UNVEILING LOOKAHEAD'S POTENTIAL IN SHARPNESS-AWARE MINIMIZATION: CONVERGENCE AND GENERALIZATION REVISITED

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models have achieved remarkable success across various applications, but over-parameterized networks often suffer from overfitting and limited generalization, limiting their real-world performance. Sharpness-Aware Lookahead (SALA), which integrates Sharpness-Aware Minimization (SAM) with the lookahead mechanism, has been proposed to improve generalization, training stability, and convergence. However, recent theoretical studies suggest that lookahead may delay convergence, which appears to contradict empirical observations. This discrepancy raises a fundamental question: how does lookahead affect the generalization and optimization behavior of SAM? To answer this question, we develop a systematic theoretical framework to derive tighter generalization and excess risk bounds for SALA under relaxed assumptions. Our results show that the lookahead parameter α affects optimization and generalization in opposite directions: 1) the optimization bound $O\left(\frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}}\right)$ decays faster as α increases, while 2) the generalization bound $O\left(\frac{T^{\alpha}\log T}{n}\right)$ increases with α through both the leading factor α and the power-law term T^{α}. We present detailed comparisons with prior theoretical studies and empirical observations to support the validity of our findings. Furthermore, the analysis extended to the stochastic gradient descent-based lookahead algorithm uncovers a similar impact of lookahead on generalization and optimization behaviors from both theoretical and empirical perspectives.

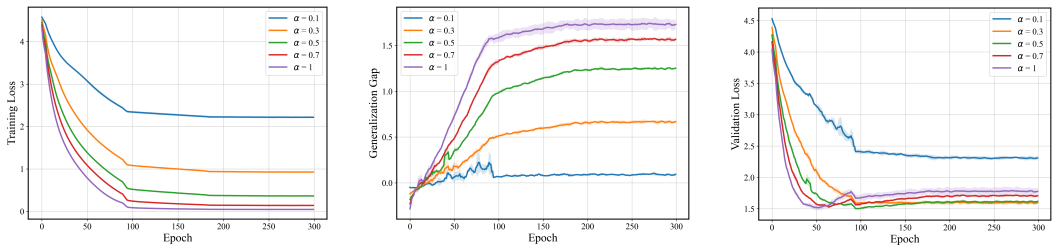
1 INTRODUCTION

Deep learning models have achieved significant success across diverse practical applications (LeCun et al., 2015; Mennella et al., 2024; Maleki Varnosfaderani & Forouzanfar, 2024). However, over-parameterized networks can overfit the training data, and achieving a low training loss does not necessarily guarantee good generalization (Srivastava et al., 2014; Zhang et al., 2016). Enhancing model generalization remains a critical challenge. A promising direction emerges from the analysis of landscapes for non-convex loss function, where empirical evidence increasingly suggests a strong correlation between the flatness of a solution and its generalization capabilities (Keskar et al., 2016; Jiang et al., 2020). To guide models toward flatter solutions, Foret et al. (2021) proposed the Sharpness-Aware Minimization (SAM) algorithm, formulating the objective as a min-max optimization problem that explicitly minimizes the loss sharpness within a parameter neighborhood:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\|\boldsymbol{\varepsilon}\|_2\leq\rho} F_S(\mathbf{w}+\boldsymbol{\varepsilon}),$$

where $\mathbf{w} \in \mathbb{R}^d$ denotes the model parameters and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is an adversarial perturbation constrained by $\|\boldsymbol{\varepsilon}\|_2 \leq \rho$. Here, $\rho > 0$ is the perturbation radius and $\|\cdot\|_2$ denotes the Euclidean norm. The objective $F_S(\cdot)$ is the empirical risk on the training set $S = \{z_i\}_{i=1}^n$, defined as $F_S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}, z_i)$, where $f(\mathbf{w}, z_i)$ denotes the per-sample training loss. Subsequently, SAM has garnered remarkable success in fields such as visual computing (Brock et al., 2021; Chen et al., 2022; Li et al., 2024), graph neural networks (Luo et al., 2024) and natural language processing (Bahri et al., 2022). Despite its success in improving model generalization, SAM has raised concerns in recent optimization studies

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



(a) Training loss (b) Generalization gap (c) Validation loss

Figure 1: Results of ResNet-18 optimized with SALA on CIFAR-100 under different outer-loop learning rates $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. Each solid curve reports the mean over three random seeds $\{0, 1, 2\}$, and the shaded region denotes one standard deviation. The training loss is recorded at the end of each epoch, and the generalization gap is computed as the validation loss minus the training loss. As α increases, convergence accelerates, while smaller α leads to a smaller generalization gap, indicating better generalization.

(Tan et al., 2025). Both empirical findings and theoretical investigations suggest that SAM may be more prone to becoming trapped in saddle points regions compared to stochastic gradient descent (SGD) (Compagnoni et al., 2023; Kim et al., 2023). Moreover, it does not always ensure arrival at the optimal solution (Si & Yun, 2023; Oikonomou & Loizou, 2025). One potential strategy to mitigate this issue is to encourage the optimizer to more thoroughly explore the local loss landscape beforehand, thereby accumulating richer information and following a more stable and favorable convergence trajectory.

To this end, the Sharpness-Aware Lookahead (SALA) scheme has been developed by (Yu et al., 2024; Tan et al., 2024), which leverages SAM as the base optimizer and incorporates the lookahead mechanism. Despite SALA has shown promising empirical results (Yu et al., 2024; Tan et al., 2024), recent theoretical works (Zhou et al., 2021; Li & Lei, 2025) indicate that integrating lookahead with SGD brings a noticeable convergence delay. This finding appears to contradict some empirical observations reported in the literature, leading the following critical question to reassess the impact of lookahead on the convergence and generalization performance of SAM.

How does lookahead actually affect the generalization and optimization of SAM?

To address this question, we aim to provide a systematic analysis that bridges the gap between theory and practice regarding the impact of lookahead on SAM. The main contributions of this paper are summarized as follows:

- For the first time, a theoretical analysis framework is established for SALA in the non-convex setting, generalizing the convex analysis of Tan et al. (2024). By introducing an almost co-coercivity condition, we derive a sharper generalization bound $O\left(\frac{T^\alpha \log T}{n}\right)$ with a weaker (sublinear) dependence on the outer-loop counts T, yielding a milder dependence on T and k than the prior convex analysis $O\left(\frac{\alpha T k}{n}\right)$. On the optimization side, we achieve a convergence rate consistent with Tan et al. (2024) without requiring bounded gradients. We also extend our analysis beyond the smooth setting by allowing Hölder continuous gradients. To the best of our knowledge, this yields the first generalization guarantee for SALA in the non-smooth setting.
- We systematically analyze how key factors (T, k, and outer-loop learning rate α) affect generalization and optimization. The generalization bound $O\left(\frac{T^\alpha \log T}{n}\right)$ reveals a sublinear power-law dependence on T with exponent α, better capturing the empirically observed growth trend than the linear bound $O\left(\frac{\alpha T k}{n}\right)$ of Tan et al. (2024). The optimization bound $O\left(\frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}}\right)$ shows that α plays the opposite role, theoretically explaining the trade-off between faster optimization and better generalization. The proposed analytical framework is extendable to assess the excess risk $O\left(\frac{\alpha T^\alpha \log T}{n} + \frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}}\right)$ of the SGD-based lookahead algorithm. This bound likewise reveals a trade-off between faster optimization and better generalization governed by α.

- We conduct numerical experiments on smooth, non-smooth, and non-convex problems to evaluate the behavior of SALA and validate our theoretical findings.

1.1 Related Works

Lookahead Optimizer. Lookahead, as a strategy that wraps base optimizers, has found widespread application across various fields of machine learning (El Shar & Jiang, 2020; Efroni et al., 2020; Protopapas & Barakat, 2024; Merlis, 2024; Merlis et al., 2024). Its robustness in handling complex tasks has spurred research into its theoretical foundations (Zhou et al., 2021; Li & Lei, 2025; Yang et al., 2024b). Zhou et al. (2021) provided the first stability-based generalization framework of lookahead and pair it with optimization behavior to explain why lookahead can outperform its inner optimizer in test accuracy. Li & Lei (2025) further analyzed how relaxing the gradient-bounded condition in convex settings influences the algorithm's generalization and optimization behavior. These two studies focused on lookahead with SGD as the base optimizer. More recently, several studies (Tan et al., 2024; Yu et al., 2024) have combined SAM with lookahead, showing exceptional performance. Tan et al. (2024) introduced the SALA algorithm, providing generalization and convergence guarantees in convex and gradient-bounded settings. However, hyperparameters (e.g., α and perturbation radius ρ) have not been thoroughly addressed in the existing literature. In particular, there is no theoretical support for SALA in non-convex settings although it has been employed widely for many real-world applications.

Algorithmic Stability and Generalization. Stability is a classical approach for proving generalization bounds, originating from the work of Rogers & Wagner (1978). A standard framework was introduced by Bousquet & Elisseeff (2002), who defined uniform stability for ERM and established generalization bounds for broad learning models. Hardt et al. (2016) extended these bounds to SGD, showing that early stopping enhances generalization. Lei & Ying (2020) introduced the concept of on-average model stability, a novel stability measure that relaxes the traditional assumption of bounded gradients in SGD. Subsequent studies have significantly advanced this framework by investigating algorithmic stability in the absence of the gradient boundedness assumption (Li et al., 2025; Schliserman & Koren, 2022; Li & Lei, 2025). Recently, generalization analyses for various stochastic optimization methods, such as SGD (Kuzborskij & Lampert, 2018; Lei et al., 2025), asynchronous SGD (Deng et al., 2025), and zeroth-order SGD (Nikolakakis et al., 2022; Chen et al., 2023; Liu et al., 2024), have been developed through stability-based techniques.

2 Preliminaries

This section introduces the background required for the subsequent analysis, including the algorithm definition, the statistical learning setup, and the assumptions underlying the theoretical results.

2.1 Optimization and Generalization in Algorithmic Analysis

Let $A(S)$ denote the model parameters returned by an algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ trained on a dataset S. The empirical risk $F_S(A(S))$ represents the average loss incurred by the model $A(S)$ over the finite training set S, and is defined as

$$F_S(A(S)) = \frac{1}{n}\sum_{i=1}^{n} f\left(A(S), z_i\right),$$

where $f\left(A(S), z_i\right)$ denotes the loss function evaluated for the i-th sample $z_i \in S$, and n is the number of samples in S. The population risk $F_{\mathcal{D}}(A(S))$, on the other hand, is the expected loss over the population distribution $\mathcal{D}$, and is given by

$$F_{\mathcal{D}}(A(S)) = \mathbb{E}_{z\sim\mathcal{D}}[f(A(S), z)],$$

where the expectation is taken with respect to the population distribution $\mathcal{D}$. The population optimal parameter and the empirical risk minimizer over $\mathcal{W}$ are defined as

$$\mathbf{w}^* \in \arg\min_{\mathbf{w}\in\mathcal{W}} F_{\mathcal{D}}(\mathbf{w}) \quad \text{and} \quad \mathbf{w}(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w}),$$

respectively. For notational convenience, the estimator produced by algorithm A on dataset S is denoted by $A(S)$.

The excess risk ϵ_{exc} of $A(S)$, given by $F_{\mathcal{D}}(A(S)) - F_{\mathcal{D}}(\mathbf{w}^*)$, can be upper bounded by decomposing it into a generalization gap ϵ_{gen} and an optimization error ϵ_{opt}. Specifically, taking expectations over the random draw of S and the internal randomness of A yields

$$\mathbb{E}_{S,A}[F_{\mathcal{D}}(A(S)) - F_{\mathcal{D}}(\mathbf{w}^*)] \leq \underbrace{\mathbb{E}_{S,A}[F_{\mathcal{D}}(A(S)) - F_S(A(S))]}_{\text{generalization gap}} + \underbrace{\mathbb{E}_{S,A}[F_S(A(S)) - F_S(\mathbf{w}(S))]}_{\text{optimization error}},$$

where $\mathbb{E}_{S,A}$ is abbreviated as $\mathbb{E}$ in the following main text.

Bounding the generalization gap is challenging due to the inaccessibility of the population distribution $\mathcal{D}$. Algorithmic stability has become a widely used and effective framework for addressing this difficulty, with successful applications in various learning scenarios, including adversarial training (Xiao et al., 2022; Cao et al., 2025), compositional optimization (Yang et al., 2024a; Chen et al., 2024), and neural network training (Deora et al., 2024; Richards & Kuzborskij, 2021; Taheri & Thrampoulidis, 2024; Wang et al., 2025).

The analysis in this work follows the stability-based approach. For completeness, the relevant definitions are recalled below.

Definition 2.1 (Lei & Ying (2020))**.** *Let $S = \{z_i\}_{i=1}^n$ and $\widetilde{S} = \{\tilde{z}_i\}_{i=1}^n$ be two datasets independently sampled from distribution $\mathcal{D}$. For any $i \in \{1, \ldots, n\}$, define*

$$S^{(i)} = \{z_1, \ldots, z_{i-1}, \tilde{z}_i, z_{i+1}, \ldots, z_n\}.$$

A randomized algorithm A is said to be on-average model ϵ-stable if

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \|A(S) - A(S^{(i)})\|_2^2\right] \leq \epsilon.$$

The following assumptions are commonly adopted to establish a formal connection between algorithmic stability and the expected generalization error $\mathbb{E}[F_{\mathcal{D}}(A(S)) - F_S(A(S))]$.

Assumption 2.2. *A differentiable loss function f is said to be L-smooth, if there exists a constant $L > 0$, for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\|\nabla f(\mathbf{w}, z) - \nabla f(\mathbf{w}', z)\| \leq L \|\mathbf{w} - \mathbf{w}'\|_2.$$

Assumption 2.2 is commonly employed in generalization analyses based on on-average model stability (Lei & Ying, 2020; Liu et al., 2024), where explicit dependence on a global Lipschitz constant can be avoided. Moreover, Assumption 2.2 is widely used in establishing convergence guarantees, including for the SALA algorithm (Tan et al., 2024) and closely related methods such as Lookahead (Li & Lei, 2025; Zhou et al., 2021) and SAM (Oikonomou & Loizou, 2025; Andriushchenko & Flammarion, 2022; Dai et al., 2023; Si & Yun, 2023).

Assumption 2.3. *A differentiable loss function f has (β, L)-Hölder continuous gradients if there exist constants $L > 0$ and $\beta \in [0, 1]$, for any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\|\nabla f(\mathbf{w}, z) - \nabla f(\mathbf{w}', z)\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2^{\beta}.$$

Assumption 2.3 generalizes the standard L-smoothness assumption: $\beta = 1$ recovers Lipschitz continuous gradients, whereas $\beta \in (0, 1)$ allows gradients that need not be Lipschitz continuous. This more general regularity condition has been widely studied in optimization and learning theory, including stability-based generalization analyses of SGD (Lei & Ying, 2020) and convergence analyses of universal gradient methods (Nesterov, 2015; Rodomanov et al., 2024) . Examples include the regression loss $|\langle \mathbf{w}, \mathbf{x}\rangle - y|^q$ and the classification loss $(1 - y\langle \mathbf{w}, \mathbf{x}\rangle)_+^q$, where $(a)_+ = \max\{a, 0\}$ and $y \in \{-1, 1\}$ for classification. For $q \in (1, 2)$ and bounded inputs, both have $(q-1)$-Hölder continuous gradients but are generally not L-smooth.

Lemma 2.4. *(Lei & Ying, 2020) Suppose the function f is nonnegative and smooth. If algorithm A is on-average model ϵ-stable, then for any $\gamma > 0$, $\epsilon_{gen} \leq \frac{L}{\gamma}\,\mathbb{E}[F_S(A(S))] + \frac{L+\gamma}{2}\,\epsilon$,*

and suppose Assumption 2.3 holds and the function f is convex and nonnegative. If algorithm A is on-average model ϵ-stable, then for any $\gamma > 0$, $\epsilon_{gen} \leq \frac{c_{\beta,1}^2}{2\gamma}\mathbb{E}\left[F_S(A(S))^{\frac{2\beta}{1+\beta}}\right] + \frac{\gamma}{2}\epsilon$,

where $c_{\beta,1} = (1 + 1/\beta)^{\frac{\beta}{1+\beta}} L^{\frac{1}{1+\beta}}$ if $\beta \in (0, 1]$ and $c_{\beta,1} = \sup_z \|\nabla f(\mathbf{0}, z)\|_2 + L$ if $\beta = 0$.

The subsequent assumptions are required by the stability and convergence analysis in Section 3.

Assumption 2.5. *We say that a differentiable loss function f satisfies the Polyak-Lojasiewicz condition if there exists a constant $\mu > 0$ for all $\mathbf{w} \in \mathbb{R}^d$,*

$$2\mu \left(F_S(\mathbf{w}) - F_S(\mathbf{w}(S))\right) \leq \|\nabla F_S(\mathbf{w})\|_2^2.$$

The Polyak-Lojasiewicz condition is a standard assumption in non-convex optimization (Zhou et al., 2021; Oikonomou & Loizou, 2025), relating the squared gradient norm $\|\nabla F_S(\mathbf{w})\|_2^2$ to the function value suboptimality $F_S(\mathbf{w}) - F_S(\mathbf{w}(S))$.

2.2 Sharpness-Aware Lookahead Algorithm

The Sharpness-Aware Lookahead method integrates Sharpness-Aware Minimization (SAM) with the lookahead optimization framework. In this approach, the inner-loop parameters are updated using SAM, which applies a perturbation to the gradient update in order to improve the model's robustness against sharp minima (Foret et al., 2021).

