}_k - \boldsymbol{H}_k'\|_2 \le \alpha\mathrm{Lip}_{\varphi,\mathcal{C}}^2 \sum_{t=0}^{k} \mathbf{1}_{\{j^\star \in I_t\}} + 2\alpha(k+1)Br_{\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}.$$

The result is proved by taking expectation over I_t. □

Lemma 10 (Metric mismatch)**.** *For any $k \in \mathbb{N}$, we obtain*

$$\begin{aligned}
&\|\pi_k u_k - \bar{\pi}_k u_k\|_{\bar{\boldsymbol{H}}_k} \\
&\qquad \le \|\bar{\boldsymbol{H}}_k^{-1/2}(\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k)\|_2 \sup_{w\in\mathcal{C}} \|w - u_k\|_2 \\
&\qquad \le \frac{\|\boldsymbol{H}_k - \boldsymbol{H}_k'\|_2}{2\lambda_{\min}^{1/2}(\bar{\boldsymbol{H}}_k)} \sup_{w\in\mathcal{C}} \|w - u_k\|_2
\end{aligned} \tag{39}$$

Also, for any $w \in \mathcal{C}$, $\|w - u_k\|_2 \le r_{\mathcal{C}} + \frac{\eta B}{\lambda}\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}$.

Proof. First,

$$\begin{aligned}
\|w - u_k\|_2 &= \|w - w_k + \eta\boldsymbol{H}_k^{-1}\Psi_k(w_k)\|_2 \\
&\le \|w - w_k\|_2 + \frac{\eta}{\lambda}\|\Psi_k(w_k)\|_2 \\
&\le r_{\mathcal{C}} + \frac{\eta}{\lambda}B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}.
\end{aligned}$$

We will use the following geometric argument due to Bubeck (2015) (Lemma 3.1):

$$\big(\pi_k u_k - \bar{\pi}_k u_k\big)^\top \bar{\boldsymbol{H}}_k\big(u_k - \bar{\pi}_k u_k\big) \le 0,$$
$$\big(\pi_k u_k - \bar{\pi}_k u_k\big)^\top \boldsymbol{H}_k\big(\pi_k u_k - u_k\big) \le 0.$$

Let $x_k := \pi_k u_k - \bar{\pi}_k u_k$. Then, the second inequality yields

$$x_k^\top \boldsymbol{H}_k(u_k - \pi_k u_k) \ge 0. \tag{40}$$

Now, the first inequality yields

$$x_k^\top \bar{\boldsymbol{H}}_k(u_k - \pi_k u_k) + \|x_k\|^2_{\bar{\boldsymbol{H}}_k} \le 0.$$

Therefore,

$$\begin{aligned}\|x_k\|^2_{\bar{\boldsymbol{H}}_k} &\le -x_k^\top \bar{\boldsymbol{H}}_k(u_k - \pi_k u_k)\\ &\le -x_k^\top(\bar{\boldsymbol{H}}_k - \boldsymbol{H}_k)(u_k - \pi_k u_k),\end{aligned}$$

where the second line stems from (40). By using Cauchy-Schwarz inequality, we obtain

$$\|x_k\|_{\bar{\boldsymbol{H}}_k} \le \|\bar{\boldsymbol{H}}_k^{-1/2}(\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k)\|_2 \, \|u_k - \pi_k u_k\|_2.$$

Hence,

$$\begin{aligned}\|\pi_k u_k - \bar{\pi}_k u_k\|_{\bar{\boldsymbol{H}}_k} &\le \lambda_{\min}^{-1/2}(\bar{\boldsymbol{H}}_k)\|\bar{\boldsymbol{H}}_k - \boldsymbol{H}_k\|_2 \|u_k - \pi_k u_k\|_2\\ &\le \lambda_{\min}^{-1/2}(\bar{\boldsymbol{H}}_k)\|\bar{\boldsymbol{H}}_k - \boldsymbol{H}_k\|_2 \sup_{w\in\mathcal{C}} \|u_k - w\|_2.\end{aligned}$$

