GENERALIZATION AND OPTIMIZATION OF SGD WITH LOOKAHEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

The Lookahead optimizer (Zhang et al., 2019) enhances deep learning models by employing a dual-weight update mechanism, which has been shown to improve the performance of underlying optimizers such as SGD. However, most theoretical studies focus on its convergence on training data, leaving its generalization capabilities less understood. Existing generalization analyses are often limited by restrictive assumptions, such as requiring the loss function to be globally Lipschitz continuous, and their bounds do not fully capture the relationship between optimization and generalization. In this paper, we address these issues by conducting a rigorous stability and generalization analysis of the Lookahead optimizer with minibatch SGD. We leverage on-average model stability to derive generalization bounds for both convex and strongly convex problems without the restrictive Lipschitzness assumption. Our analysis demonstrates a linear speedup with respect to the batch size in the convex setting.

1 INTRODUCTION

Stochastic optimization has become the method of choice to train modern machine learning models due to its efficiency and scalability (Kingma & Ba, 2014). A simple stochastic optimization method is the minibatch stochastic gradient descent (minibatch SGD) (Cotter et al., 2011b; Dekel et al., 2012; Li et al., 2014; Shamir & Srebro, 2014), where a minibatch of training examples are randomly sampled to build gradient estimates with a reduced variance. Due to its simplicity, computational efficiency and strong generalization in practice (Zhou et al., 2020; Bottou et al., 2018), minibatch SGD remains one of the most preferable algorithms. Another representative stochastic optimization method is Adam (Kingma & Ba, 2014), which augments SGD with coordinate-wise adaptive learning rates and momentum, often accelerating convergence and improving robustness to ill-conditioning.

To further enhance generalization performance, the Lookahead optimizer (Zhang et al., 2019) was introduced as an orthogonal method. It introduces a two-timescale updating framework of two parameters: the fast weights $\mathbf{v}$ and the slow weights $\mathbf{w}$. In the inner loop, starting from the slow weights $\mathbf{w}$, the fast weights are updated by applying a standard optimizer $\mathcal{A}$ for k times and output $\mathbf{v}_k$; for the outer loop, the slow weights are updated towards the fast weights by $\mathbf{w}_+ = \alpha\mathbf{v}_k + (1-\alpha)\mathbf{w}$, where $\alpha \in (0, 1]$ is an interpolation parameter. This mechanism dampens oscillations, reduces sensitivity to learning-rate schedules and synchronization periods, and improves robustness across tasks with negligible overhead, often matching or improving the accuracy of the underlying base optimizer (Zhang et al., 2019).

The empirical efficiency of the Lookahead optimizer motivates a lot of theoretical studies to understand its behavior. However, most of existing studies focus on their convergence to minimize the training errors (Yang et al., 2024; Chen et al., 2022b; Zhang et al., 2019). As a comparison, there are far less studies on how the training behavior generalizes to testing examples, which is a concept of central interest in machine learning. To our best knowledge, the only work on the generalization analysis is Zhou et al. (2021), which conducted a stability analysis to argue that the Lookahead optimizer can generalize better than SGD and Adam. While these results provide a sound foundation on the use of the Lookahead mechanism, there are still some issues to be addressed. For example, their analysis hinges on the Lipschitzness condition on the loss, which is often restrictive in high-dimensional problems where gradients can be unbounded and the loss landscapes are non-Lipschitz globally.

Furthermore, their stability bounds are not optimistic and cannot fully capture the connection between generalization and optimization.

This paper aims to address the above issues by improving the existing stability and generalization analysis of the Lookahead optimizer. Our main contributions can be summarized as follows.