Let $\mathbf{w}_{t-1} \in \mathbb{R}^d$ represent the outer-loop parameters and $\{\eta_{\tau,t}\}_{\tau=1}^k$ be the sequence of step sizes for the k iterations within the inner loop. At each step τ, a sample $z_{i_{\tau,t}}$ is randomly and independently drawn from the training set $S = \{z_i\}_{i=1}^n$. Given the perturbation scale $\rho > 0$, SAM updates the inner-loop parameters $\boldsymbol{v}_{\tau,t}$ by adding a scaled gradient perturbation, which is then used to guide the update of $\boldsymbol{v}_{\tau+1,t}$:

$$\boldsymbol{v}_{\tau+1,t} = \boldsymbol{v}_{\tau,t} - \eta_{\tau+1,t} \nabla f \left(\boldsymbol{v}_{\tau,t} + \rho \frac{\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})}{\|\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\|_2}, z_{i_{\tau,t}} \right), \tag{1}$$

This update incorporates the sharpness-aware gradient, which modifies the inner-loop parameters to mitigate sharp minima, thereby encouraging a more stable optimization trajectory. After k inner-loop iterations, the final inner-loop parameter $\boldsymbol{v}_{k+1,t}$ is transferred to the outer-loop. The outer-loop then uses these parameters to update the outer-loop parameters through lookahead interpolation, controlled by the outer-loop step size $\alpha \in (0, 1)$:

$$\boldsymbol{w}_t = (1 - \alpha)\boldsymbol{w}_{t-1} + \alpha \boldsymbol{v}_{k+1,t}$$

Algorithm 1 Sharpness-Aware Lookahead Algorithm

Input: Training data $S = \{z_i\}_{i=1}^n$, initial model $\mathbf{w}_0 \in \mathbf{R}^d$, objective function $f(\mathbf{w}, z)$, perturbation radius ρ, inner-loop step number k, learning rates $\{\eta_{\tau,t}\}_{\tau=1,t=1}^{k,T}$, outer-loop step number T and learning rate α.

```
for t = 1, ..., T do
    v_{1,t} = w_{t-1}
    for τ = 1, ..., k do
        Sample i_{τ,t} independently and uniformly from {1, ..., n}.
        v̂_{τ,t} = v_{τ,t} + ρ ∇f(v_{τ,t}, z_{i_{τ,t}}) / ||∇f(v_{τ,t}, z_{i_{τ,t}})||_2
        v_{τ+1,t} = v_{τ,t} − η_{τ,t} ∇f(v̂_{τ,t}, z_{i_{τ,t}})
    end for
    w_t = (1 − α) w_{t-1} + α v_{k+1,t}
end for
```

Output: $\mathbf{w}_T$

Here, α dictates the influence of the inner-loop parameters on the outer-loop parameters. The lookahead step smooths the parameter updates by mixing the outer-loop and inner-loop weights, preventing overfitting to sharp minima while allowing faster convergence towards a generalizable solution. For the degenerate case where $\|\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\|_2 = 0$, we define the normalized perturbation direction as the zero vector, i.e., $\frac{\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})}{\|\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\|_2} = \mathbf{0}$. The details of SALA are provided in Algorithm 1.

3 Main Results

This section presents the theoretical guarantees of two lookahead algorithms under several cases. All proofs of the theorems and corollaries are provided in the appendix.

3.1 LEARNING GUARANTEES FOR SALA

Theorem 3.1. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0,1]$. Let $\{\mathbf{w}_t\}$ and $\left\{\mathbf{w}_t^{(i)}\right\}$ be generated by the SALA algorithm based on S and $S^{(i)}$, respectively. By setting $\eta_{\tau,t} \leq \frac{1}{4Lkt}$, then we have*

$$\epsilon_{gen} \leq \frac{e^3(L+\gamma)}{2} T^{\alpha} \left(\frac{(2\sqrt{L}+3L\rho)\rho(\log T+1)}{L} + \frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2\pi^2}{6nkL^2} \right) + \frac{L}{\gamma}\,\mathbb{E}[F_S(A(S))]\,.$$

Remark 3.2. *Existing generalization guarantees of SALA are mainly developed for convex and gradient-bounded loss function settings (Tan et al., 2024), leading to the bounds with the order $O\big(\frac{\alpha Tk}{n}\big)$. By contrast, Theorem 3.1 targets a weaker condition, i.e., non-convex and gradient-unbounded settings, providing a tighter characterization of generalization bound in terms of T. As shown in Theorem 3.1, the generalization bound derived from on-average model stability depends explicitly on the optimization error $F_S(A(S))$. During the initial phase of training, the generalization error typically decreases as optimization error improves, leading to a temporary narrowing of the generalization gap. However, as training proceeds, the generalization gap will increase, reflecting the emergence of overfitting. In general, we focus on the regimes where optimization has essentially converged. Therefore, we assume $\mathbb{E}\left[F_S(A(S))\right] = O\big(\frac{1}{n}\big)$ as previous theoretical works. Under this case, the first term becomes the dominant contributor to the bound. By setting $\rho = \frac{1}{n}$, the bound becomes $O\Big(\frac{T^{\alpha}\log T}{n}\Big)$. Compared with the prior linear dependence on Tk, this bound exhibits a milder power-law scaling with respect to the effective iteration count T, where the exponent is controlled by $\alpha \in (0,1)$. Empirical validation of the derived bound is provided in Figures 1 and 2, supporting its consistency with observed behavior. Detailed discussion is shown in Section 4.*

Theorem 3.3. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0,1]$. Let $\{\mathbf{w}_t\}$ be generated by the SALA algorithm based on S. By setting $\eta_{\tau,t} \leq \frac{1}{\mu\left(tk+\tau+1\right)}$ and $\rho_{\tau,t} \leq \frac{1}{\mu\left(tk+\tau+1\right)}$, then we have*

$$\epsilon_{opt} \leq \frac{4}{(Tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right) - F_S\left(\mathbf{w}(S)\right)\right] + \frac{2\alpha(1-\alpha)L^2(2+L^{\frac{1}{2}})(k+1)}{\mu^2(Tk+1)^{2\alpha-1}} + \frac{+2\alpha L^{\frac{3}{2}} + \alpha^2 L(2\sqrt{L}+L)^2}{\mu^2(Tk+1)^{2\alpha-1}}.$$

Remark 3.4. *Recent convergence analyses indicate that SAM can reach the neighborhood of an optimum, the radius of which is bounded by the perturbation level ρ (Oikonomou & Loizou, 2025; Si & Yun, 2023). Under gradient-bounded assumptions, previous theoretical work (Zhou et al., 2021) derived the convergence rate with the order $O\Big(\frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}}\Big)$ for the non-convex SGD-based lookahead algorithm. Our Theorem 3.3 can provide a comparable convergence rate under the weaker conditions, i.e., gradient-unbounded assumption and a perturbation radius decaying as $\rho_{k,t} \leq O\big(\frac{1}{tk}\big)$ (Oikonomou & Loizou, 2025). We acknowledge that our bound inherits the limitation of Zhou et al. (2021), wherein the term $O\Big(\frac{k}{(Tk+1)^{2\alpha-1}}\Big)$ is valid for $\alpha \in [0.5, 1]$. Moreover, the bound reveals that the outer-loop step size α plays a critical role in the convergence of the training objective. A detailed analysis of this relationship is presented in Section 4.*

Corollary 3.5. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is convex and $\nabla f(\mathbf{w}, z)$ is (β, L)-Hölder continuous with $\beta \in [0,1]$. Let $\{\mathbf{w}_t\}$ and $\left\{\mathbf{w}_t^{(i)}\right\}$ be generated by the SALA algorithm based on S and $S^{(i)}$, respectively. Then, we have*

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{w}_T - \mathbf{w}_T^{(i)}\right\|_2^2\right] \leq & \frac{e\alpha(1-\beta)}{1+\beta} L^{\frac{2}{1-\beta}} \sum_{t=1}^{T}\sum_{j=1}^{k} \eta_{j,t}^{\frac{2}{1-\beta}} + \frac{2^{2+\beta}e\alpha L}{1+\beta}\rho^{1+\beta}\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t} \\ & + \frac{8e\alpha}{n}\left(1+\frac{kT}{n}\right)\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}} + L^2\rho^{2\beta}\right].\end{aligned}$$

Remark 3.6. *The Hölder continuous extends the applicability of our stability analysis beyond L-smooth losses. Relaxing smoothness introduces an additional stability term proportional to $\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^{\frac{2}{1-\beta}}$, whose magnitude is controlled by the step sizes. Moreover, expressing the*

bound in terms of the empirical risks $F_S(\boldsymbol{v}_{j,t})$ allows us to avoid imposing a uniform bounded-gradient assumption. Let $\eta = c(kT)^{-\theta}$ with fixed $c > 0$, and suppose $kT \asymp n$, $\rho = O(n^{-1})$, $\mathbb{E}_S\|\mathbf{w}_1 - \mathbf{w}(S)\|_2^2 = O(1)$, and $\mathbb{E}_S[F_S(\mathbf{w}(S))] = O(1)$. For $\frac{1}{2} \le \beta < 1$, choosing $\theta = \frac{1}{2}$ yields the second-moment stability bound $O\left(\frac{\alpha}{n} + \alpha^{\frac{1-\beta}{1+\beta}} n^{-\frac{1+2\beta}{1+\beta}}\right)$. For $0 \le \beta < \frac{1}{2}$, choosing $\theta = 1 - \beta$ gives $O\left(\frac{\alpha}{n} + \alpha^{\frac{1-\beta}{1+\beta}} n^{-\frac{2}{1+\beta}}\right)$. In both regimes, a smaller α yields a tighter stability bound and, under uniform control of the relevant risk moments, a tighter generalization guarantee. Thus, the beneficial effect of a smaller α on the generalization bound persists in the non-smooth setting.

3.2 Byproduct: Learning Guarantees for SGD-based Lookahead Optimizer

We apply our theoretical analysis framework of SALA to the SGD-based Lookahead optimizer.

Theorem 3.7. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0, 1]$. Let $\{\mathbf{w}_t\}$ and $\left\{\mathbf{w}_t^{(i)}\right\}$ be generated by SGD-based lookahead optimizer based on S and $S^{(i)}$, respectively. By setting $\eta_{\tau,t} \le \frac{1}{2Lkt}$, then we have*

$$\epsilon_{gen} \le \frac{\pi^2 e^{2\alpha+\frac{5}{2}}\alpha(L+\gamma)}{3nkL}\left(1 + \frac{kT}{n}\right)T^{\alpha} + \frac{L}{\gamma}\mathbb{E}[F_S(A(S))]\,.$$

Remark 3.8. *Theorem 3.7 establishes the first generalization bound for SGD-based lookahead algorithm in non-convex settings without assuming gradient boundedness. By leveraging on-average model stability together with the almost co-coercivity property of smooth losses, the resulting guarantee scales as $O\left(\frac{\alpha T^{\alpha}}{n}\right)$, showing that α controls the power-law exponent with respect to the effective iteration count T.*

Theorem 3.9. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0, 1]$. Let $\{\mathbf{w}_t\}$ be generated by SGD-based lookahead on S. By setting $\eta_{\tau,t} \le \frac{1}{\mu\left(tk+\tau+1\right)}$, then we have $\epsilon_{opt} \le \frac{4}{(Tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right) - F_S\left(\mathbf{w}(S)\right)\right] + \frac{4\alpha(1-\alpha)L^2(k+1)+2\alpha L^2}{\mu^2(Tk+1)^{2\alpha-1}}$.*

Remark 3.10. *Under weaker assumptions, Theorem 3.9 establishes a convergence guarantee for SGD-based lookahead that matches the rate of Zhou et al. (2021), namely $O\left(\frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}}\right)$.*

4 Impact of Lookahead

This section provides a detailed analysis of the impact of the lookahead strategy on the generalization and optimization of SAM and SGD. Specifically, we analyze the theoretical results from Section 3 and complement them with the intuitive analysis of our experimental results to support our findings.

4.1 Empirical Setup

We evaluate the generalization and convergence performance of SALA under different outer-loop learning-rate parameter α settings. ResNet-18 (He et al., 2016) is trained on CIFAR-100 (Krizhevsky et al., 2009) as the primary benchmark. SALA is trained with $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$, where $\alpha = 1.0$ recovers the base SAM update without the outer-loop interpolation effect. The dataset is split into 80% for training and 20% for testing. All reported curves are averaged over three random seeds $\{0, 1, 2\}$, with shaded regions indicating one standard deviation. The model is trained for 300 epochs with a batch size of 64, an initial learning rate of 1×10^{-3}, and a multi-step scheduler with decay factor $\gamma = 0.1$ at 50% and 80% of total epochs. The perturbation radius is set to $\rho = 0.0005$ and inner-loop steps $k = 5$. Additional results on VGG-11/CIFAR-10, LSTM/AG_News, extended comparisons, and implementation details are provided in Appendix A.

4.2 Trade-off between Generalization and Optimization

Based on the decomposition of excess risk presented earlier, we have $\epsilon_{exc} \le \epsilon_{opt} + \epsilon_{gen}$.This decomposition highlights that to achieve a model that performs well on unseen data, both training

Table 1: A list of the generalization error bounds developed here in for lookahead algorithms. (T represents the total number of outer-loop iterations, k denotes the number of inner-loop iterations, α is the learning rate associated with the outer-loop, η is the learning rate associated with the inner-loop, and n indicates the total number of samples.)

Algorithm	Reference	Assumptions		Tool	Generalization Gap
		Convex	Lipschitz		
SALA	Tan et al. (2024)	✓	✓	Uniform Stability	$O(\frac{\alpha T k}{n})$
	Ours	×	×	On-Average Model Stability	$O(\frac{T^{\alpha}\log T}{n})$
SGD-based Lookahead	Zhou et al. (2021)	✓	✓	Uniform Stability	$O(\frac{\alpha T k}{n})$
	Zhou et al. (2021)	×	✓	Uniform Stability	$O(\frac{\alpha (Tk)^{\frac{\gamma}{1+\gamma}}}{n})$
	Li & Lei (2025)	✓	×	On-Average Model Stability	$O(\frac{\alpha^2\eta^2 T k}{n})$
	Ours	×	×	On-Average Model Stability	$O(\frac{\alpha T^{\alpha}}{n})$

and generalization must be effectively optimized. However, as established in previous literature, optimization and generalization often involve a trade-off in many aspects (Xiao et al., 2022; Li et al., 2025). In the following, we analyze the trade-off inherent in excess risk across various lookahead algorithms, using the theoretical results from this paper as well as the corresponding experimental observations to illustrate this phenomenon.

Theorem 4.1. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0,1]$. Let $\{\mathbf{w}_t\}$ be generated by the SALA algorithm based on S. By setting $\eta_{\tau,t} \leq \min\left\{\frac{1}{\mu\left(tk+\tau+1\right)}, \frac{1}{4Lkt}\right\}$ and $\rho_{\tau,t} \leq \min\left\{\frac{1}{n}, \frac{1}{\mu\left(tk+\tau+1\right)}\right\}$, then*

$$\epsilon_{exc} \leq O\left(\frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}} + \frac{T^{\alpha}\log T}{n}\right).$$

Remark 4.2. *The bound in Theorem 4.1 shows that α governs the trade-off between optimization and generalization. Balancing the dominant α-dependent terms yields the heuristic estimate $\tilde{\alpha} = \frac{\log\left(\frac{nk(Tk+1)}{\log T}\right)}{2\log(Tk+1)+\log T}$. Assuming $T > 1$, this estimate lies in the admissible interval $(\frac{1}{2}, 1)$ when $n \in \left(\frac{\sqrt{T}\log T}{k}, \frac{T(Tk+1)\log T}{k}\right)$. This estimate offers an additional insight into the α trade-off established above. Within this range, increasing n with T and k fixed suppresses the generalization term through the factor $\frac{1}{n}$, favoring a larger α to improve the optimization convergence rate. In contrast, increasing T with n and k fixed favors a smaller α to keep the growing generalization term in check. We emphasize that $\tilde{\alpha}$ should be interpreted as a bound-induced qualitative guideline rather than a precise prescription.*

Trade-off of iteration counts T, k. The first two terms of Theorem 4.1 shows that the model's fit to the training data improves progressively as the number of iterations increases. However, the third term, the generalization error, reveals a growing divergence between performance on unseen data and the training data as iterations proceed. This overfitting phenomenon is consistent with the findings in prior literature (Xiao et al., 2022; Rice et al., 2020), underscoring the importance of early stopping. Our experimental results similarly demonstrate a progressive increase of test error during the later stages of training (Figures 2 (c), (f) and Figure 3), further validating the necessity of early stopping strategy.

Trade-off of learning rates η, α. Given fixed model architecture, data, loss function, and number of iterations, the selection of hyperparameters emerges as the primary determinant of the model's ultimate optimization efficiency and generalization capability. Among these, learning rates are particularly critical, exerting significant impacts on both generalization behavior and convergence dynamics. For the lookahead module, the impacts of the inner-loop learning rate η on generalization and optimization are relatively consistent. Detailedly, employing a decaying η, as our theoretical

analysis, facilitates improved convergence and generalization. However, the outer-loop learning rate α directly shapes the derived theoretical bounds, exhibiting a distinct dual impact. Specifically, its value exhibits a positive correlation with generalization performance while displaying a negative correlation with optimization efficiency.

(1) Degradation in optimization for α. Theorem 4.1 presents an optimization bound with the order $O\left(\frac{1}{(Tk+1)^{2\alpha}} + \frac{k}{(Tk+1)^{2\alpha-1}}\right)$. Overall, this bound decreases as α increases, indicating faster optimization with a larger outer-loop learning rate. As previously noted, when $\alpha < 0.5$, the bound fails to guarantee theoretical convergence—a limitation shared by existing related works. Nonetheless, it also yields insights that align closely with our empirical findings, as illustrated in Figure 2 (a), (d). Firstly, it can be observed that convergence deteriorates as α decreases. Secondly, the corresponding convergence performance is particularly poor when α approaches 0. In contrast, α is close to 1, the theoretical bound captures the exponential convergence trend effectively.

(2) Advancement in generalization for α. Theorem 4.1 establishes a generalization bound with the order $O\left(\frac{T^{\alpha}\log T}{n}\right)$. Overall, this bound grows with α, since α controls the power-law exponent with respect to the effective iteration count Tk. In contrast, existing works (Tan et al., 2024; Zhou et al., 2021) derive the generalization guarantees with the linear dependence αTk and the dependence $(Tk)^{\frac{\gamma}{1+\gamma}}$, respectively. Empirical evidence from Figure 2 (b), (e) aligns closely with our theoretical findings. Specifically, we observe that: (1) the generalization gap grows exponentially with the number of iterations, albeit with an exponent less than 1; (2) generalization performance improves as α decreases; (3) the derivative of the generalization gap trend becomes less pronounced for smaller values of α, indicating a deceleration in the deterioration rate.

Thus, there exists a trade-off between optimization and generalization with respect to the choice of the outer-loop learning rate α. A smaller value of α leads to improved generalization, while a larger α allows for faster fitting of the training data. Observing the experimental results in Figure 2, the trend of the validation loss shows that when α is small, the initial decline is slower. However, due to the relatively lower generalization error, the model exhibits better performance on unseen data in the later stages of training.

Corollary 4.3. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0,1]$. Let $\{\mathbf{w}_t\}$ be generated by SGD-based lookahead algorithm based on S. By setting $\eta_{\tau,t} \leq \min\left\{\frac{1}{\mu\left(tk+\tau+1\right)}, \frac{1}{2Lkt}\right\}$, then*

$$\epsilon_{exc} \leq \frac{\pi^2 e^{2\alpha+\frac{5}{2}}\alpha(L+\gamma)(1+\frac{kT}{n})}{3nkL}T^{\alpha} + \frac{4}{(Tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right) - F_S\left(\mathbf{w}(S)\right)\right] + \frac{4\alpha(1-\alpha)L^2(k+1)+2\alpha L^2}{\mu^2(Tk+1)^{2\alpha-1}}.$$

Remark 4.4. *As indicated by Corollary 4.3, the SGD-based lookahead algorithm also exhibits a trade-off between optimization and generalization. Specifically, as summarized in Table 1, the generalization bound of this algorithm exhibits a linear positive correlation with the parameter α: decreasing α leads to a corresponding reduction in the generalization bound. Compared with prior bounds, our bound reflects a more pronounced dependence on α. Furthermore, prior studies have noted that smaller values of α can adversely affect the algorithm's convergence behavior. These theoretical insights are consistent with empirical observations (Figure 3), which further validates the existence of an α-dependent trade-off between optimization and generalization performance.*

5 Conclusion

This paper develops a novel theoretical analysis framework based on the almost co-coercivity property of smooth losses, which applies to SALA and extends to SGD-based lookahead. Based on this framework, we establish tighter generalization and optimization guarantees for SALA and SGD-based lookahead in non-convex settings. Theoretical and empirical evidence further elucidates the trade-off between optimization and generalization induced by lookahead and clarifies the role of the outer-loop learning rate α.