Finally, since

$$\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k = \frac{1}{2}(\boldsymbol{H}_k - \boldsymbol{H}_k'),$$

we obtain

$$\begin{aligned}\|\bar{\boldsymbol{H}}_k^{-1/2}(\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k)\|_2 &\le \|\bar{\boldsymbol{H}}_k^{-1/2}\|_2 \, \|\boldsymbol{H}_k - \bar{\boldsymbol{H}}_k\|_2\\ &\le \frac{1}{2\lambda_{\min}^{1/2}(\bar{\boldsymbol{H}}_k)}\|\boldsymbol{H}_k - \boldsymbol{H}_k'\|_2,\end{aligned}$$

which concludes the proof. □

Define the event

$$\bar{\mathcal{E}} = \Big\{ \min\{\lambda_{\min}(\boldsymbol{H}_t), \lambda_{\min}(\boldsymbol{H}_t')\} \ge \lambda_t,\ t \in \mathbb{N}\Big\} \tag{41}$$

for $\{\lambda_t : t \in \mathbb{N}\}$. We always have $\inf_{t\in\mathbb{N}} \lambda_t \ge \lambda > 0$. Now, we summarize the bounds found earlier in this proof as follows:

$$\mathbb{E}[(A_{t,1}); \bar{\mathcal{E}}] \le \mathbb{E}\big[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}}; \bar{\mathcal{E}}\big] + \underbrace{2r_{\mathcal{C}}\sqrt{\eta\varepsilon} + 2\sqrt{2\alpha}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} B r_{\mathcal{C}}\Big(\frac{r_{\mathcal{C}}}{2} + \frac{\sqrt{B}\mathrm{Lip}_{\ell,\mathcal{C}}}{\mu_0\nu}\Big)}_{=:Z_1},$$

$$\mathbb{E}[(A_{t,2}); \bar{\mathcal{E}}] \le \frac{2B^2}{\lambda_t^{1/2} n}\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}} =: \frac{1}{n}\lambda_t^{-1/2}\underbrace{2B^2\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}_{=:Z_2},$$

$$\mathbb{E}[(A_{t,3}) + (A_{t,3}'); \bar{\mathcal{E}}] \le \alpha(t+1)\lambda_t^{-3/2}\underbrace{B^2\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}_{=:Z_3}\underbrace{\Big(2r_{\mathcal{C}}\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \mathrm{Lip}^2_{\varphi,\mathcal{C}}\frac{1}{n}\Big)}_{=:Z_0^{(m,n)}},$$

$$\mathbb{E}[(B_t) + (B_t'); \bar{\mathcal{E}}] \le \alpha(t+1)\lambda_t^{-1/2}\underbrace{\Big(r_{\mathcal{C}} + \frac{\eta B}{\lambda}\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\Big)B}_{=:Z_4}\underbrace{\Big(2r_{\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{\mathrm{Lip}^2_{\varphi,\mathcal{C}}}{n}\Big)}_{=Z_0^{(m,n)}}$$

Consequently, we have

$$\mathbb{E}[\|\Delta_{t+1}\|_{\boldsymbol{H}_t};\bar{\mathcal{E}}] \le \mathbb{E}[\|\Delta_t\|_{\boldsymbol{H}_{t-1}};\bar{\mathcal{E}}] + Z_1 + \eta Z_2 \lambda_t^{-1/2}\frac{1}{n} + \alpha\frac{t+1}{\lambda_t^{1/2}}\Big(\eta\frac{Z_3}{\lambda_t} + Z_4\Big)Z_0^{(m,n)}.$$

Telescoping sum over $t = 0, 1, 2, \ldots, k-1$ yields

$$\mathbb{E}[\|\Delta_k\|_{\boldsymbol{H}_{k-1}};\bar{\mathcal{E}}] \le kZ_1 + \frac{\eta Z_2}{n}\sum_{t=0}^{k-1}\lambda_t^{-1/2} + \alpha\eta Z_3 Z_0^{(m,n)}\sum_{t=0}^{k-1}\frac{t+1}{\lambda_t^{3/2}} + \alpha Z_4 Z_0^{(m,n)}\sum_{t=0}^{k-1}\frac{t+1}{\lambda_t^{1/2}}. \tag{42}$$