1. We leverage the on-average model stability to analyze the generalization behavior of the Lookahead methods for both convex and strongly convex problems. Our analysis removes the restrictive Lipschitzness assumptions of the loss functions, which can imply effective generalization bounds in the case with unbounded gradients. Furthermore, our analysis clearly shows how the interpolation parameter α strengthens the stability, which shows a clear benefit of the Lookahead mechanism.
2. Our stability bounds are optimistic, meaning that they depend on the empirical risk of the iterates produced by the algorithm. As the optimizer minimizes the empirical risk during the optimization process, our bounds become progressively tighter, offering a more refined and practical characterization of stability compared to existing bounds that rely on worst-case global constants.
3. By carefully combining our stability bounds with the convergence rates, we establish optimal excess risk rates for SGD with Lookahead. We show that it achieves a rate of $\mathcal{O}(1/n)$ for convex problems and a rate of $\mathcal{O}(1/(n\mu))$ for μ-strongly convex problems, where n is the sample size. Furthermore, our analysis shows a linear speedup with respect to the batch size b, meaning that the number of required iterations is decreased by a factor of b to achieve the optimal excess risk bounds.

The paper is organized as follows. We review the related work in Section 2 and introduce the problem formulation in Section 3. We present our main theoretical results in Section 5. The detailed proofs are provided in Appendix A. We conclude the paper in Section 6.

2 Related Work

Stability and Generalization Analysis A central challenge in machine learning is ensuring that models generalize well from finite training data to unseen examples. Algorithmic stability is an effective concept to study the generalization gap of learning algorithms, which can incorporate the special property of learning algorithms to derive algorithm-dependent generalization bounds (Bousquet & Elisseeff, 2002). A most widely used stability measure is the uniform stability, which is frequently used to analyze the generalization of regularization methods (Bousquet & Elisseeff, 2002) and stochastic optimization methods (Hardt et al., 2016). This stability concept was relaxed to on-average stability and on-average model stability to derive data-dependent generalization bounds (Shalev-Shwartz et al., 2010; Kuzborskij & Lampert, 2018; Lei & Ying, 2020; Schliserman & Koren, 2022). Recently, algorithm stability has found very successful applications in understanding the generalization behavior of complex models and training paradigms, including zeroth-order SGD (Nikolakakis et al., 2022; Chen et al., 2023), differential privacy (Bassily et al., 2019; 2020), asynchronous SGD (Deng et al., 2025) and neural network training (Richards & Kuzborskij, 2021; Wang et al., 2025a; Taheri & Thrampoulidis, 2024; Deora et al., 2024).

Lookahead Optimizer The Lookahead optimizer (Zhang et al., 2019) represents a significant advancement in optimization techniques for deep learning by employing a dual-weight update mechanism that separates "fast weights" (updated via a base optimizer) and "slow weights" (updated through exponential moving averaging). It reduces sensitivity to hyperparameters such as learning rates and synchronization periods, making it particularly robust in complex training scenarios where conventional optimizers struggle with oscillation or divergence (Nag, 2020; Zuo et al., 2024). Lookahead is widely adopted and extended across diverse domains including online learning (Chen et al., 2022a), aircraft maintenance scheduling (Deng & Santos, 2022), reinforcement learning (Merlis, 2024; Winnicki et al., 2025; Zhang et al., 2025), precision path tracking (Wang et al., 2025b), and healthcare prediction (Chen et al., 2022c; Adeshina & Adedigba, 2022). Various algorithmic extensions for Lookahead have also been introduced, including Multilayer Lookahead (Pushkin & Barba, 2021), Sharpness-Aware Lookahead (SALA) (Tan et al., 2024), Multi-step Lookahead Bayesian Optimization (Byun et al., 2022), and Lookaround Optimizer (Zhang et al., 2023).