ACKNOWLEDGMENTS

We used AI tools to assist with writing, grammar correction, and checking mathematical derivations. The authors take full responsibility for the accuracy and integrity of the manuscript.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning (ICML)*, pp. 639–668, 2022.

Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7360–7371, 2022.

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

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning (ICML)*, pp. 1059–1071, 2021.

Chang Cao, Han Li, Yulong Wang, Rui Wu, and Hong Chen. Adversarial robust generalization of graph neural networks. In *International Conference on Machine Learning (ICML)*, 2025.

Jun Chen, Hong Chen, Bin Gu, and Hao Deng. Fine-grained theoretical analysis of federated zeroth-order optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 54496–54508, 2023.

Jun Chen, Hong Chen, and Bin Gu. How does black-box impact the learning guarantee of stochastic compositional optimization? *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 107745–107794, 2024.

Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022.

Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto, Frank Norbert Proske, Hans Kersting, and Aurelien Lucchi. An sde for modeling sam: Theory and insights. In *International Conference on Machine Learning (ICML)*, pp. 25209–25253, 2023.

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:67741–67770, 2023.

Xiaoge Deng, Li Shen, Shengwei Li, Tao Sun, Dongsheng Li, and Dacheng Tao. Towards understanding the generalizability of delayed stochastic gradient descent. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024.

Yonathan Efroni, Mohammad Ghavamzadeh, and Shie Mannor. Online planning with lookahead policies. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:14024–14033, 2020.

Ibrahim El Shar and Daniel Jiang. Lookahead-bounded q-learning. In *International Conference on Machine Learning (ICML)*, pp. 8665–8675, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*, 2020.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Hoki Kim, Jinseong Park, Yujin Choi, and Jaewook Lee. Stability analysis of sharpness-aware minimization. *arXiv preprint arXiv:2301.06308*, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, pp. 2815–2824, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

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

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 38557–38570, 2022.

Yunwen Lei, Tao Sun, and Mingrui Liu. Minibatch and local sgd: Algorithmic stability and linear speedup in generalization. *Applied and Computational Harmonic Analysis*, pp. 101795, 2025.

Kangcheng Li and Yunwen Lei. Generalization and optimization of sgd with lookahead. *arXiv preprint arXiv:2509.15776*, 2025.

Qinglun Li, Yingqi Liu, Miao Zhang, Xiaochun Cao, Quanjun Yin, and Li Shen. Unveiling the power of multiple gossip steps: A stability-based generalization analysis in decentralized training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5631–5640, 2024.

Xinyue Liu, Hualin Zhang, Bin Gu, and Hong Chen. General stability analysis for zeroth-order optimization algorithms. In *International Conference on Learning Representations (ICLR)*, 2024.

Yihong Luo, Yuhan Chen, Siya Qiu, Yiwei Wang, Chen Zhang, Yan Zhou, Xiaochun Cao, and Jing Tang. Fast graph sharpness-aware minimization for enhancing and accelerating few-shot node classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:132364–132387, 2024.

Shiva Maleki Varnosfaderani and Mohamad Forouzanfar. The role of ai in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering*, 11(4):337, 2024.

Ciro Mennella, Umberto Maniscalco, Giuseppe De Pietro, and Massimo Esposito. Ethical and regulatory challenges of ai technologies in healthcare: A narrative review. *Heliyon*, 10(4), 2024.

Nadav Merlis. Reinforcement learning with lookahead information. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:64523–64581, 2024.

Nadav Merlis, Dorian Baudry, and Vianney Perchet. The value of reward lookahead in reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:83627–83664, 2024.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

Konstantinos Nikolakakis, Farzin Haddadpour, Dionysis Kalogerias, and Amin Karbasi. Black-box generalization: Stability of zeroth-order learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:31525–31541, 2022.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *International Conference on Learning Representations (ICLR)*, 2025.

Kimon Protopapas and Anas Barakat. Policy mirror descent with lookahead. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:26443–26481, 2024.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8609–8621, 2021.

Anton Rodomanov, Ali Kavis, Yongtao Wu, Kimon Antonakopoulos, and Volkan Cevher. Universal gradient methods for stochastic convex optimization. *arXiv preprint arXiv:2402.03210*, 2024.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pp. 506–514, 1978.

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory (CoLT)*, pp. 3380–3394, 2022.

Dongkuk Si and Chulhee Yun. Practical sharpness-aware minimization cannot converge all the way to optima. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:26190–26228, 2023.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems (NeurIPS)*, 23, 2010.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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

Chengli Tan, Jiangshe Zhang, Junmin Liu, and Yihong Gong. Sharpness-aware lookahead for accelerating convergence and improving generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Chengli Tan, Jiangshe Zhang, Junmin Liu, Yicheng Wang, and Yunda Hao. Stabilizing sharpness-aware minimization through a simple renormalization strategy. *Journal of Machine Learning Research*, 26(68):1–35, 2025.

Puyu Wang, Yunwen Lei, Di Wang, Yiming Ying, and Ding-Xuan Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2):344–402, 2025.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:15446–15459, 2022.

Ming Yang, Xiyuan Wei, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic compositional gradient descent algorithms. In *International Conference on Machine Learning (ICML)*, pp. 56542–56593, 2024a.

Shangda Yang, Vitaly Zankin, Maximilian Balandat, Stefan Scherer, Kevin Carlberg, Neil Walton, and Kody JH Law. Accelerating look-ahead in bayesian optimization: Multilevel monte carlo is all you need. *arXiv preprint arXiv:2402.02111*, 2024b.

Runsheng Yu, Youzhi Zhang, and James Kwok. Improving sharpness-aware minimization by lookahead. In *International Conference on Machine Learning (ICML)*, 2024.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Pan Zhou, Hanshu Yan, Xiaotong Yuan, Jiashi Feng, and Shuicheng Yan. Towards understanding why lookahead generalizes better than sgd and beyond. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:27290–27304, 2021.

A Additional Experimental Details

A.1 Implementation Details

All experiments are implemented with Python 3.14, PyTorch 2.9.1, and CUDA 12.8, running on a Windows machine equipped with an NVIDIA GeForce RTX 4060 GPU and an Intel Core i5-13400F CPU @ 2.50GHz.

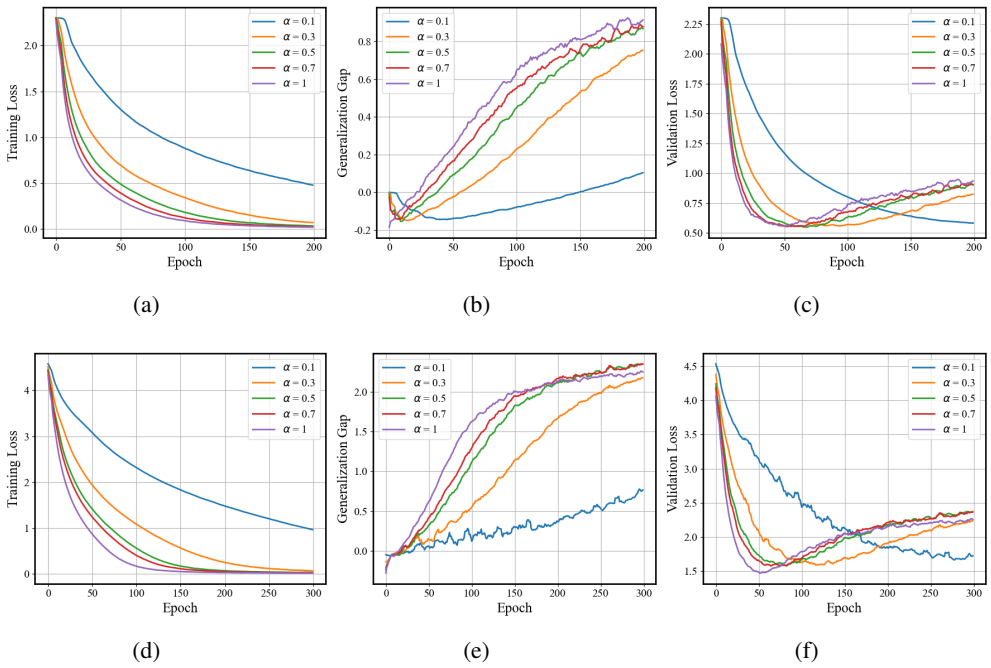


Figure 2: Results obtained with SALA under different outer-loop learning rates α. Subfigures (a)–(c) report the training, generalization gap, and validation loss of VGG-11 on CIFAR-10, while subfigures (d)–(f) show the corresponding results of ResNet-18 on CIFAR-100. The train loss is evaluated and recorded at the end of each epoch and the generalization gap is measured by the difference between the validation loss and the training loss.

A.2 Additional Results on VGG-11/CIFAR-10

To further validate our findings across different architectures and datasets, we train VGG-11 (Simonyan & Zisserman, 2014) on CIFAR-10 (Krizhevsky et al., 2009) under the same experimental protocol as the main text, with 200 training epochs. As shown in Figure 2, the results are consistent with those reported for ResNet-18/CIFAR-100, confirming that the trade-off between optimization and generalization with respect to α generalizes across architectures and datasets.

A.3 Additional Results on SGD-based Lookahead

To further validate the generalizability of our theoretical framework, we extend the experiments to the SGD-based Lookahead algorithm. ResNet-18 (He et al., 2016) is trained on CIFAR-100 (Krizhevsky et al., 2009) following the same experimental protocol as the main text: the dataset is split into 80% for training and 20% for testing, training runs for 300 epochs with a batch size of 64, an initial learning rate of 1×10^{-3}, and a multi-step scheduler with decay factor $\gamma = 0.1$ at 50% and 80% of total epochs. The inner-loop steps are set to $k = 5$. The algorithm is evaluated under $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$, where $\alpha = 1.0$ recovers the standard SGD update without the outer-loop interpolation effect.

As shown in Figure 3, the SGD-based Lookahead algorithm exhibits a trade-off between optimization and generalization consistent with our theoretical predictions in Corollary 4.3: a smaller α leads to a reduced generalization gap at the cost of slower convergence in training loss, while a larger α accelerates optimization but results in a larger generalization gap. These observations further confirm that the α-induced trade-off between optimization and generalization is not specific to SAM-based methods but is a fundamental property of the lookahead mechanism itself.

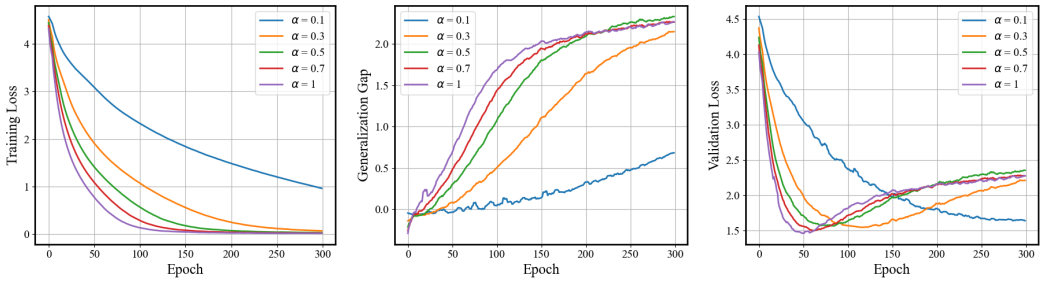


Figure 3: Results of ResNet-18 optimized with the SGD-based lookahead algorithm on the CIFAR-100 dataset for different outer-loop learning rates (α) and the generalization gap is measured by the difference between the validation loss and the training loss. The train loss is evaluated and recorded at the end of each epoch.

A.4 Additional Results on LSTM/AG_News

To assess the generalizability of SALA beyond image classification, we further conduct experiments on a text classification task. Specifically, an LSTM-based model is trained on the AG_News dataset, which consists of news articles categorized into four classes. The same experimental protocol as the main text is adopted: the dataset is split into 80% for training and 20% for testing, training runs for 300 epochs with a batch size of 64, an initial learning rate of 5×10^{-4}, and a multi-step scheduler with decay factor $\gamma = 0.2$ at 50% and 80% of total epochs. The perturbation radius is set to $\rho = 0.01$ and inner-loop steps $k = 5$. Performance is evaluated using Top-1 accuracy, macro-averaged F1 score, precision, and recall.

As shown in Figure 4, the results are consistent with our theoretical predictions and the image classification findings: a smaller α leads to a reduced generalization gap, while a larger α facilitates faster convergence in training loss. This confirms that the trade-off between optimization and generalization induced by the outer-loop learning rate α is not specific to vision tasks but holds more broadly across different model architectures and data modalities.

A.5 Additional Results under Convex Hölder-Continuous Gradients

To investigate the behavior of SALA in a convex setting with Hölder-continuous gradients, we conduct additional experiments using a linear classifier on the MNIST dataset. We use the standard split of 60,000 training samples and 10,000 test samples. Each 28×28 grayscale image is scaled to $[0, 1]$ and flattened into a 784-dimensional vector, which is mapped to ten class scores by a single linear layer. No data augmentation is applied. We adopt the multiclass q-hinge loss

$$f(\mathbf{w}, (\mathbf{x}, y)) = \sum_{c \neq y} [1 + s_b(\mathbf{w}; \mathbf{x}) - s_y(\mathbf{w}; \mathbf{x})]_+^{1+\beta},$$

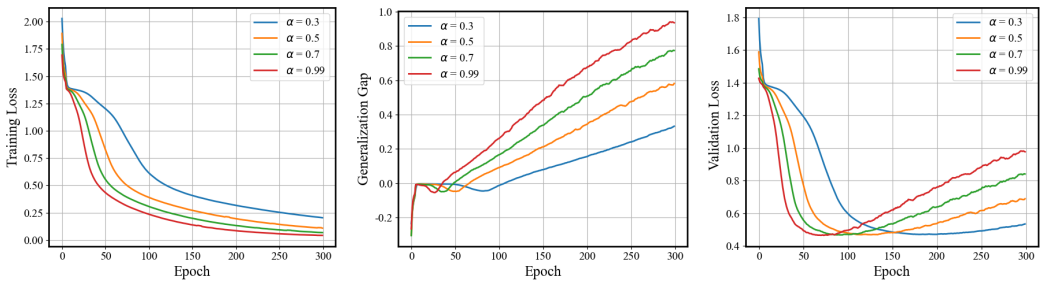


Figure 4: Results of LSTM-based optimized with the SALA algorithm on the AG_News dataset for different outer-loop learning rates (α), and the generalization gap is measured by the difference between the validation loss and the training loss.

where $s_b(\mathbf{w}; \mathbf{x})$ denotes the score for class b and $[a]_+ = \max\{a, 0\}$. We set $\beta = 0.5$, corresponding to $q = 1.5$. With bounded inputs and a linear classifier, this loss is convex in the model parameters and has β-Hölder-continuous gradients, while generally not satisfying the standard L-smoothness condition.

Training runs for 300 epochs with a batch size of 64 and an initial learning rate of 10^{-3}. A multi-step scheduler reduces the learning rate by a factor of $\gamma = 0.1$ at 30%, 60%, and 80% of the total training epochs. The perturbation radius is set to $\rho = 5 \times 10^{-4}$, and the number of inner-loop steps is fixed at $k = 5$. After every k SAM updates, the slow parameters are updated through lookahead interpolation with coefficient α. For reproducibility, we conduct three runs for each $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ using the same set of fixed random seeds $\{0, 1, 2\}$. Solid lines represent the mean across the three runs, while shaded regions indicate $\pm$ one standard deviation around the mean at each epoch.

At the end of each epoch, training and test losses are evaluated using the same slow-parameter model. The empirical generalization gap is defined as the test loss minus the training loss. Comparing these quantities across different values of α allows us to examine the trade-off between optimization and generalization in a convex setting beyond the standard smoothness assumption.

As shown in Figure 5, the results reveal a clear optimization and generalization trade-off controlled by the outer-loop learning rate α. Larger values of α enable faster optimization and yield lower training losses within the same training budget. However, larger α also yields a larger empirical generalization gap, qualitatively consistent with the linear dependence of our stability bound on α at a fixed training horizon. At a fixed training horizon, smaller α favors generalization-gap control, while larger α facilitates optimization. Intermediate values, such as $\alpha = 0.3$ and $\alpha = 0.5$, balance these effects and achieve lower validation losses.