Now, note that $\mathbb{P}(\bar{\mathcal{E}}) = 1$ if $\lambda_t = \lambda$ for all $t \in \mathbb{N}$. Then, we obtain the worst-case bound:

$$\mathbb{E}\|\Delta_k\|_{\boldsymbol{H}_{k-1}} \le kZ_1 + \frac{\eta k Z_2}{\lambda^{1/2} n} + \alpha\eta Z_3 Z_0^{(m,n)}\frac{k^2}{\lambda^{3/2}} + \alpha Z_4 Z_0^{(m,n)}\frac{k^2}{\lambda^{1/2}}.$$

By symmetry, the same recursion holds with $\boldsymbol{H}_t'$ in place of $\boldsymbol{H}_t$. By averaging the two recursions and using $\|v\|_{\bar{\boldsymbol{H}}_t}^2 = \frac{1}{2}\|v\|_{\boldsymbol{H}_t}^2 + \frac{1}{2}\|v\|_{\boldsymbol{H}_t'}^2$, we obtain the corresponding bound with respect to the midpoint-metric. □

Remark 8. *Fix $\lambda > 0$, and set $\eta = \frac{C}{k}$ and $\alpha = \frac{\xi C}{k}$ for some $C > 0$ that satisfies the conditions in (8). Then, we obtain*

$$\mathbb{E}\|\Delta_k\|_{\boldsymbol{H}_{k-1}} \lesssim_C kZ_1 + \frac{Z_2}{\lambda^{1/2} n} + Z_3 Z_0^{(m,n)}\frac{1}{\lambda^{3/2}} + \frac{Z_4 Z_0^{(m,n)} k}{\lambda^{1/2}} \text{ for } k \ge 1.$$

Now, note that

$$Z_1 \lesssim r_{\mathcal{C}}\frac{1}{\sqrt{k}}\left(\sqrt{B\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}} + Br_{\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}}\Big(\frac{r_{\mathcal{C}}}{2} + \frac{\sqrt{B}\mathrm{Lip}_{\ell,\mathcal{C}}}{\mu_0\nu}\Big)\right) \lesssim \frac{\sqrt{B}}{\sqrt{k\sqrt{m}}} + \frac{Br_{\mathcal{C}}}{\sqrt{m}}(r_{\mathcal{C}} + \sqrt{B}),$$

$$Z_0^{(m,n)} \lesssim r_{\mathcal{C}}\mathrm{Lip}_{\nabla\varphi,\mathcal{C}} + \frac{1}{n} \lesssim \frac{r_{\mathcal{C}}}{\sqrt{m}} + \frac{1}{n}.$$

Substituting these into the stability bound, we obtain

$$\begin{aligned}\mathbb{E}\|\Delta_k\|_{\boldsymbol{H}_{k-1}} &\lesssim \frac{\sqrt{B\cdot k}}{m^{1/4}} + \frac{B\sqrt{k}}{\sqrt{m}}(1+\sqrt{B}) + \frac{B^2}{\lambda^{1/2}n} + \frac{B^2}{\lambda^{3/2}}\Big(\frac{1}{\sqrt{m}} + \frac{1}{n}\Big) + \frac{k}{\lambda^{1/2}}\Big(\frac{1}{\sqrt{m}} + \frac{1}{n}\Big),\\ &\lesssim \frac{\sqrt{Bk}}{m^{1/4}} + \frac{1}{\sqrt{m}}\Big(B\sqrt{k} + B^{3/2}\sqrt{k} + B^2\lambda^{-3/2} + k\lambda^{-1/2}\Big) + \frac{1}{n}\Big(\frac{B^2}{\lambda^{1/2}} + \frac{B^2}{\lambda^{3/2}} + \frac{k}{\lambda^{1/2}}\Big).\end{aligned}$$

Proposition 2 is proved via exactly the same methodology, whereby we substitute the lower bound for λ_t into the stability bound.