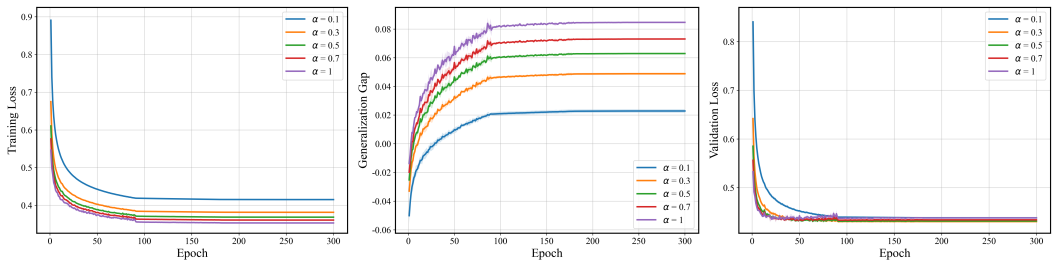


Figure 5: Results of a linear classifier trained with SALA on MNIST using the convex q-hinge loss with $q = 1.5$, for different outer-loop learning rates α. The generalization gap is measured by the difference between the validation loss and the training loss

B PROOFS OF THE MAIN RESULTS

B.1 LEMMA AND DEFINITION

Lemma B.1 (Richards & Kuzborskij (2021); Lei et al. (2022)). *Consider the gradient-based optimization method with update rule $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t, z_t)$. For two iteration sequences $\{\mathbf{w}_t\}_{t\in[T]}$ and $\{\mathbf{w}'_t\}_{t\in[T]}$, assume that $f(\mathbf{w}_t, z_t)$ is L-smooth, the step size satisfies $\eta_t \leq 1/(2L)$, and the minimum eigenvalue condition $\lambda_{\min}\big(\nabla^2 f(\mathbf{w}_t, z_t)\big) \geq -\epsilon$ holds. Then, for all $\mathbf{w}_t, \mathbf{w}'_t \in \mathbb{R}^d$ and $z_t \in \mathcal{Z}$, the following inequality is valid:*

$$\begin{aligned}&\langle \mathbf{w}_t - \mathbf{w}'_t,\, \nabla f(\mathbf{w}_t, z_t) - \nabla f(\mathbf{w}'_t, z_t)\rangle \\ &\geq 2\eta_t\left(1 - \frac{\eta_t L}{2}\right)\|\nabla f(\mathbf{w}_t, z_t) - \nabla f(\mathbf{w}'_t, z_t)\|_2^2 \; - \epsilon\,\|\mathbf{w}_t - \mathbf{w}'_t - \eta_t \nabla f(\mathbf{w}_t, z_t) + \eta_t \nabla f(\mathbf{w}'_t, z_t)\|_2^2 .\end{aligned}$$

Lemma B.2 (Srebro et al. (2010)). *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is nonnegative and L-smooth. Then we have*

$$\|\nabla f(\mathbf{w}, z)\|_2 \leq 2\sqrt{L f(\mathbf{w}, z)}.$$

Lemma B.3. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is nonnegative and L-smooth. For $\hat{\mathbf{w}} = \mathbf{w} + \rho\frac{\nabla f(\mathbf{w},z)}{\|\nabla f(\mathbf{w},z)\|_2}$, we have*

$$\|\nabla f(\hat{\mathbf{w}}, z)\|_2 \leq \sqrt{4L f(\mathbf{w}, z)} + L\rho.$$

Proof.

$$\begin{aligned}&\|\nabla f(\hat{\mathbf{w}}, z)\|_2 \\ \leq&\|\nabla f(\mathbf{w}, z)\|_2 + \|\nabla f(\mathbf{w} + \rho\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2}, z) - \nabla f(\mathbf{w}, z)\|_2 \\ \leq&\|\nabla f(\mathbf{w}, z)\|_2 + L\|\mathbf{w} + \rho\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \mathbf{w}\|_2 \\ \leq&\sqrt{4L f(\mathbf{w}, z)} + \rho L.\end{aligned}$$

□

Lemma B.4. *Assume that, for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0, 1]$. Let $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ be generated by SGD. Then, for $\eta < \frac{1}{2L}$, we have*

$$\|\mathbf{w}' - \eta\nabla f(\mathbf{w}', z) - \big(\mathbf{w} - \eta\nabla f(\mathbf{w}, z)\big)\|_2^2 \leq \frac{1}{1 - 2\eta L}\|\mathbf{w}' - \mathbf{w}\|_2^2.$$

Proof.

$$\begin{aligned}&\|\mathbf{w}' - \eta\nabla f(\mathbf{w}', z) - \big(\mathbf{w} - \eta\nabla f(\mathbf{w}, z)\big)\|_2^2 \\ =&\|\mathbf{w}' - \mathbf{w}\|_2^2 + \eta^2\|\nabla f(\mathbf{w}', z) - \nabla f(\mathbf{w}, z)\|_2^2 - 2\eta\langle \mathbf{w}' - \mathbf{w}, \nabla f(\mathbf{w}', z) - \nabla f(\mathbf{w}, z)\rangle.\end{aligned}$$

Since $\left\|\nabla^2 f\left(\mathbf{w}, z\right)\right\|$ equals to the largest singular value of $\nabla^2 f\left(\mathbf{w}, z\right)$, we have $\lambda_{\min}\left(\nabla^2 f\left(\mathbf{w}, z\right)\right) \geq -\left\|\nabla^2 f\left(\mathbf{w}, z\right)\right\| \geq -L$. Thus, by Lemma B.1, it follows that

$$\begin{aligned}&\|\mathbf{w}' - \eta\nabla f(\mathbf{w}', z) - \big(\mathbf{w} - \eta\nabla f(\mathbf{w}, z)\big)\|_2^2 \\ \leq&\|\mathbf{w}' - \mathbf{w}\|_2^2 + \eta^2\|\nabla f(\mathbf{w}', z) - \nabla f(\mathbf{w}, z)\|_2^2 - 4\eta^2(1 - \frac{\eta L}{2})\|\nabla f(\mathbf{w}', z) - \nabla f(\mathbf{w}, z)\|_2^2 \\ &+ 2\eta L\|\mathbf{w}' - \eta\nabla f(\mathbf{w}', z) - \big(\mathbf{w} - \eta\nabla f(\mathbf{w}, z)\big)\|_2^2 \\ =&\|\mathbf{w}' - \mathbf{w}\|_2^2 + (2\eta^3 L - 3\eta^2)\|\nabla f(\mathbf{w}', z) - \nabla f(\mathbf{w}, z)\|_2^2 \\ &+ 2\eta L\|\mathbf{w}' - \eta\nabla f(\mathbf{w}', z) - \big(\mathbf{w} - \eta\nabla f(\mathbf{w}, z)\big)\|_2^2,\end{aligned}$$

then, by setting $\eta < \frac{1}{2L}$, we have

$$\|\mathbf{w}' - \eta\nabla f(\mathbf{w}', z) - \big(\mathbf{w} - \eta\nabla f(\mathbf{w}, z)\big)\|_2^2 \leq \frac{1}{1 - 2\eta L}\|\mathbf{w}' - \mathbf{w}\|_2^2.$$

□

Lemma B.5. *Assume that, for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0, 1]$. Let $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ be generated by SAM. Then, for $\eta < \frac{1}{4L}$, we have*

$$\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 \leq \frac{1}{1-4\eta L}\|\mathbf{w}' - \mathbf{w}\|_2^2 + \frac{(8\sqrt{L} + 12L\rho)\rho\eta}{1-4\eta L}.$$

Proof. Since $\left\|\nabla^2 f(\mathbf{w}, z)\right\|$ equals to the largest singular value of $\nabla^2 f(\mathbf{w}, z)$, we have $\lambda_{\min}\left(\nabla^2 f(\mathbf{w}, z)\right) \geq -\left\|\nabla^2 f(\mathbf{w}, z)\right\| \geq -L$. Thus, by Lemma B.1, it follows that

$$\begin{aligned}
&\langle \hat{\mathbf{w}} - \hat{\mathbf{w}}', \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\rangle \\
\geq& 2\eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 - L\|\hat{\mathbf{w}} - \hat{\mathbf{w}}' - \eta\nabla f(\hat{\mathbf{w}}, z) + \eta\nabla f(\hat{\mathbf{w}}', z)\|_2^2 \\
\geq& 2\eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 - 2L\|\mathbf{w} - \mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}, z) + \eta\nabla f(\hat{\mathbf{w}}', z)\|_2^2 \\
&- 2L\rho^2\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}\|_2^2.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\langle \mathbf{w} - \mathbf{w}', \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\rangle \\
\geq& 2\eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 - 2L\|\mathbf{w} - \mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}, z) + \eta\nabla f(\hat{\mathbf{w}}', z)\|_2^2 \\
&- 2L\rho^2\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}\|_2^2 \\
&- \rho\left\langle \frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}, \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\right\rangle.
\end{aligned}$$

Substituting the above inequality yields

$$\begin{aligned}
&\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 \\
\leq& \|\mathbf{w} - \mathbf{w}'\|_2^2 + \eta^2\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|^2 + 4\eta L\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 \\
&+ 2\rho\eta\left\langle \frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}, \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\right\rangle \\
&- 4\eta^2\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 + 4L\eta\rho^2\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|}\|_2 \\
\leq& \|\mathbf{w} - \mathbf{w}'\|_2^2 + \eta^2\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|^2 + 4\eta L\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 \\
&+ 2\rho\eta\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}\|_2 \cdot \|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2 \\
&- 4\eta^2\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 + 4L\eta\rho^2\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|}\|_2 \\
\leq& \|\mathbf{w} - \mathbf{w}'\|_2^2 + 4\eta L\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 \\
&+ 2\rho\eta\left(\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2}\|_2 + \|\frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}\|_2\right) \cdot (\|\nabla f(\hat{\mathbf{w}}, z)\|_2 + \|\nabla f(\hat{\mathbf{w}}', z)\|_2) \\
&+ \left(2L\eta^3 - 3\eta^2\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 + 4L\eta\rho^2\|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|}\|_2 \\
\leq& \|\mathbf{w} - \mathbf{w}'\|_2^2 + 4\eta L\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 \\
&+ \left(2L\eta^3 - 3\eta^2\right)\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\|_2^2 + \big(8\sqrt{L} + 12L\rho\big)\rho\eta \\
\leq& \|\mathbf{w} - \mathbf{w}'\|_2^2 + 4\eta L\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\|_2^2 + (8\sqrt{L} + 12L\rho)\rho\eta.
\end{aligned}$$

The last inequality holds since we assume $\eta < \frac{1}{4L}$. The above inequality implies

$$\left\|\mathbf{w}' - \eta\nabla f(\hat{\mathbf{w}}', z) - \big(\mathbf{w} - \eta\nabla f(\hat{\mathbf{w}}, z)\big)\right\|_2^2 \leq \frac{1}{1-4\eta L}\left\|\mathbf{w} - \mathbf{w}'\right\|_2^2 + \frac{(8\sqrt{L} + 12L\rho)\rho\eta}{1-4\eta L}.$$

□

Assumption B.6. *We say that a differentiable loss function f is convex if for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$f(\mathbf{w}', z) \geq f(\mathbf{w}, z) + \langle \nabla f(\mathbf{w}, z), \mathbf{w}' - \mathbf{w} \rangle .$$

Lemma B.7 ((Lei & Ying, 2020))**.** *Suppose Assumption 2.3 holds and $f(\cdot, z)$ is convex for every $z \in \mathcal{Z}$. Then, for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$ and $z \in \mathcal{Z}$,*

$$\langle \mathbf{w} - \mathbf{w}', \nabla f(\mathbf{w}, z) - \nabla f(\mathbf{w}', z) \rangle \geq \frac{2\beta L^{-1/\beta}}{1+\beta} \left\| \nabla f(\mathbf{w}, z) - \nabla f(\mathbf{w}', z) \right\|_2^{\frac{1+\beta}{\beta}} . \tag{2}$$

Lemma B.8. *Assume that, for every $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is convex and differentiable, and its gradient is (β, L)-Hölder continuous with $\beta \in [0, 1]$. For any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, $\rho \geq 0$ and $\eta > 0$, we have*

$$\left\| \mathbf{w} - \eta \nabla f(\hat{\mathbf{w}}, z) - \mathbf{w}' + \eta \nabla f(\hat{\mathbf{w}}', z) \right\|_2^2 \leq \left\| \mathbf{w} - \mathbf{w}' \right\|_2^2 + \frac{1-\beta}{1+\beta} L^{\frac{2}{1-\beta}} \eta^{\frac{2}{1-\beta}} + \frac{2^{2+\beta}}{1+\beta} L \eta \rho^{1+\beta}. \tag{3}$$

Proof. By the Lemma B.7 and the definitions of $\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}'$, we have

$$\begin{aligned} &\langle \mathbf{w} - \mathbf{w}', \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \rangle \\ &+ \rho \left\langle \frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2}, \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\rangle \\ \geq &\frac{2\beta L^{-1/\beta}}{1+\beta} \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^{\frac{1+\beta}{\beta}} . \end{aligned} \tag{4}$$

Since

$$\left\| \frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2} - \frac{\nabla f(\mathbf{w}', z)}{\|\nabla f(\mathbf{w}', z)\|_2} \right\|_2 \leq 2,$$

expanding the squared distance and applying equation 4 together with the Cauchy–Schwarz inequality yield

$$\begin{aligned} &\left\| \mathbf{w} - \eta \nabla f(\hat{\mathbf{w}}, z) - \mathbf{w}' + \eta \nabla f(\hat{\mathbf{w}}', z) \right\|_2^2 \\ &= \left\| \mathbf{w} - \mathbf{w}' \right\|_2^2 + \eta^2 \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^2 \\ &\quad - 2\eta \left\langle \mathbf{w} - \mathbf{w}', \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\rangle \\ \leq &\left\| \mathbf{w} - \mathbf{w}' \right\|_2^2 + \eta^2 \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^2 \\ &\quad - \frac{4\beta\eta L^{-1/\beta}}{1+\beta} \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^{\frac{1+\beta}{\beta}} \\ &\quad + 4\eta\rho \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2 . \end{aligned} \tag{5}$$

We use half of the negative term in equation 5 to absorb the quadratic term. By Young's inequality with conjugate exponents $\frac{1+\beta}{2\beta}$ and $\frac{1+\beta}{1-\beta}$,

$$\begin{aligned} &\eta^2 \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^2 \\ &= \left(\eta L^{-1/\beta} \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^{\frac{1+\beta}{\beta}} \right)^{\frac{2\beta}{1+\beta}} (\eta L)^{\frac{2}{1+\beta}} \\ &\leq \frac{2\beta\eta L^{-1/\beta}}{1+\beta} \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2^{\frac{1+\beta}{\beta}} + \frac{1-\beta}{1+\beta} (\eta L)^{\frac{2}{1-\beta}}. \end{aligned} \tag{6}$$

The remaining half absorbs the perturbation term. Applying Young's inequality with conjugate exponents $\frac{1+\beta}{\beta}$ and $1+\beta$, we obtain

$$4\eta\rho \left\| \nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z) \right\|_2$$

$$
\begin{aligned}
&= \left(2\eta L^{-1/\beta} \left\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\right\|_2^{\frac{1+\beta}{\beta}}\right)^{\frac{\beta}{1+\beta}} \left(2^{2+\beta}\eta L \rho^{1+\beta}\right)^{\frac{1}{1+\beta}} \\
&\leq \frac{2\beta\eta L^{-1/\beta}}{1+\beta} \left\|\nabla f(\hat{\mathbf{w}}, z) - \nabla f(\hat{\mathbf{w}}', z)\right\|_2^{\frac{1+\beta}{\beta}} + \frac{2^{2+\beta}}{1+\beta} L\eta\rho^{1+\beta}. \qquad (7)
\end{aligned}
$$

Substituting equation 6 and equation 7 into equation 5 cancels all terms involving the gradient difference and gives equation 3. The proof is complete. □

Lemma B.9 (Lei & Ying (2020))**.** *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is nonnegative and convex and $\nabla f(\mathbf{w}, z)$ is (β, L)-Hölder continuous with $\beta \in [0, 1]$. Then we have*

$$
\|\nabla f(\mathbf{w}, z)\|_2 \leq c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\mathbf{w}, z), \quad \forall \mathbf{w} \in \mathbb{R}^d, z \in \mathcal{Z},
$$

where $c_{\beta,1} = \begin{cases} (1+1/\beta)^{\frac{\beta}{1+\beta}} L^{\frac{1}{1+\beta}}, & \text{if } \beta > 0 \\ \sup_z \|\nabla f(\mathbf{0}, z)\|_2 + L, & \text{if } \beta = 0 \end{cases}$.

Lemma B.10. *Assume for all $z \in \mathcal{Z}$, $f(\mathbf{w}, z)$ is nonnegative and convex and $\nabla f(\mathbf{w}, z)$ is (β, L)-Hölder continuous with $\beta \in [0, 1]$. Define $\hat{\mathbf{w}} = \mathbf{w} + \rho\nabla f(\mathbf{w}, z)/\|\nabla f(\mathbf{w}, z)\|_2$. Then we have*

$$
\|\nabla f(\hat{\mathbf{w}}, z)\|_2 \leq c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\mathbf{w}, z) + \rho^\beta L.
$$

Proof. Consider Lemma B.9

$$
\begin{aligned}
&\|\nabla f(\mathbf{w} + \rho\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2}, z)\|_2 \\
\leq &\|\nabla f(\mathbf{w}, z)\|_2 + \|\nabla f(\mathbf{w} + \rho\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2}, z) - \nabla f(\mathbf{w}, z)\|_2 \\
\leq &c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\mathbf{w}, z) + \rho^\beta L \|\frac{\nabla f(\mathbf{w}, z)}{\|\nabla f(\mathbf{w}, z)\|_2}\|_2^\beta \\
\leq &c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\mathbf{w}, z) + \rho^\beta L,
\end{aligned}
$$

Here if we set $\beta = 1$, $\|\nabla f(\hat{\mathbf{w}}, z)\|_2 \leq \sqrt{2L} f(\mathbf{w}, z)^{\frac{1}{2}} + \rho L$. □

B.2 Proof of Theorem 3.1

Theorem B.11. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0, 1]$. Let $\{\mathbf{w}_t\}$ and $\left\{\mathbf{w}_t^{(i)}\right\}$ be generated by the SALA algorithm based on S and $S^{(i)}$, respectively. Setting $\eta_{\tau,t} \leq \frac{1}{4Lkt}$, then we have*

$$
\mathbb{E}\left[\left\|\mathbf{w}_T - \mathbf{w}_T^{(i)}\right\|_2^2\right] \leq T^\alpha e^3 \left(\frac{(2\sqrt{L} + 3L\rho)\rho(\log T + 1)}{L} + \frac{(1 + \frac{kT}{n})(2\sqrt{L} + \rho L)^2\pi^2}{6nkL^2}\right).
$$

Proof. Firstly, we formulate the lookahead update as

$$
\begin{aligned}
\mathbf{w}_t &= (1-\alpha)\mathbf{w}_{t-1} + \alpha \boldsymbol{v}_{k+1,t} \\
&= (1-\alpha)\mathbf{w}_{t-1} + \alpha\left(\boldsymbol{v}_{k,t} - \eta_{k,t}\nabla f\left(\hat{\boldsymbol{v}}_{k,t}, z_{i_{k,t}}\right)\right), \\
\mathbf{w}_t^{(i)} &= (1-\alpha)\mathbf{w}_{t-1}^{(i)} + \alpha\left(\boldsymbol{v}_{k,t}^{(i)} - \eta_{k,t}\nabla f\left(\hat{\boldsymbol{v}}_{k,t}^{(i)}, z_{i_{k,t}}^{(i)}\right)\right).
\end{aligned}
$$

It is important to note that in the proof, the superscript (i) on the inner and outer-loop parameters indicates that they are derived from the dataset $S^{(i)}$, where $S^{(i)}$ is the dataset obtained by replacing the original dataset S with a new data point $\tilde{z}_i$.

Then according to lemma B.5, we have

$$
\left\|\boldsymbol{v}_{\tau+1,t} - \boldsymbol{v}_{\tau+1,t}^{(i)}\right\|_2^2 \leq \frac{1}{1 - 4\eta_{\tau,t}L}\left\|\boldsymbol{v}_{\tau,t} - \boldsymbol{v}_{\tau,t}^{(i)}\right\|_2^2 + \frac{(8\sqrt{L} + 12L\rho)\rho\eta_{\tau,t}}{1 - 4\eta_{\tau,t}L}.
$$

Updating the parameters with different data z_i, the above relationship can be reshaped as follow

$$\begin{aligned}\left\|\boldsymbol{v}_{\tau+1,t}-\boldsymbol{v}_{\tau+1,t}^{(i)}\right\|_2 &\leq \left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+\eta_{\tau,t}\left\|\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_i)-\nabla f(\hat{\boldsymbol{v}}_{\tau,t}^{(i)},\tilde{z}_i)\right\|_2\\ &\leq \left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+\eta_{\tau,t}\big(\|\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_i)\|_2+\|\nabla f(\hat{\boldsymbol{v}}_{\tau,t}^{(i)},\tilde{z}_i)\|_2\big)\\ &\leq \left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+2\eta_{\tau,t}(2\sqrt{L}+\rho L).\end{aligned}$$

The last inequality holds due to the Lemma B.3. Throughout the following derivation, $\mathbb{E}$ denotes expectation over the datasets and the shared sampling randomness. At each inner-loop step, conditional on the datasets and the preceding sampling history, the shared sample index equals i with probability $\frac{1}{n}$. Combining the above two situations and applying the law of total expectation, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{v}_{k+1,t}-\boldsymbol{v}_{k+1,t}^{(i)}\right\|^2\right]\\ \leq&(\frac{1+p}{n}+(1-\frac{1}{n})\frac{1}{1-4\eta_{k,t}L})\mathbb{E}\left[\left\|\boldsymbol{v}_{k,t}-\boldsymbol{v}_{k,t}^{(i)}\right\|_2^2\right]\\ +&\frac{(8\sqrt{L}+12L\rho)\rho\eta_{k,t}}{1-4\eta_{k,t}L}+\frac{4\eta_{k,t}^2(1+p^{-1})(2\sqrt{L}+\rho L)^2}{n}\\ \leq&\left(\prod_{j=1}^{k}(\frac{p+1}{n}+(1-\frac{1}{n})\frac{1}{1-4L\eta_{j,t}})\right)\mathbb{E}\left[\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2\right]\\ &+\sum_{j=1}^{k}\left(\prod_{j'=j+1}^{k}(\frac{p+1}{n}+(1-\frac{1}{n})\frac{1}{1-4L\eta_{j',t}})\right)\left(\frac{(8\sqrt{L}+12L\rho)\rho\eta_{j,t}}{1-4\eta_{j,t}L}+\frac{4(1+p^{-1})\eta_{j,t}^2(2\sqrt{L}+\rho L)^2}{n}\right)\\ \leq&\prod_{j=1}^{k}\left(\frac{1}{kT}+\frac{1}{n}+(1-\frac{1}{n})\frac{1}{1-4L\eta_{j,t}}\right)\mathbb{E}\left[\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2\right]\\ &+\sum_{j=1}^{k}\left(\prod_{j'=j+1}^{k}(\frac{1}{kT}+\frac{1}{n}+(1-\frac{1}{n})\frac{1}{1-4L\eta_{j',t}})\right)\left(\frac{(8\sqrt{L}+12L\rho)\rho\eta_{j,t}}{1-4\eta_{j,t}L}+\frac{4(1+\frac{kT}{n})\eta_{j,t}^2(2\sqrt{L}+\rho L)^2}{n}\right)\\ \leq&\left((1+\frac{1}{kT})^k(1+\frac{1}{kt-1})^k\right)\mathbb{E}\left[\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2\right]\\ &+\left((1+\frac{1}{kT})^k(1+\frac{1}{kt-1})^k\right)\left(\frac{(2\sqrt{L}+3L\rho)\rho k}{L(kt-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2k}{nk^2t^2L^2}\right)\\ \leq&\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\mathbb{E}\left[\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2\right]+\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\left(\frac{(2\sqrt{L}+3L\rho)\rho}{L(t-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2}{nkt^2L^2}\right),\end{aligned}$$

where the third inequality holds due to the setting of $p=\frac{n}{kT}$. And by setting $\eta_{j,t}\leq\frac{1}{4Lkt}$, we have the last inequality.

Since $\|\cdot\|_2^2$ is convex and the Lookahead update is a convex combination, taking expectations gives

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\right]=&\mathbb{E}\left[\|(1-\alpha)\big(\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\big)+\alpha\big(\boldsymbol{v}_{k+1,t}-\boldsymbol{v}_{k+1,t}^{(i)}\big)\|_2^2\right]\\ \leq&(1-\alpha)\mathbb{E}\left[\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2\right]+\alpha\mathbb{E}\left[\|\boldsymbol{v}_{k+1,t}-\boldsymbol{v}_{k+1,t}^{(i)}\|_2^2\right].\end{aligned}$$

Assume that $\mathbf{w}_1=\mathbf{w}_1^{(i)}$. For $T\geq 2$, iterating the above recursion from $t=2$ to T, we obtain

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\|_2^2\right]\\ \leq&\left(1-\alpha+\alpha\exp(\frac{1}{T}+\frac{1}{T-1})\right)\mathbb{E}\left[\|\mathbf{w}_{T-1}-\mathbf{w}_{T-1}^{(i)}\|_2^2\right]\end{aligned}$$

$$
\begin{aligned}
&+\alpha\exp\left(\frac{1}{T}+\frac{1}{T-1}\right)\left(\frac{(2\sqrt{L}+3L\rho)\rho}{L(T-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2}{nkT^2L^2}\right)\\
\leq&\prod_{t=2}^{T}\left(1-\alpha+\alpha\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\right)\sum_{t=2}^{T}\left(\frac{(2\sqrt{L}+3L\rho)\rho}{L(t-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2}{nkt^2L^2}\right)\\
=&\prod_{t=2}^{T}\exp\left(\log\left(1-\alpha+\alpha\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\right)\right)\sum_{t=2}^{T}\left(\frac{(2\sqrt{L}+3L\rho)\rho}{L(t-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2}{nkt^2L^2}\right)\\
\leq&\prod_{t=2}^{T}\exp\left(\alpha\left(\frac{1}{T}+\frac{1}{t-1}\right)+\frac{1}{8}\left(\frac{1}{T}+\frac{1}{t-1}\right)^2\right)\sum_{t=2}^{T}\left(\frac{(2\sqrt{L}+3L\rho)\rho}{L(t-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2}{nkt^2L^2}\right)\\
=&\exp\left(\sum_{t=2}^{T}\alpha\left(\frac{1}{T}+\frac{1}{t-1}\right)+\sum_{t=2}^{T}\frac{1}{8}\left(\frac{1}{T}+\frac{1}{t-1}\right)^2\right)\sum_{t=2}^{T}\left(\frac{(2\sqrt{L}+3L\rho)\rho}{L(t-1)}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2}{nkt^2L^2}\right)\\
\leq&T^{\alpha}\exp\left(2\alpha+\frac{2\log T+3}{8T}+\frac{\pi^2}{48}\right)\left(\frac{(2\sqrt{L}+3L\rho)\rho(\log T+1)}{L}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2\pi^2}{6nkL^2}\right)\\
\leq&T^{\alpha}e^3\left(\frac{(2\sqrt{L}+3L\rho)\rho(\log T+1)}{L}+\frac{(1+\frac{kT}{n})(2\sqrt{L}+\rho L)^2\pi^2}{6nkL^2}\right).
\end{aligned}
$$

The second inequality follows by iterating the recursion with $\mathbf{w}_1=\mathbf{w}_1^{(i)}$ and using $\alpha e^x\leq 1-\alpha+\alpha e^x$ and $1-\alpha+\alpha e^x\geq 1$ for $x\geq 0$ and $\alpha\in[0,1]$. The third inequality follows from $\log(1-\alpha+\alpha e^x)\leq \alpha x+\frac{x^2}{8}$, $\alpha\in[0,1]$. The fourth inequality is obtained using $\sum_{t=2}^{T}\frac{1}{t-1}\leq 1+\log T$, $\sum_{t=2}^{T}\frac{1}{t^2}\leq\frac{\pi^2}{6}$, and $\sum_{t=2}^{T}\left(\frac{1}{T}+\frac{1}{t-1}\right)^2\leq\frac{3+2\log T}{T}+\frac{\pi^2}{6}$. The last inequality follows from $2\alpha+\frac{2\log T+3}{8T}+\frac{\pi^2}{48}\leq 2\alpha+\frac{3}{8}+\frac{\pi^2}{48}<3$. □

B.3 Proof of Theorem 3.3

Theorem B.12. *Assume for all $z\in\mathcal{Z}$, the function $f(\mathbf{w},z)$ is L-smooth, nonnegative and bounded in $[0,1]$. Suppose $\alpha\in[\frac{1}{2},1]$. Let $\{\mathbf{w}_t\}$ be generated by SALA based on S. By setting $\eta_{\tau,t}=\frac{1}{\mu\left((t-1)k+\tau+1\right)}$ and $\rho_{\tau,t}\leq\frac{1}{\mu\left((t-1)k+\tau+1\right)}$, then we have*

$$
\begin{aligned}
&\mathbb{E}[F_S(\mathbf{w}_T)-F_S(\mathbf{w}(S))]\\
\leq&\frac{4}{(Tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right)-F_S\left(\mathbf{w}(S)\right)\right]+\frac{2\alpha(1-\alpha)L^2(2+L^{\frac{1}{2}})(k+1)}{\mu^2(Tk+1)^{2\alpha-1}}+\frac{2\alpha L^{\frac{3}{2}}+\alpha^2L(2\sqrt{L}+L)^2}{\mu^2(Tk+1)^{2\alpha-1}}.
\end{aligned}
$$

Proof. Firstly, we formulate the lookahead update as

$$
\begin{aligned}
\mathbf{w}_t&=(1-\alpha)\mathbf{w}_{t-1}+\alpha\boldsymbol{v}_{k+1,t}\\
&=(1-\alpha)\mathbf{w}_{t-1}+\alpha\left(\boldsymbol{v}_{k,t}-\eta_{k,t}\nabla f\left(\hat{\boldsymbol{v}}_{k,t},z_{i_{k,t}}\right)\right),\\
\mathbf{w}_t^{(i)}&=(1-\alpha)\mathbf{w}_{t-1}^{(i)}+\alpha\left(\boldsymbol{v}_{k,t}^{(i)}-\eta_{k,t}\nabla f\left(\hat{\boldsymbol{v}}_{k,t}^{(i)},z_{i_{k,t}}^{(i)}\right)\right).
\end{aligned}
$$

It is important to note that in the proof, the superscript (i) on the inner and outer loop parameters indicates that they are derived from the dataset $S^{(i)}$, where $S^{(i)}$ is the dataset obtained by replacing the original dataset S with a new data point $\tilde{z}_i$.

Based on the update rule of above algorithm, we define the variable $\boldsymbol{u}_{\tau,t}$

$$
\boldsymbol{u}_{\tau+1,t}=\alpha\boldsymbol{v}_{\tau+1,t}+(1-\alpha)\boldsymbol{v}_{1,t}=\boldsymbol{u}_{\tau,t}-\alpha\eta_{\tau,t}\hat{\boldsymbol{g}}_{\tau,t},
$$

where the $\hat{\boldsymbol{g}}_{\tau,t}$ is the gradient of the inner-loop parameter $\hat{\boldsymbol{v}}_{\tau,t}$. When τ equals $k+1$, the variable $\boldsymbol{u}_{\tau,t}$ is equivalent to $\mathbf{w}_t$. Therefore,

$$
F_S\left(\boldsymbol{u}_{\tau+1,t}\right)\leq F_S\left(\boldsymbol{u}_{\tau,t}\right)+\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\boldsymbol{u}_{\tau+1,t}-\boldsymbol{u}_{\tau,t}\right\rangle+\frac{L}{2}\left\|\boldsymbol{u}_{\tau+1,t}-\boldsymbol{u}_{\tau,t}\right\|^2
$$

$$=F_S\left(\boldsymbol{u}_{\tau,t}\right)-\alpha\eta_{\tau,t}\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\rangle+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2}{2}\left\|\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\|_2^2.$$

By taking expectation, we obtain:

$$\begin{aligned}
&\mathbb{E}[F_S(\boldsymbol{u}_{\tau+1,t})-F_S(\mathbf{w}(S))]\\
\leq&\mathbb{E}\Big[F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))-\alpha\eta_{\tau,t}\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\rangle+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2}{2}\left\|\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\|_2^2\Big]\\
=&\mathbb{E}\Big[F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))-\alpha\eta_{\tau,t}\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\nabla f(\boldsymbol{u}_{\tau,t},z_{i_{\tau,t}})\right\rangle\\
&+\alpha\eta_{\tau,t}\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\nabla f(\boldsymbol{u}_{\tau,t},z_{i_{\tau,t}})-\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\rangle+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2}{2}\left\|\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\|_2^2\Big]\\
\leq&\mathbb{E}\Big[F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))-\alpha\eta_{\tau,t}\|\nabla F_S(\boldsymbol{u}_{\tau,t})\|_2^2+\alpha\eta_{\tau,t}\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\nabla f(\boldsymbol{u}_{\tau,t},z_{i_{\tau,t}})-\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\rangle\\
&+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+\alpha\eta_{\tau,t}\left\langle\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right),\nabla f(\boldsymbol{u}_{\tau,t},z_{i_{\tau,t}})-\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\right\rangle\\
&+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+\alpha\eta_{\tau,t}\|\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right)\|_2\cdot\|\nabla f(\boldsymbol{u}_{\tau,t},z_{i_{\tau,t}})-\nabla f(\hat{\boldsymbol{v}}_{\tau,t},z_{i_{\tau,t}})\|_2\\
&+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+2\alpha\eta_{\tau,t}L^{\frac{3}{2}}\|\boldsymbol{u}_{\tau,t}-\boldsymbol{v}_{\tau,t}-\rho_{\tau,t}\frac{\nabla f(\boldsymbol{v}_{\tau,t},z_{i_{\tau,t}})}{\|\nabla f(\boldsymbol{v}_{\tau,t},z_{i_{\tau,t}})\|_2}\|_2\\
&+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+2\alpha\eta_{\tau,t}L^{\frac{3}{2}}\|\boldsymbol{u}_{\tau,t}-\boldsymbol{v}_{\tau,t}\|_2\\
&+2\alpha\rho_{\tau,t}\eta_{\tau,t}L^{\frac{3}{2}}+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+2\alpha(1-\alpha)\eta_{\tau,t}L^{\frac{3}{2}}\|\boldsymbol{v}_{1,t}-\boldsymbol{v}_{\tau,t}\|_2\\
&+2\alpha\rho_{\tau,t}\eta_{\tau,t}L^{\frac{3}{2}}+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+2\alpha(1-\alpha)\eta_{\tau,t}L^{\frac{3}{2}}\sum_{i=1}^{\tau-1}\eta_{i,t}\|\hat{\boldsymbol{g}}_{i,t}\|_2\\
&+2\alpha\rho_{\tau,t}\eta_{\tau,t}L^{\frac{3}{2}}+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big]\\
\leq&\mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t})-F_S(\mathbf{w}(S))\big)+2\alpha(1-\alpha)\eta_{\tau,t}L^{\frac{3}{2}}(2\sqrt{L}+\rho_{\tau,t}L)\sum_{i=1}^{\tau-1}\eta_{i,t}\\
&+2\alpha\rho_{\tau,t}\eta_{\tau,t}L^{\frac{3}{2}}+\frac{L\alpha^2\left(\eta_{\tau,t}\right)^2(2\sqrt{L}+L\rho_{\tau,t})^2}{2}\Big].
\end{aligned}$$

The third inequality holds since that $F_S(\boldsymbol{u}_{\tau,t})$ satisfies PL condition $2\mu\left(F_S(\boldsymbol{u}_{\tau,t})-F_S\left(\mathbf{w}(S)\right)\right)\leq\|\nabla F_S(\boldsymbol{u}_{\tau,t})\|^2$. Hence

$$\begin{aligned}
&\mathbb{E}[F_S(\boldsymbol{u}_{k+1,t})-F_S(\mathbf{w}(S))]\\
\leq&\big(1-2\alpha\mu\eta_{k,t}\big)\mathbb{E}[F_S(\boldsymbol{u}_{k,t})-F_S(\mathbf{w}(S))]+2\alpha(1-\alpha)\eta_{k,t}L^{\frac{3}{2}}(2\sqrt{L}+\rho_{k,t}L)\sum_{i=1}^{k-1}\eta_{i,t}\\
&+2\alpha\rho_{k,t}\eta_{k,t}L^{\frac{3}{2}}+\frac{L\alpha^2\left(\eta_{k,t}\right)^2(2\sqrt{L}+L\rho_{k,t})^2}{2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \prod_{j=1}^{k}(1-2\alpha\mu\eta_{j,t})\mathbb{E}[F_S(\boldsymbol{u}_{1,t})-F_S(\mathbf{w}(S))] \\
&\quad+\sum_{j=1}^{k}\prod_{j'=j+1}^{k}(1-2\alpha\mu\eta_{j',t})(2\alpha\eta_{j,t}L^{\frac{3}{2}}\rho_{j,t}+\frac{L\alpha^2\left(\eta_{j,t}\right)^2\left(2\sqrt{L}+L\rho_{j,t}\right)^2}{2}) \\
&\quad+\sum_{j=1}^{k}\prod_{j'=j+1}^{k}(1-2\alpha\mu\eta_{j',t})\big(2\alpha(1-\alpha)L^2(2+\rho_{j,t}L^{\frac{1}{2}})\eta_{j,t}\sum_{i=1}^{j-1}\eta_{i,t}\big) \\
&\leq \prod_{j=1}^{k}\left(1-\frac{2\alpha}{(t-1)k+j+1}\right)\mathbb{E}[F_S(\boldsymbol{u}_{1,t})-F_S(\mathbf{w}(S))] \\
&\quad+\sum_{j=1}^{k}\prod_{j'=j+1}^{k}\left(1-\frac{2\alpha}{(t-1)k+j'+1}\right)\frac{2\alpha L^{\frac{3}{2}}+\alpha^2L(2\sqrt{L}+L\rho_{j,t})^2}{(\mu(t-1)k+\mu j+\mu)^2} \\
&\quad+\sum_{j=1}^{k}\prod_{j'=j+1}^{k}\left(1-\frac{2\alpha}{(t-1)k+j'+1}\right)\frac{2\alpha(1-\alpha)L^2\left(2+\rho_{j,t}L^{\frac{1}{2}}\right)}{(\mu(t-1)k+\mu j+\mu)}\sum_{i=1}^{j-1}\frac{1}{\mu((t-1)k+i+1)} \\
&\leq \prod_{j=1}^{k}\left(\frac{(t-1)k+j}{(t-1)k+j+1}\right)^{2\alpha}\mathbb{E}[F_S(\boldsymbol{u}_{1,t})-F_S(\mathbf{w}(S))]+ \\
&\quad\sum_{j=1}^{k}\prod_{j'=j+1}^{k}\left(\frac{(t-1)k+j'}{(t-1)k+j'+1}\right)^{2\alpha}\frac{2\alpha L^{\frac{3}{2}}+\alpha^2L(2\sqrt{L}+L\rho_{j,t})^2}{(\mu(t-1)k+\mu j+\mu)^2} \\
&\quad+\sum_{j=1}^{k}\left[\frac{2\alpha(1-\alpha)L^2(2+\rho_{j,t}L^{\frac{1}{2}})}{\mu((t-1)k+j+1)}\sum_{i=1}^{j-1}\frac{1}{\mu((t-1)k+i+1)}\right]\prod_{j'=j+1}^{k}\left(\frac{(t-1)k+j'}{(t-1)k+j'+1}\right)^{2\alpha} \\
&\leq \left(\frac{(t-1)k+1}{tk+1}\right)^{2\alpha}\mathbb{E}[F_S(\boldsymbol{u}_{1,t})-F_S(\mathbf{w}(S))]+\frac{2\alpha L^{\frac{3}{2}}k+\alpha^2kL(2\sqrt{L}+L)^2}{\mu^2(tk+1)^{2\alpha}} \\
&\quad+\frac{2\alpha(1-\alpha)L^2(2+L^{\frac{1}{2}})k(k+1)}{\mu^2(tk+1)^{2\alpha}},
\end{aligned}
$$