Proof of Corollary 3. For each $i \in [n]$, let $\mathcal{S}^{(i)}$ denote the sample obtained by replacing the i-th example, and let $\widehat{w}_k, \widehat{w}_k^{(i)}$ be the corresponding Polyak-Ruppert averaged SGN iterates driven by the same realization of the subsampling process $\{I_k : k \in \mathbb{N}\}$. For arbitrary $\boldsymbol{z}$, let

$$X_i := \ell(\varphi(\boldsymbol{x};\widehat{w}_k);\boldsymbol{z}) - \ell(\varphi(\boldsymbol{x};\widehat{w}_k^{(i)});\boldsymbol{z}).$$

Then, we decompose the expectation as

$$\mathbb{E}[|X_i|] = \mathbb{E}[|X_i|;\mathcal{E}_i] + \mathbb{E}[|X_i|;\mathcal{E}_i^c].$$

Lemma 2 implies that, on the event $\mathcal{E}_i$,

$$\mathbb{E}\big[|X_i|\big|\mathcal{S},\mathcal{S}'\big] \le \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}k}\sum_{t=1}^{k}\mathbb{E}[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}}|\mathcal{S},\mathcal{S}'].$$

On the event $\mathcal{E}_i^c$, since both iterates remain in $\mathcal{C}$ as a consequence of projection, we have

$$|X_i| \le \mathrm{Lip}_{\ell,\mathcal{C}}\,|\varphi(\boldsymbol{x};\widehat{w}_k) - \varphi(\boldsymbol{x};\widehat{w}_k^{(i)})| \le \mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}\,r_{\mathcal{C}} = \Upsilon_{\mathcal{C}}.$$

In both cases, the upper bound is independent of $\boldsymbol{z}$, which implies that

$$\sup_{\boldsymbol{z}} \mathbb{E}\big[|\ell(\varphi(\boldsymbol{x};\widehat{w}_k);\boldsymbol{z}) - \ell(\varphi(\boldsymbol{x};\widehat{w}_k^{(i)});\boldsymbol{z})\big|\mathcal{S},\mathcal{S}'\big] \leq \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}k}\sum_{t=1}^{k}\mathbb{E}[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}}|\mathcal{S},\mathcal{S}']\mathbf{1}_{\mathcal{E}_i} + \Upsilon_{\mathcal{C}}\mathbf{1}_{\mathcal{E}_i^c}.$$

By taking expectation and noting that

$$\mathbb{E}[|X_i|\mathbf{1}_{\mathcal{E}_i^c}] \leq \Upsilon_{\mathcal{C}}\,\mathbb{P}(\mathcal{E}_i^c) \leq \Upsilon_{\mathcal{C}}\delta,$$

we obtain

$$\mathbb{E}\Big[\sup_{\boldsymbol{z}} \mathbb{E}\big[|\ell(\varphi(\boldsymbol{x};\widehat{w}_k);\boldsymbol{z}) - \ell(\varphi(\boldsymbol{x};\widehat{w}_k^{(i)});\boldsymbol{z})\big|\mathcal{S},\mathcal{S}'\big]\Big] \leq \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}k}\sum_{t=1}^{k}\mathbb{E}[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}}\mathbf{1}_{\mathcal{E}_i}] + \Upsilon_{\mathcal{C}}\delta.$$

Hence, the bound in (6) yields

$$\big|\mathbb{E}[\mathrm{gen}(\varphi_{\widehat{w}_k};\mathcal{S})]\big| \leq \frac{\mathrm{Lip}_{\ell,\mathcal{C}}\mathrm{Lip}_{\varphi,\mathcal{C}}}{\sqrt{\lambda}kn}\sum_{t=1}^{k}\sum_{i=1}^{n}\mathbb{E}[\|\Delta_t\|_{\bar{\boldsymbol{H}}_{t-1}}\mathbf{1}_{\mathcal{E}_i}] + \delta\Upsilon_{\mathcal{C}}.$$

□