where the third inequality holds due to that we assume $\eta_{\tau,t}=\frac{1}{\mu\left((t-1)k+\tau+1\right)}$ and $\rho_{\tau,t}\leq\frac{1}{\mu\left((t-1)k+\tau+1\right)}$. The fourth inequality follows from $1-2\alpha/x\leq(1-1/x)^{2\alpha}$ for $x\geq 2$ and $\alpha\in[\frac{1}{2},1]$. Since $\boldsymbol{u}_{1,t}=\mathbf{w}_{t-1}$, then we can obtain

$$
\begin{aligned}
&\mathbb{E}\left[F_S\left(\mathbf{w}_t\right)-F_S\left(\mathbf{w}(S)\right)\right] \\
&\leq\left(\frac{(t-1)k+1}{tk+1}\right)^{2\alpha}\mathbb{E}\left[F_S\left(\mathbf{w}_{t-1}\right)-F_S\left(\mathbf{w}(S)\right)\right] \\
&\quad+\frac{2\alpha L^{\frac{3}{2}}k+\alpha^2kL(2\sqrt{L}+L)^2}{\mu^2(tk+1)^{2\alpha}} \\
&\quad+\frac{2\alpha(1-\alpha)L^2(2+L^{\frac{1}{2}})k(k+1)}{\mu^2(tk+1)^{2\alpha}} \\
&\leq\prod_{i=1}^{t}\left(\frac{(i-1)k+1}{ik+1}\right)^{2\alpha}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right)-F_S\left(\mathbf{w}(S)\right)\right] \\
&\quad+\sum_{i=1}^{t}\frac{2\alpha(1-\alpha)L^2(2+L^{\frac{1}{2}})k(k+1)}{\mu^2(ik+1)^{2\alpha}}\cdot\prod_{j=i+1}^{t}\left(\frac{(j-1)k+1}{jk+1}\right)^{2\alpha}
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{i=1}^{t}\frac{2\alpha L^{\frac{3}{2}}k+\alpha^2kL(2\sqrt{L}+L)^2}{\mu^2(ik+1)^{2\alpha}}\cdot\prod_{j=i+1}^{t}\left(\frac{(j-1)k+1}{jk+1}\right)^{2\alpha}\\
\leq&\frac{4}{(tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right)-F_S\left(\mathbf{w}(S)\right)\right]\\
&+\frac{2\alpha(1-\alpha)L^2(2+L^{\frac{1}{2}})(k+1)+2\alpha L^{\frac{3}{2}}+\alpha^2L(2\sqrt{L}+L)^2}{\mu^2(tk+1)^{2\alpha-1}}.
\end{aligned}
$$

□

B.4 Proof of Theorem 3.7

Theorem B.13. *Assume for all $z\in\mathcal{Z}$, the function $f(\mathbf{w},z)$ is L-smooth, nonnegative and bounded in $[0,1]$, where $L>0$. Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(i)}\}$ be generated by the Lookahead-SGD algorithm based on S and $S^{(i)}$, respectively, with $\mathbf{w}_0=\mathbf{w}_0^{(i)}$ and the same sequence of independent, uniformly sampled indices. Setting $\alpha\in(0,1]$ and $\eta_{j,t}\leq\frac{1}{2Lkt}$ for all $j=1,\ldots,k$, then we have*

$$
\mathbb{E}\left[\left\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\right\|_2^2\right]\leq\frac{2\pi^2e^{2\alpha+\frac{5}{2}}\alpha}{3nkL}\left(1+\frac{kT}{n}\right)T^{\alpha}.
$$

Proof. Firstly, we formulate the lookahead update as

$$
\begin{aligned}
\mathbf{w}_t&=(1-\alpha)\mathbf{w}_{t-1}+\alpha\boldsymbol{v}_{k+1,t}\\
&=(1-\alpha)\mathbf{w}_{t-1}+\alpha\left(\boldsymbol{v}_{k,t}-\eta_{k,t}\nabla f\left(\boldsymbol{v}_{k,t},z_{i_{k,t}}\right)\right),\\
\mathbf{w}_t^{(i)}&=(1-\alpha)\mathbf{w}_{t-1}^{(i)}+\alpha\left(\boldsymbol{v}_{k,t}^{(i)}-\eta_{k,t}\nabla f\left(\boldsymbol{v}_{k,t}^{(i)},z_{i_{k,t}}^{(i)}\right)\right).
\end{aligned}
$$

It is important to note that in the proof, the superscript (i) on the inner and outer-loop parameters indicates that they are derived from the dataset $S^{(i)}$, where $S^{(i)}$ is obtained by replacing z_i in S with an independent copy $\tilde{z}_i$. Accordingly, $z_r^{(i)}=\tilde{z}_i$ if $r=i$, and $z_r^{(i)}=z_r$ otherwise. The inner loops are initialized with $\boldsymbol{v}_{1,t}=\mathbf{w}_{t-1}$ and $\boldsymbol{v}_{1,t}^{(i)}=\mathbf{w}_{t-1}^{(i)}$.

Then according to the expansiveness inequality for SGD, when $i_{\tau,t}\neq i$, we have

$$
\left\|\boldsymbol{v}_{\tau+1,t}-\boldsymbol{v}_{\tau+1,t}^{(i)}\right\|_2^2\leq\frac{1}{1-2\eta_{\tau,t}L}\left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2^2.
$$

Updating the parameters with different data z_i and $\tilde{z}_i$, the above relationship can be reshaped as follows:

$$
\begin{aligned}
\left\|\boldsymbol{v}_{\tau+1,t}-\boldsymbol{v}_{\tau+1,t}^{(i)}\right\|_2&\leq\left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+\eta_{\tau,t}\left\|\nabla f(\boldsymbol{v}_{\tau,t},z_i)-\nabla f(\boldsymbol{v}_{\tau,t}^{(i)},\tilde{z}_i)\right\|_2\\
&\leq\left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+\eta_{\tau,t}\big(\|\nabla f(\boldsymbol{v}_{\tau,t},z_i)\|_2+\|\nabla f(\boldsymbol{v}_{\tau,t}^{(i)},\tilde{z}_i)\|_2\big)\\
&\leq\left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+4\sqrt{L}\,\eta_{\tau,t}.
\end{aligned}
$$

The last inequality holds because nonnegativity and L-smoothness imply $\|\nabla f(\mathbf{w},z)\|_2^2\leq 2Lf(\mathbf{w},z)\leq 2L$, and hence $\|\nabla f(\mathbf{w},z)\|_2\leq 2\sqrt{L}$. Since the probability of selecting different samples is $\frac{1}{n}$, combining the above two situations and applying Young's inequality, for $t\geq 2$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{v}_{k+1,t}-\boldsymbol{v}_{k+1,t}^{(i)}\right\|_2^2\\
\leq&\left(\frac{1+p}{n}+\left(1-\frac{1}{n}\right)\frac{1}{1-2\eta_{k,t}L}\right)\mathbb{E}\left\|\boldsymbol{v}_{k,t}-\boldsymbol{v}_{k,t}^{(i)}\right\|_2^2\\
&+\frac{16L\eta_{k,t}^2(1+p^{-1})}{n}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left(\prod_{j=1}^{k}\left(\frac{p+1}{n}+\left(1-\frac{1}{n}\right)\frac{1}{1-2L\eta_{j,t}}\right)\right)\mathbb{E}\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2 \\
&\quad+\sum_{j=1}^{k}\left(\prod_{j'=j+1}^{k}\left(\frac{p+1}{n}+\left(1-\frac{1}{n}\right)\frac{1}{1-2L\eta_{j',t}}\right)\right)\frac{16L(1+p^{-1})\eta_{j,t}^2}{n} \\
&\leq \prod_{j=1}^{k}\left(\frac{1}{kT}+\frac{1}{n}+\left(1-\frac{1}{n}\right)\frac{1}{1-2L\eta_{j,t}}\right)\mathbb{E}\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2 \\
&\quad+\sum_{j=1}^{k}\left(\prod_{j'=j+1}^{k}\left(\frac{1}{kT}+\frac{1}{n}+\left(1-\frac{1}{n}\right)\frac{1}{1-2L\eta_{j',t}}\right)\right)\frac{16L(1+\frac{kT}{n})\eta_{j,t}^2}{n} \\
&\leq \left(\left(1+\frac{1}{kT}\right)^k\left(1+\frac{1}{kt-1}\right)^k\right)\mathbb{E}\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2 \\
&\quad+\left(\left(1+\frac{1}{kT}\right)^k\left(1+\frac{1}{kt-1}\right)^k\right)\frac{4(1+\frac{kT}{n})}{nkt^2L} \\
&\leq \exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\mathbb{E}\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2+\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\frac{4(1+\frac{kT}{n})}{nkt^2L},
\end{aligned}
$$

where the third inequality holds by setting $p=\frac{n}{kT}$. The fourth inequality follows from $\eta_{j,t}\leq\frac{1}{2Lkt}$ and the fact that every factor in the products is at least one. The last inequality uses $(1+x)^k\leq e^{kx}$.

Since $\|\cdot\|_2^2$ is convex and the lookahead update is a convex combination, we have

$$
\begin{aligned}
\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2 =&\|(1-\alpha)(\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)})+\alpha(\boldsymbol{v}_{k+1,t}-\boldsymbol{v}_{k+1,t}^{(i)})\|_2^2 \\
\leq&(1-\alpha)\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2+\alpha\|\boldsymbol{v}_{k+1,t}-\boldsymbol{v}_{k+1,t}^{(i)}\|_2^2.
\end{aligned}
$$

For the first outer iteration, $\mathbf{w}_0=\mathbf{w}_0^{(i)}$, and each suffix product in the inner-loop recursion satisfies

$$
\prod_{j'=j+1}^{k}\left(\frac{1}{kT}+\frac{1}{n}+\left(1-\frac{1}{n}\right)\frac{1}{1-2L\eta_{j',1}}\right)\leq\left(1+\frac{1}{kT}\right)^{k-1}\left(1+\frac{1}{k-1}\right)^{k-1}\leq e^2.
$$

Consequently,

$$
\mathbb{E}\|\mathbf{w}_1-\mathbf{w}_1^{(i)}\|_2^2=\alpha^2\mathbb{E}\|\boldsymbol{v}_{k+1,1}-\boldsymbol{v}_{k+1,1}^{(i)}\|_2^2\leq\frac{4e^2\alpha^2(1+\frac{kT}{n})}{nkL}\leq\frac{4e^2\alpha(1+\frac{kT}{n})}{nkL}.
$$

Combining the above results, for $T\geq 2$ we have

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\|_2^2 \\
&\leq\left(1-\alpha+\alpha\exp\left(\frac{1}{T}+\frac{1}{T-1}\right)\right)\mathbb{E}\|\mathbf{w}_{T-1}-\mathbf{w}_{T-1}^{(i)}\|_2^2 \\
&\quad+\exp\left(\frac{1}{T}+\frac{1}{T-1}\right)\frac{4\alpha(1+\frac{kT}{n})}{nkT^2L} \\
&\leq\prod_{t=2}^{T}\left(1-\alpha+\alpha\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\right)\left(\frac{4e^2\alpha(1+\frac{kT}{n})}{nkL}+\sum_{t=2}^{T}\frac{4\alpha(1+\frac{kT}{n})}{nkt^2L}\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\right) \\
&\leq\frac{4e^2\alpha(1+\frac{kT}{n})}{nkL}\prod_{t=2}^{T}\left(1-\alpha+\alpha\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\right)\sum_{t=1}^{T}\frac{1}{t^2} \\
&=\frac{4e^2\alpha(1+\frac{kT}{n})}{nkL}\prod_{t=2}^{T}\exp\left(\log\left(1-\alpha+\alpha\exp\left(\frac{1}{T}+\frac{1}{t-1}\right)\right)\right)\sum_{t=1}^{T}\frac{1}{t^2}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{4e^2\alpha(1+\frac{kT}{n})}{nkL}\prod_{t=2}^{T}\exp\left(\alpha\left(\frac{1}{T}+\frac{1}{t-1}\right)+\frac{1}{8}\left(\frac{1}{T}+\frac{1}{t-1}\right)^2\right)\sum_{t=1}^{T}\frac{1}{t^2}\\
&= \frac{4e^2\alpha(1+\frac{kT}{n})}{nkL}\exp\left(\sum_{t=2}^{T}\alpha\left(\frac{1}{T}+\frac{1}{t-1}\right)+\sum_{t=2}^{T}\frac{1}{8}\left(\frac{1}{T}+\frac{1}{t-1}\right)^2\right)\sum_{t=1}^{T}\frac{1}{t^2}\\
&\leq \frac{2\pi^2 e^{2\alpha+\frac{5}{2}}\alpha}{3nkL}\left(1+\frac{kT}{n}\right)T^{\alpha},
\end{aligned}
$$

where the second inequality follows by unrolling the recursion and enlarging the product and summation ranges, since all factors are at least one and all summands are nonnegative. The third inequality uses $\exp(\frac{1}{T}+\frac{1}{t-1}) \leq e^2$ for $t \geq 2$, and the fourth inequality follows from $\log(1-\alpha+\alpha e^x) \leq \alpha x + \frac{x^2}{8}$ for $\alpha \in [0,1]$. The last inequality is obtained using $\sum_{t=2}^{T}\left(\frac{1}{T}+\frac{1}{t-1}\right) \leq 2+\log T, \sum_{t=1}^{T}\frac{1}{t^2} \leq \frac{\pi^2}{6}, \sum_{t=2}^{T}\left(\frac{1}{T}+\frac{1}{t-1}\right)^2 \leq \frac{2(T-1)}{T^2}+2\sum_{t=2}^{T}\frac{1}{(t-1)^2} \leq \frac{1}{2}+\frac{\pi^2}{3} < 4$. The cases $T=0$ and $T=1$ follow directly from the common initialization and the first-iteration bound, respectively. □

B.5 Proof of Theorem 3.9

Theorem B.14. *Assume for all $z \in \mathcal{Z}$, the function $f(\mathbf{w}, z)$ is L-smooth, nonnegative and bounded in $[0,1]$. Suppose $\alpha \in [\frac{1}{2}, 1]$. Let $\{\mathbf{w}_t\}$ be generated by stochastic gradient descent-based lookahead on S. By setting $\eta_{\tau,t} = \frac{1}{\mu\left((t-1)k+\tau+1\right)}$, then we have*

$$
\mathbb{E}[F_S(\mathbf{w}_T) - F_S(\mathbf{w}(S))] \leq \frac{4}{(Tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right) - F_S\left(\mathbf{w}(S)\right)\right] + \frac{4\alpha(1-\alpha)L^2(k+1)+2\alpha L^2}{\mu^2(Tk+1)^{2\alpha-1}}.
$$

Proof. We formulate the lookahead update as

$$
\begin{aligned}
\mathbf{w}_t &= (1-\alpha)\mathbf{w}_{t-1} + \alpha \boldsymbol{v}_{k+1,t}\\
&= (1-\alpha)\mathbf{w}_{t-1} + \alpha\left(\boldsymbol{v}_{k,t} - \eta_{k,t}\nabla f\left(\boldsymbol{v}_{k,t}, z_{i_{k,t}}\right)\right),\\
\mathbf{w}_t^{(i)} &= (1-\alpha)\mathbf{w}_{t-1}^{(i)} + \alpha\left(\boldsymbol{v}_{k,t}^{(i)} - \eta_{k,t}\nabla f\left(\boldsymbol{v}_{k,t}^{(i)}, z_{i_{k,t}}^{(i)}\right)\right).
\end{aligned}
$$

Based on the update rule of above algorithm, we define the variable $\boldsymbol{u}_{\tau,t}$

$$
\boldsymbol{u}_{\tau+1,t} = \alpha\boldsymbol{v}_{\tau+1,t} + (1-\alpha)\boldsymbol{v}_{1,t} = \boldsymbol{u}_{\tau,t} - \alpha\eta_{\tau,t}\boldsymbol{g}_{\tau,t},
$$

where the $\boldsymbol{g}_{\tau,t}$ is the gradient of the inner-loop parameter $\boldsymbol{v}_{\tau,t}$. When τ equals $k+1$, the variable $\boldsymbol{u}_{\tau,t}$ is equivalent to $\mathbf{w}_t$.

$$
\begin{aligned}
F_S\left(\boldsymbol{u}_{\tau+1,t}\right) \leq& F_S\left(\boldsymbol{u}_{\tau,t}\right) + \left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \boldsymbol{u}_{\tau+1,t} - \boldsymbol{u}_{\tau,t}\right\rangle + \frac{L}{2}\left\|\boldsymbol{u}_{\tau+1,t} - \boldsymbol{u}_{\tau,t}\right\|^2\\
=& F_S\left(\boldsymbol{u}_{\tau,t}\right) - \alpha\eta_{\tau,t}\left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\rangle + \frac{L\alpha^2\left(\eta_{\tau,t}\right)^2}{2}\left\|\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\|_2^2.
\end{aligned}
$$

By taking expectation, we obtain:

$$
\begin{aligned}
&\mathbb{E}[F_S(\boldsymbol{u}_{\tau+1,t}) - F_S(\mathbf{w}(S))]\\
\leq& \mathbb{E}\Big[F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S)) - \alpha\eta_{\tau,t}\left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\rangle + \frac{L\alpha^2\left(\eta_{\tau,t}\right)^2}{2}\left\|\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\|_2^2\Big]\\
=& \mathbb{E}\Big[F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S)) - \alpha\eta_{\tau,t}\left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \nabla f(\boldsymbol{u}_{\tau,t}, z_{i_{\tau,t}})\right\rangle\\
&+ \alpha\eta_{\tau,t}\left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \nabla f(\boldsymbol{u}_{\tau,t}, z_{i_{\tau,t}}) - \nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\rangle + \frac{L\alpha^2\left(\eta_{\tau,t}\right)^2}{2}\left\|\nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\|_2^2\Big]\\
\leq& \mathbb{E}\Big[F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S)) - \alpha\eta_{\tau,t}\|\nabla F_S(\boldsymbol{u}_{\tau,t})\|_2^2 + \alpha\eta_{\tau,t}\left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \nabla f(\boldsymbol{u}_{\tau,t}, z_{i_{\tau,t}}) - \nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\quad + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big] \\
&\leq \mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S))\big) + \alpha\eta_{\tau,t}\left\langle \nabla F_S\left(\boldsymbol{u}_{\tau,t}\right), \nabla f(\boldsymbol{u}_{\tau,t}, z_{i_{\tau,t}}) - \nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\right\rangle \\
&\quad + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big] \\
&\leq \mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S))\big) + \alpha\eta_{\tau,t}\|\nabla F_S\left(\boldsymbol{u}_{\tau,t}\right)\|_2 \cdot \|\nabla f(\boldsymbol{u}_{\tau,t}, z_{i_{\tau,t}}) - \nabla f(\boldsymbol{v}_{\tau,t}, z_{i_{\tau,t}})\|_2 \\
&\quad + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big] \\
&\leq \mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S))\big) + 2\alpha\eta_{\tau,t}L^{\frac{3}{2}}\|\boldsymbol{u}_{\tau,t} - \boldsymbol{v}_{\tau,t}\|_2 + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big] \\
&\leq \mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S))\big) + 2\alpha(1-\alpha)\eta_{\tau,t}L^{\frac{3}{2}}\|\boldsymbol{v}_{1,t} - \boldsymbol{v}_{\tau,t}\|_2 + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big] \\
&\leq \mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S))\big) + 2\alpha(1-\alpha)\eta_{\tau,t}L^{\frac{3}{2}}\sum_{i=1}^{\tau-1}\eta_{i,t}\|\boldsymbol{g}_{i,t}\|_2 + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big] \\
&\leq \mathbb{E}\Big[(1-2\alpha\mu\eta_{\tau,t})\big(F_S(\boldsymbol{u}_{\tau,t}) - F_S(\mathbf{w}(S))\big) + 4\alpha(1-\alpha)\eta_{\tau,t}L^2\sum_{i=1}^{\tau-1}\eta_{i,t} + 2L^2\alpha^2\left(\eta_{\tau,t}\right)^2\Big].
\end{aligned}
$$

The third inequality holds since that $F_S(\boldsymbol{u}_{\tau,t})$ satisfies PL condition $2\mu\left(F_S(\boldsymbol{u}_{\tau,t}) - F_S\left(\mathbf{w}(S)\right)\right) \leq \|\nabla F_S(\boldsymbol{u}_{\tau,t})\|^2$. Hence

$$
\begin{aligned}
&\mathbb{E}[F_S(\boldsymbol{u}_{k+1,t}) - F_S(\mathbf{w}(S))] \\
&\leq \big(1-2\alpha\mu\eta_{k,t}\big)\mathbb{E}[F_S(\boldsymbol{u}_{k,t}) - F_S(\mathbf{w}(S))] + 4\alpha(1-\alpha)\eta_{k,t}L^2\sum_{i=1}^{k-1}\eta_{i,t} + 2L^2\alpha^2\left(\eta_{k,t}\right)^2 \\
&\leq \prod_{j=1}^{k}(1-2\alpha\mu\eta_{j,t})\mathbb{E}[F_S(\boldsymbol{u}_{1,t}) - F_S(\mathbf{w}(S))] + \sum_{j=1}^{k}\prod_{j'=j+1}^{k}(1-2\alpha\mu\eta_{j',t})(2L^2\alpha^2\eta_{j,t}^2) \\
&\quad + \sum_{j=1}^{k}\prod_{j'=j+1}^{k}(1-2\alpha\mu\eta_{j',t})\big(4\alpha(1-\alpha)L^2\eta_{j,t}\sum_{i=1}^{j-1}\eta_{i,t}\big) \\
&\leq \prod_{j=1}^{k}\left(1 - \frac{2\alpha}{(t-1)k+j+1}\right)\mathbb{E}[F_S(\boldsymbol{u}_{1,t}) - F_S(\mathbf{w}(S))] \\
&\quad + \sum_{j=1}^{k}\prod_{j'=j+1}^{k}\left(1 - \frac{2\alpha}{(t-1)k+j'+1}\right)\frac{2L^2\alpha^2}{(\mu(t-1)k+\mu j+\mu)^2} \\
&\quad + \sum_{j=1}^{k}\prod_{j'=j+1}^{k}\left(1 - \frac{2\alpha}{(t-1)k+j'+1}\right)\frac{4\alpha(1-\alpha)L^2}{(\mu(t-1)k+\mu j+\mu)}\sum_{i=1}^{j-1}\frac{1}{\mu((t-1)k+i+1)} \\
&\leq \prod_{j=1}^{k}\left(\frac{(t-1)k+j}{(t-1)k+j+1}\right)^{2\alpha}\mathbb{E}[F_S(\boldsymbol{u}_{1,t}) - F_S(\mathbf{w}(S))] \\
&\quad + \sum_{j=1}^{k}\prod_{j'=j+1}^{k}\left(\frac{(t-1)k+j'}{(t-1)k+j'+1}\right)^{2\alpha}\frac{2L^2\alpha^2}{(\mu(t-1)k+\mu j+\mu)^2} \\
&\quad + \sum_{j=1}^{k}\left[\frac{4\alpha(1-\alpha)L^2}{\mu((t-1)k+j+1)}\sum_{i=1}^{j-1}\frac{1}{\mu((t-1)k+i+1)}\right]\prod_{j'=j+1}^{k}\left(\frac{(t-1)k+j'}{(t-1)k+j'+1}\right)^{2\alpha} \\
&\leq \left(\frac{(t-1)k+1}{tk+1}\right)^{2\alpha}\mathbb{E}[F_S(\boldsymbol{u}_{1,t}) - F_S(\mathbf{w}(S))] + \frac{2L^2\alpha^2 k}{\mu^2(tk+1)^{2\alpha}} + \frac{4\alpha(1-\alpha)L^2k(k+1)}{\mu^2(tk+1)^{2\alpha}}.
\end{aligned}
$$

The fourth inequality follows from $1 - 2\alpha/x \leq (1-1/x)^{2\alpha}$ for $x \geq 2$ and $\alpha \in [\frac{1}{2}, 1]$. Since $\boldsymbol{u}_{1,t} = \mathbf{w}_{t-1}$, then we can obtain

$$
\begin{aligned}
&\mathbb{E}\left[F_S\left(\mathbf{w}_t\right)-F_S\left(\mathbf{w}(S)\right)\right] \\
\leq&\left(\frac{(t-1)k+1}{tk+1}\right)^{2\alpha}\mathbb{E}\left[F_S\left(\mathbf{w}_{t-1}\right)-F_S\left(\mathbf{w}(S)\right)\right]+\frac{2L^2\alpha^2k}{\mu^2(tk+1)^{2\alpha}}+\frac{4\alpha(1-\alpha)L^2k(k+1)}{\mu^2(tk+1)^{2\alpha}} \\
\leq&\prod_{i=1}^{t}\left(\frac{(i-1)k+1}{ik+1}\right)^{2\alpha}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right)-F_S\left(\mathbf{w}(S)\right)\right]+\sum_{i=1}^{t}\frac{4\alpha(1-\alpha)L^2k(k+1)}{\mu^2(ik+1)^{2\alpha}}\cdot\prod_{j=i+1}^{t}\left(\frac{(j-1)k+1}{jk+1}\right)^{2\alpha} \\
&+\sum_{i=1}^{t}\frac{2L^2\alpha^2k}{\mu^2(ik+1)^{2\alpha}}\cdot\prod_{j=i+1}^{t}\left(\frac{(j-1)k+1}{jk+1}\right)^{2\alpha} \\
\leq&\frac{4}{(tk+1)^{2\alpha}}\mathbb{E}\left[F_S\left(\mathbf{w}_0\right)-F_S\left(\mathbf{w}(S)\right)\right]+\frac{4\alpha(1-\alpha)L^2(k+1)+2\alpha L^2}{\mu^2(tk+1)^{2\alpha-1}}.
\end{aligned}
$$

□

B.6 Proof of Corollary 3.5

Theorem B.15. *Assume for all $z\in\mathcal{Z}$, the function $f(\mathbf{w},z)$ is nonnegative and convex and $\nabla f(\mathbf{w},z)$ is (β,L)-Hölder continuous with $\beta\in[0,1]$. Let $\{\mathbf{w}_t\}$ and $\left\{\mathbf{w}_t^{(i)}\right\}$ be generated by the SALA algorithm based on S and $S^{(i)}$, respectively. Then, we have*

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\right\|_2\right]\leq&\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\sqrt{\frac{1-\beta}{1+\beta}}L^{\frac{1}{1-\beta}}\eta_{j,t}^{\frac{1}{1-\beta}}+\frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}}L^{\frac{1}{2}}\eta_{j,t}^{\frac{1}{2}}\rho^{\frac{1+\beta}{2}}\right) \\
&+\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\frac{2\eta_{j,t}c_{\beta,1}}{n}\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{\beta}{1+\beta}}+\frac{2\eta_{j,t}\rho^{\beta}L}{n}\right).
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\right\|_2^2\right]\leq&\frac{e\alpha(1-\beta)}{1+\beta}L^{\frac{2}{1-\beta}}\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}e\alpha L}{1+\beta}\rho^{1+\beta}\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t} \\
&+\frac{8e\alpha}{n}\left(1+\frac{kT}{n}\right)\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right].
\end{aligned}
$$

Proof. Firstly, we formulate the lookahead update as

$$
\begin{aligned}
\mathbf{w}_t&=(1-\alpha)\mathbf{w}_{t-1}+\alpha\boldsymbol{v}_{k+1,t} \\
&=(1-\alpha)\mathbf{w}_{t-1}+\alpha\left(\boldsymbol{v}_{k,t}-\eta_{k,t}\nabla f\left(\hat{\boldsymbol{v}}_{k,t},z_{i_{k,t}}\right)\right), \\
\mathbf{w}_t^{(i)}&=(1-\alpha)\mathbf{w}_{t-1}^{(i)}+\alpha\left(\boldsymbol{v}_{k,t}^{(i)}-\eta_{k,t}\nabla f\left(\hat{\boldsymbol{v}}_{k,t}^{(i)},z_{i_{k,t}}^{(i)}\right)\right).
\end{aligned}
$$

Here, the superscript (i) indicates that the parameters are generated from $S^{(i)}=(z_1,\ldots,z_{i-1},\tilde{z}_i,z_{i+1},\ldots,z_n)$, where $\tilde{z}_i$ is an independent copy of z_i. In particular, $z_r^{(i)}=z_r$ for $r\neq i$ and $z_i^{(i)}=\tilde{z}_i$. The two runs use the same initialization and the same sequence of independent uniform sampling indices, with $\boldsymbol{v}_{1,t}=\mathbf{w}_{t-1}$ and $\boldsymbol{v}_{1,t}^{(i)}=\mathbf{w}_{t-1}^{(i)}$.

When $i_{\tau,t}\neq i$, according to Lemma B.8, we have

$$
\left\|\boldsymbol{v}_{\tau+1,t}-\boldsymbol{v}_{\tau+1,t}^{(i)}\right\|_2\leq\left\|\boldsymbol{v}_{\tau,t}-\boldsymbol{v}_{\tau,t}^{(i)}\right\|_2+\sqrt{\frac{1-\beta}{1+\beta}}L^{\frac{1}{1-\beta}}\eta_{\tau,t}^{\frac{1}{1-\beta}}+\frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}}L^{\frac{1}{2}}\eta_{\tau,t}^{\frac{1}{2}}\rho^{\frac{1+\beta}{2}}.
$$

When $i_{\tau,t}=i$, the two runs update using z_i and $\tilde{z}_i$, respectively. By the self-bounding property and Hölder continuity, we obtain

$$
\left\|\boldsymbol{v}_{\tau+1,t}-\boldsymbol{v}_{\tau+1,t}^{(i)}\right\|_2
$$

$$\leq \left\| \boldsymbol{v}_{\tau,t} - \boldsymbol{v}_{\tau,t}^{(i)} \right\|_2 + \eta_{\tau,t} \|\nabla f(\hat{\boldsymbol{v}}_{\tau,t}, z_i)\|_2 + \eta_{\tau,t} \|\nabla f(\hat{\boldsymbol{v}}_{\tau,t}^{(i)}, \tilde{z}_i)\|_2$$
$$\leq \left\| \boldsymbol{v}_{\tau,t} - \boldsymbol{v}_{\tau,t}^{(i)} \right\|_2 + \eta_{\tau,t} c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{\tau,t}, z_i) + \eta_{\tau,t} c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{\tau,t}^{(i)}, \tilde{z}_i) + 2\eta_{\tau,t}\rho^{\beta} L.$$

Combining the above two cases before taking expectations, and using the nonnegativity of the same-sample additive terms, we obtain

$$\begin{aligned}
&\left\| \boldsymbol{v}_{k+1,t} - \boldsymbol{v}_{k+1,t}^{(i)} \right\|_2 \\
\leq &\left\| \boldsymbol{v}_{k,t} - \boldsymbol{v}_{k,t}^{(i)} \right\|_2 + \sqrt{\frac{1-\beta}{1+\beta}} L^{\frac{1}{1-\beta}} \eta_{k,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}} L^{\frac{1}{2}} \eta_{k,t}^{\frac{1}{2}} \rho^{\frac{1+\beta}{2}} \\
&+ \eta_{k,t} \mathbf{1}_{\{i_{k,t}=i\}} \left(c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{k,t}, z_i) + c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{k,t}^{(i)}, \tilde{z}_i) + 2L\rho^{\beta} \right) \\
\leq &\|\mathbf{w}_{t-1} - \mathbf{w}_{t-1}^{(i)}\|_2 + \sum_{j=1}^{k} \left(\sqrt{\frac{1-\beta}{1+\beta}} L^{\frac{1}{1-\beta}} \eta_{j,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}} L^{\frac{1}{2}} \eta_{j,t}^{\frac{1}{2}} \rho^{\frac{1+\beta}{2}} \right) \\
&+ \sum_{j=1}^{k} \eta_{j,t} \mathbf{1}_{\{i_{j,t}=i\}} \left(c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i) + c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}^{(i)}, \tilde{z}_i) + 2L\rho^{\beta} \right).
\end{aligned}$$

Since $\|\cdot\|_2$ is convex and the Lookahead update is a convex combination, we have

$$\begin{aligned}
\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 =& \|(1-\alpha)\big(\mathbf{w}_{t-1} - \mathbf{w}_{t-1}^{(i)}\big) + \alpha\big(\boldsymbol{v}_{k+1,t} - \boldsymbol{v}_{k+1,t}^{(i)}\big)\|_2 \\
\leq& (1-\alpha)\|\mathbf{w}_{t-1} - \mathbf{w}_{t-1}^{(i)}\|_2 + \alpha\|\boldsymbol{v}_{k+1,t} - \boldsymbol{v}_{k+1,t}^{(i)}\|_2.
\end{aligned}$$

Combining the above results and using the same initialization, we have

$$\begin{aligned}
&\left\| \mathbf{w}_T - \mathbf{w}_T^{(i)} \right\|_2 \\
\leq &\left\| \mathbf{w}_{T-1} - \mathbf{w}_{T-1}^{(i)} \right\|_2 + \alpha \sum_{j=1}^{k} \left(\sqrt{\frac{1-\beta}{1+\beta}} L^{\frac{1}{1-\beta}} \eta_{j,T}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}} L^{\frac{1}{2}} \eta_{j,T}^{\frac{1}{2}} \rho^{\frac{1+\beta}{2}} \right) \\
&+ \alpha \sum_{j=1}^{k} \eta_{j,T} \mathbf{1}_{\{i_{j,T}=i\}} \left(c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,T}, z_i) + c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,T}^{(i)}, \tilde{z}_i) + 2L\rho^{\beta} \right) \\
\leq &\alpha \sum_{t=1}^{T} \sum_{j=1}^{k} \left(\sqrt{\frac{1-\beta}{1+\beta}} L^{\frac{1}{1-\beta}} \eta_{j,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}} L^{\frac{1}{2}} \eta_{j,t}^{\frac{1}{2}} \rho^{\frac{1+\beta}{2}} \right) \\
&+ \alpha \sum_{t=1}^{T} \sum_{j=1}^{k} \eta_{j,t} \mathbf{1}_{\{i_{j,t}=i\}} \left(c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i) + c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}^{(i)}, \tilde{z}_i) + 2L\rho^{\beta} \right).
\end{aligned}$$

Averaging over the replaced sample index gives

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{w}_T - \mathbf{w}_T^{(i)} \right\|_2 \\
\leq &\alpha \sum_{t=1}^{T} \sum_{j=1}^{k} \left(\sqrt{\frac{1-\beta}{1+\beta}} L^{\frac{1}{1-\beta}} \eta_{j,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}} L^{\frac{1}{2}} \eta_{j,t}^{\frac{1}{2}} \rho^{\frac{1+\beta}{2}} \right) \\
&+ \frac{\alpha}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{k} \eta_{j,t} \mathbf{1}_{\{i_{j,t}=i\}} \left(c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i) + c_{\beta,1} f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}^{(i)}, \tilde{z}_i) + 2L\rho^{\beta} \right).
\end{aligned}$$

Conditional on the datasets and all preceding sampling indices, the parameters $\boldsymbol{v}_{j,t}$ and $\boldsymbol{v}_{j,t}^{(i)}$ are determined before the current draw, and the conditional probability of $i_{j,t} = i$ is $1/n$. Moreover,

exchange symmetry gives $\mathbb{E}[f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i)] = \mathbb{E}[f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}^{(i)}, \tilde{z}_i)]$. Consequently, taking expectations in the preceding inequality and applying Jensen's inequality yields

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{w}_T - \mathbf{w}_T^{(i)}\right\|_2 \\
\leq&\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\sqrt{\frac{1-\beta}{1+\beta}}L^{\frac{1}{1-\beta}}\eta_{j,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}}L^{\frac{1}{2}}\eta_{j,t}^{\frac{1}{2}}\rho^{\frac{1+\beta}{2}}\right) \\
&+\frac{\alpha}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\frac{2\eta_{j,t}c_{\beta,1}}{n}\mathbb{E}[f^{\frac{\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i)] + \frac{2\eta_{j,t}L\rho^{\beta}}{n}\right) \\
\leq&\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\sqrt{\frac{1-\beta}{1+\beta}}L^{\frac{1}{1-\beta}}\eta_{j,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}}L^{\frac{1}{2}}\eta_{j,t}^{\frac{1}{2}}\rho^{\frac{1+\beta}{2}}\right) \\
&+\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\frac{2\eta_{j,t}c_{\beta,1}}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(\boldsymbol{v}_{j,t}, z_i)]\right)^{\frac{\beta}{1+\beta}} + \frac{2\eta_{j,t}L\rho^{\beta}}{n}\right) \\
=&\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\sqrt{\frac{1-\beta}{1+\beta}}L^{\frac{1}{1-\beta}}\eta_{j,t}^{\frac{1}{1-\beta}} + \frac{2^{\frac{2+\beta}{2}}}{\sqrt{1+\beta}}L^{\frac{1}{2}}\eta_{j,t}^{\frac{1}{2}}\rho^{\frac{1+\beta}{2}}\right) \\
&+\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(\frac{2\eta_{j,t}c_{\beta,1}}{n}\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{\beta}{1+\beta}} + \frac{2\eta_{j,t}L\rho^{\beta}}{n}\right).
\end{aligned}
$$

We next establish the ℓ_2-on-average stability. When $i_{j,t} \neq i$, Lemma B.8 yields

$$
\|\boldsymbol{v}_{j+1,t} - \boldsymbol{v}_{j+1,t}^{(i)}\|_2^2 \leq \|\boldsymbol{v}_{j,t} - \boldsymbol{v}_{j,t}^{(i)}\|_2^2 + \frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}} + \frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta}.
$$

When $i_{j,t} = i$, applying Young's inequality to the inner-loop update, followed by Hölder continuity and the self-bounding property, gives

$$
\begin{aligned}
&\|\boldsymbol{v}_{j+1,t} - \boldsymbol{v}_{j+1,t}^{(i)}\|_2^2 \\
\leq&\left(1+\frac{n}{kT}\right)\|\boldsymbol{v}_{j,t} - \boldsymbol{v}_{j,t}^{(i)}\|_2^2 + \left(1+\frac{kT}{n}\right)\eta_{j,t}^2\|\nabla f(\hat{\boldsymbol{v}}_{j,t}, z_i) - \nabla f(\hat{\boldsymbol{v}}_{j,t}^{(i)}, \tilde{z}_i)\|_2^2 \\
\leq&\left(1+\frac{n}{kT}\right)\|\boldsymbol{v}_{j,t} - \boldsymbol{v}_{j,t}^{(i)}\|_2^2 + 4\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2 f^{\frac{2\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i) + c_{\beta,1}^2 f^{\frac{2\beta}{1+\beta}}(\boldsymbol{v}_{j,t}^{(i)}, \tilde{z}_i) + 2L^2\rho^{2\beta}\right].
\end{aligned}
$$

Conditional on the datasets and all preceding sampling indices, the parameters $\boldsymbol{v}_{j,t}$ and $\boldsymbol{v}_{j,t}^{(i)}$ are determined before the current draw, and $\mathbb{P}(i_{j,t} = i) = 1/n$. Combining the two cases, taking expectations, and averaging over i, we obtain

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{v}_{j+1,t} - \boldsymbol{v}_{j+1,t}^{(i)}\|_2^2 \\
\leq&\left(1+\frac{1}{kT}\right)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{v}_{j,t} - \boldsymbol{v}_{j,t}^{(i)}\|_2^2 + \frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}} + \frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta} \\
&+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[\frac{c_{\beta,1}^2}{n}\sum_{i=1}^{n}\mathbb{E}\left[f^{\frac{2\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_i)\right] + L^2\rho^{2\beta}\right] \\
\leq&\left(1+\frac{1}{kT}\right)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{v}_{j,t} - \boldsymbol{v}_{j,t}^{(i)}\|_2^2 + \frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}} + \frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta} \\
&+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}} + L^2\rho^{2\beta}\right].
\end{aligned}
$$

Here, the first inequality uses exchange symmetry and $(1-1/n)+(1+n/(kT))/n = 1+1/(kT)$. The second inequality follows from Jensen's inequality.

Iterating over $j=1,\dots,k$, using $\boldsymbol{v}_{1,t}=\mathbf{w}_{t-1}$, and applying the convexity of the squared norm to the outer-loop update, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\\
\le&\left[1-\alpha+\alpha\left(1+\frac{1}{kT}\right)^k\right]\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{w}_{t-1}-\mathbf{w}_{t-1}^{(i)}\|_2^2\\
&+\alpha\sum_{j=1}^{k}\left(1+\frac{1}{kT}\right)^{k-j}\left[\frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta}\right.\\
&\left.+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right]\right].
\end{aligned}$$

Since $\mathbf{w}_0=\mathbf{w}_0^{(i)}$, iterating over $t=1,\dots,T$ gives

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\|_2^2\\
\le&\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left[1-\alpha+\alpha\left(1+\frac{1}{kT}\right)^k\right]^{T-t}\left(1+\frac{1}{kT}\right)^{k-j}\\
&\times\left[\frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta}+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right]\right]\\
\le&\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left(1+\frac{1}{kT}\right)^{k(T-t+1)-j}\left[\frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta}\right.\\
&\left.+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right]\right]\\
\le&\alpha\left(1+\frac{1}{kT}\right)^{kT}\sum_{t=1}^{T}\sum_{j=1}^{k}\left[\frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta}\right.\\
&\left.+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right]\right]\\
\le&e\alpha\sum_{t=1}^{T}\sum_{j=1}^{k}\left[\frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}L}{1+\beta}\eta_{j,t}\rho^{1+\beta}\right.\\
&\left.+\frac{8}{n}\left(1+\frac{kT}{n}\right)\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right]\right]\\
=&\frac{e\alpha(1-\beta)}{1+\beta}L^{\frac{2}{1-\beta}}\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^{\frac{2}{1-\beta}}+\frac{2^{2+\beta}e\alpha L}{1+\beta}\rho^{1+\beta}\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}\\
&+\frac{8e\alpha}{n}\left(1+\frac{kT}{n}\right)\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}}+L^2\rho^{2\beta}\right].
\end{aligned}$$

Here, the second inequality follows from $1-\alpha+\alpha(1+1/(kT))^k \le (1+1/(kT))^k$ for $\alpha\in(0,1]$. The third inequality uses $k(T-t+1)-j\le kT$, and the fourth follows from $1+x\le e^x$. □

Lemma B.16. *Assume for all $z\in\mathcal{Z}$, the function $f(\mathbf{w},z)$ is nonnegative and convex, and $\nabla f(\mathbf{w},z)$ is (β,L)-Hölder continuous with $\beta\in(0,1)$. Let $\{\boldsymbol{v}_{j,t}\}$ and $\{\mathbf{w}_t\}$ be generated by the SALA*

algorithm based on S, with deterministic initialization $\mathbf{w}_0$. Let $\mathbf{w}(S) = \arg\min_{\mathbf{w}} F_S(\mathbf{w})$ and $\mathbb{E}_S[F_S(\mathbf{w}(S))] < \infty$, we have

$$\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}\mathbb{E}[F_S(\boldsymbol{v}_{j,t})] \leq \frac{\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2}{\alpha} + \left[2\mathbb{E}_S[F_S(\mathbf{w}(S))] + \left(1 + \frac{2L}{1+\beta}\right)\rho^{1+\beta}\right]\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t} + \frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left[(2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right]\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^{\frac{2}{1-\beta}},$$

where $c_{\beta,1} = (1 + 1/\beta)^{\beta/(1+\beta)}L^{1/(1+\beta)}$.

Proof. The perturbation is set to zero whenever the gradient vanishes. By Hölder continuity and $\|\hat{\boldsymbol{v}}_{j,t} - \boldsymbol{v}_{j,t}\|_2 \leq \rho$, we have

$$f(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) \leq f(\hat{\boldsymbol{v}}_{j,t}, z_{i_{j,t}}) + \left\langle \nabla f(\hat{\boldsymbol{v}}_{j,t}, z_{i_{j,t}}), \boldsymbol{v}_{j,t} - \hat{\boldsymbol{v}}_{j,t}\right\rangle + \frac{L}{1+\beta}\rho^{1+\beta}.$$

Combining this with the convexity of $f(\cdot, z_{i_{j,t}})$ gives

$$\begin{aligned}\left\langle \boldsymbol{v}_{j,t} - \mathbf{w}(S), \nabla f(\hat{\boldsymbol{v}}_{j,t}, z_{i_{j,t}})\right\rangle &\geq f(\hat{\boldsymbol{v}}_{j,t}, z_{i_{j,t}}) - f(\mathbf{w}(S), z_{i_{j,t}}) + \left\langle \boldsymbol{v}_{j,t} - \hat{\boldsymbol{v}}_{j,t}, \nabla f(\hat{\boldsymbol{v}}_{j,t}, z_{i_{j,t}})\right\rangle \\ &\geq f(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) - f(\mathbf{w}(S), z_{i_{j,t}}) - \frac{L}{1+\beta}\rho^{1+\beta}.\end{aligned}$$

Furthermore, the self-bounding property and Hölder continuity imply

$$\|\nabla f(\hat{\boldsymbol{v}}_{j,t}, z_{i_{j,t}})\|_2^2 \leq 2\|\nabla f(\boldsymbol{v}_{j,t}, z_{i_{j,t}})\|_2^2 + 2L^2\rho^{2\beta} \leq 2c_{\beta,1}^2 f^{\frac{2\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) + 2L^2\rho^{2\beta}.$$

It follows from the inner-loop update that

$$\begin{aligned}\|\boldsymbol{v}_{j+1,t} - \mathbf{w}(S)\|_2^2 \leq &\|\boldsymbol{v}_{j,t} - \mathbf{w}(S)\|_2^2 - 2\eta_{j,t}\left(f(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) - f(\mathbf{w}(S), z_{i_{j,t}})\right) \\ &+ 2c_{\beta,1}^2\eta_{j,t}^2 f^{\frac{2\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) + \frac{2L}{1+\beta}\eta_{j,t}\rho^{1+\beta} + 2L^2\eta_{j,t}^2\rho^{2\beta}. \qquad (8)\end{aligned}$$

By Young's inequality with conjugate exponents $(1+\beta)/(2\beta)$ and $(1+\beta)/(1-\beta)$,

$$\begin{aligned}2c_{\beta,1}^2\eta_{j,t}f^{\frac{2\beta}{1+\beta}}(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) \leq &f(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) + \frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}(2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}}\eta_{j,t}^{\frac{1+\beta}{1-\beta}}, \\ 2L^2\eta_{j,t}\rho^{2\beta} \leq &\rho^{1+\beta} + \frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}(2L^2)^{\frac{1+\beta}{1-\beta}}\eta_{j,t}^{\frac{1+\beta}{1-\beta}}.\end{aligned}$$

Multiplying these inequalities by $\eta_{j,t}$ and substituting them into equation 8, we obtain

$$\begin{aligned}&\eta_{j,t}f(\boldsymbol{v}_{j,t}, z_{i_{j,t}}) \\ \leq &\|\boldsymbol{v}_{j,t} - \mathbf{w}(S)\|_2^2 - \|\boldsymbol{v}_{j+1,t} - \mathbf{w}(S)\|_2^2 + 2\eta_{j,t}f(\mathbf{w}(S), z_{i_{j,t}}) \\ &+ \left(1 + \frac{2L}{1+\beta}\right)\eta_{j,t}\rho^{1+\beta} + \frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left[(2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right]\eta_{j,t}^{\frac{2}{1-\beta}}.\end{aligned}$$

Since $\mathbf{w}(S)$ depends only on S, conditional uniform sampling gives $\mathbb{E}[f(\boldsymbol{v}_{j,t}, z_{i_{j,t}})] = \mathbb{E}[F_S(\boldsymbol{v}_{j,t})]$ and $\mathbb{E}[f(\mathbf{w}(S), z_{i_{j,t}})] = \mathbb{E}_S[F_S(\mathbf{w}(S))]$. Moreover, the outer-loop update and the convexity of the squared norm imply $\|\mathbf{w}_t - \mathbf{w}(S)\|_2^2 \leq (1-\alpha)\|\mathbf{w}_{t-1} - \mathbf{w}(S)\|_2^2 + \alpha\|\boldsymbol{v}_{k+1,t} - \mathbf{w}(S)\|_2^2$. Taking expectations in equation 9, summing over the inner and outer iterations, and using $\boldsymbol{v}_{1,t} = \mathbf{w}_{t-1}$, we derive

$$\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}\mathbb{E}[F_S(\boldsymbol{v}_{j,t})] \leq \frac{\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2}{\alpha} + \left[2\mathbb{E}_S[F_S(\mathbf{w}(S))] + \left(1 + \frac{2L}{1+\beta}\right)\rho^{1+\beta}\right]\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t} + \frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left[(2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right]\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^{\frac{2}{1-\beta}}.$$

Setting $\eta_{j,t} = c(kT)^{-\theta}$ with $c > 0$, Lemma B.16 gives

$$\frac{1}{kT}\sum_{t=1}^{T}\sum_{j=1}^{k}\mathbb{E}[F_S(\boldsymbol{v}_{j,t})] \leq \frac{\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2}{\alpha c(kT)^{1-\theta}} + 2\mathbb{E}_S[F_S(\mathbf{w}(S))] + \left(1 + \frac{2L}{1+\beta}\right)\rho^{1+\beta} + \frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left[(2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right]c^{\frac{1+\beta}{1-\beta}}(kT)^{-\frac{(1+\beta)\theta}{1-\beta}}.$$

By Jensen's inequality and the subadditivity of $x \mapsto x^{2\beta/(1+\beta)}$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}} + L^2\rho^{2\beta}\right] \\
\leq& c^2c_{\beta,1}^2(kT)^{1-2\theta}\left(\frac{1}{kT}\sum_{t=1}^{T}\sum_{j=1}^{k}\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}} + c^2L^2\rho^{2\beta}(kT)^{1-2\theta} \\
\leq& c_{\beta,1}^2c^{\frac{2}{1+\beta}}\alpha^{-\frac{2\beta}{1+\beta}}\left(\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2\right)^{\frac{2\beta}{1+\beta}}(kT)^{\frac{1-\beta-2\theta}{1+\beta}} \\
&+ c^2(kT)^{1-2\theta}\left[2^{\frac{2\beta}{1+\beta}}c_{\beta,1}^2\left(\mathbb{E}_S[F_S(\mathbf{w}(S))]\right)^{\frac{2\beta}{1+\beta}} + \left(c_{\beta,1}^2\left(1 + \frac{2L}{1+\beta}\right)^{\frac{2\beta}{1+\beta}} + L^2\right)\rho^{2\beta}\right] \\
&+ c_{\beta,1}^2c^{\frac{2}{1-\beta}}\left[\frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left((2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right)\right]^{\frac{2\beta}{1+\beta}}(kT)^{1-\frac{2\theta}{1-\beta}}.
\end{aligned}$$

Consequently, the empirical-risk contribution in the stability bound satisfies

$$\begin{aligned}
&\frac{8e\alpha}{n}\left(1 + \frac{kT}{n}\right)\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^2\left[c_{\beta,1}^2\left(\mathbb{E}[F_S(\boldsymbol{v}_{j,t})]\right)^{\frac{2\beta}{1+\beta}} + L^2\rho^{2\beta}\right] \\
\leq& \frac{8ec_{\beta,1}^2c^{\frac{2}{1+\beta}}}{n}\left(1 + \frac{kT}{n}\right)\alpha^{\frac{1-\beta}{1+\beta}}\left(\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2\right)^{\frac{2\beta}{1+\beta}}(kT)^{\frac{1-\beta-2\theta}{1+\beta}} \\
&+ \frac{8e\alpha c^2}{n}\left(1 + \frac{kT}{n}\right)(kT)^{1-2\theta}\left[2^{\frac{2\beta}{1+\beta}}c_{\beta,1}^2\left(\mathbb{E}_S[F_S(\mathbf{w}(S))]\right)^{\frac{2\beta}{1+\beta}} + \left(c_{\beta,1}^2\left(1 + \frac{2L}{1+\beta}\right)^{\frac{2\beta}{1+\beta}} + L^2\right)\rho^{2\beta}\right] \\
&+ \frac{8e\alpha c_{\beta,1}^2c^{\frac{2}{1-\beta}}}{n}\left(1 + \frac{kT}{n}\right)\left[\frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left((2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right)\right]^{\frac{2\beta}{1+\beta}}(kT)^{1-\frac{2\theta}{1-\beta}}.
\end{aligned}$$

Since $\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t}^{2/(1-\beta)} = c^{2/(1-\beta)}(kT)^{1-2\theta/(1-\beta)}$ and $\sum_{t=1}^{T}\sum_{j=1}^{k}\eta_{j,t} = c(kT)^{1-\theta}$, substituting the preceding estimate into the stability bound and collecting the terms with the same power of kT yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{w}_T - \mathbf{w}_T^{(i)}\|_2^2 \\
\leq& e\alpha c^{\frac{2}{1-\beta}}(kT)^{1-\frac{2\theta}{1-\beta}}\left\{\frac{1-\beta}{1+\beta}L^{\frac{2}{1-\beta}} + \frac{8c_{\beta,1}^2}{n}\left(1 + \frac{kT}{n}\right)\right. \\
&\left.\times\left[\frac{1-\beta}{1+\beta}\left(\frac{2\beta}{1+\beta}\right)^{\frac{2\beta}{1-\beta}}\left((2c_{\beta,1}^2)^{\frac{1+\beta}{1-\beta}} + (2L^2)^{\frac{1+\beta}{1-\beta}}\right)\right]^{\frac{2\beta}{1+\beta}}\right\} + \frac{2^{2+\beta}e\alpha Lc}{1+\beta}\rho^{1+\beta}(kT)^{1-\theta} \\
&+ \frac{8ec_{\beta,1}^2c^{\frac{2}{1+\beta}}}{n}\left(1 + \frac{kT}{n}\right)\alpha^{\frac{1-\beta}{1+\beta}}\left(\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2\right)^{\frac{2\beta}{1+\beta}}(kT)^{\frac{1-\beta-2\theta}{1+\beta}}
\end{aligned}$$

$$+\frac{8e\alpha c^2}{n}\left(1+\frac{kT}{n}\right)(kT)^{1-2\theta}\left[2^{\frac{2\beta}{1+\beta}}c_{\beta,1}^2\left(\mathbb{E}_S[F_S(\mathbf{w}(S))]\right)^{\frac{2\beta}{1+\beta}}+\left(c_{\beta,1}^2\left(1+\frac{2L}{1+\beta}\right)^{\frac{2\beta}{1+\beta}}+L^2\right)\rho^{2\beta}\right].$$

Suppose $kT \asymp n$, $\rho = O(n^{-1})$, $\mathbb{E}_S\|\mathbf{w}_0 - \mathbf{w}(S)\|_2^2 = O(1)$, and $\mathbb{E}_S[F_S(\mathbf{w}(S))] = O(1)$. The preceding stability bound yields the following results.

(a) If $\frac{1}{2} \leq \beta < 1$, we take $\theta = \frac{1}{2}$ to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{w}_T - \mathbf{w}_T^{(i)}\|_2^2 \leq O\left(\frac{\alpha}{n} + \alpha^{\frac{1-\beta}{1+\beta}}n^{-\frac{1+2\beta}{1+\beta}}\right).$$

(b) If $0 < \beta < \frac{1}{2}$, we take $\theta = 1 - \beta$ to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{w}_T - \mathbf{w}_T^{(i)}\|_2^2 \leq O\left(\frac{\alpha}{n} + \alpha^{\frac{1-\beta}{1+\beta}}n^{-\frac{2}{1+\beta}}\right).$$

□