3 NOTATIONS AND PRELIMINARIES

Let $\mathcal{D}$ be a probability measure defined on a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is an output space. Let $S = \{z_1, z_2, \ldots, z_n\}$ be a sample drawn independently and identically (i.i.d.) from $\mathcal{D}$, based on which we aim to learn a model $h : \mathcal{X} \mapsto \mathbb{R}$ for prediction. We assume the model is characterized by a parameter $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$, where $\mathcal{W}$ is a parameter space. The performance of a model $\mathbf{w}$ on a single data point z is measured by a non-negative loss function $f(\mathbf{w}; z)$, from which we can define empirical risks $F_S(\mathbf{w})$ and population risks $F(\mathbf{w})$ to measure the behavior of $\mathbf{w}$ on training and testing datasets, respectively

$$F_S(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i) \quad \text{and} \quad F(\mathbf{w}) := \mathbb{E}_{z\sim\mathcal{D}}[f(\mathbf{w}; z)],$$

where $\mathbb{E}_z[\cdot]$ means the expectation w.r.t. z.

We often apply a randomized optimizer A to approximately minimize F_S to train a model. We use $A(S)$ to denote the model produced by applying A to S, and are interested in its relative performance w.r.t. the best model $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$, which is quantified by the excess risk defined by $\mathbb{E}[F(A(S)) - F(\mathbf{w}^*)]$. A powerful method to study the excess risk is to decompose it into two components (Bousquet & Bottou, 2008):

$$\mathbb{E}[F(A(S)) - F(\mathbf{w}^*)] = \underbrace{\mathbb{E}[F(A(S)) - F_S(A(S))]}_{\text{Generalization Error}} + \underbrace{\mathbb{E}[F_S(A(S)) - F_S(\mathbf{w}^*)]}_{\text{Optimization Error}}, \tag{3.1}$$

where the expectation is taken over the randomness of the training set S and any randomness within the algorithm itself. Here we use the identity $\mathbb{E}[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$. We refer to $\mathbb{E}[F(A(S)) - F_S(A(S))]$ as the generalization gap, which shows the cost we suffer when we generalize the behavior from training to testing. A small generalization gap indicates that the model does not overfit the training data and its performance is likely to be representative of its true performance. We refer to $\mathbb{E}[F_S(A(S)) - F_S(\mathbf{w}^*)]$ as the optimization error, which measures the gap between the estimated model and the true optimal model on empirical risk.

We introduce the following necessary definitions for our analysis. Let $\|\cdot\|_2$ denote the Euclidean norm.

Definition 1. Let $g : \mathcal{W} \mapsto \mathbb{R}$, $G, L > 0$ and $\mu \geq 0$. We denote the gradient of g by ∇g.

1. A function $g(\mathbf{w})$ is μ-strongly convex for some $\mu > 0$ if it satisfies:

$$g(\mathbf{w}_1) \geq g(\mathbf{w}_2) + \langle \nabla g(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle + \frac{\mu}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}.$$

   A function $g(\mathbf{w})$ is convex if it is μ-strongly convex with $\mu = 0$.

2. A function $g(\mathbf{w})$ is G-Lipschitz continuous if the function value is bounded in its change:

$$|g(\mathbf{w}_1) - g(\mathbf{w}_2)| \leq G\|\mathbf{w}_1 - \mathbf{w}_2\|_2, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}.$$

3. A differentiable function $g(\mathbf{w})$ is L-smooth if its gradient is Lipschitz continuous with the constant L:

$$\|\nabla g(\mathbf{w}_1) - \nabla g(\mathbf{w}_2)\|_2 \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|_2, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}.$$

4 ALGORITHMIC STABILITY

To control the generalization gap, we analyze the stability of our learning algorithm. We say an algorithm is on-average stable if its output model does not change significantly when a single data point in the training set is modified. Let A be a learning algorithm that takes a dataset S and outputs a model $A(S)$. We denote $S \sim S'$ if S and S' differ by at most one data point. Specifically, we let $S^{(i)}$ be a dataset identical to S except that the i-th data z_i is replaced with a new point z_i', drawn from the same distribution $\mathcal{D}$. That is, $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$.

Definition 2 (Uniform Stability)**.** An algorithm A has uniform stability ϵ if

$$\sup_{z\in\mathcal{Z}}\sup_{S\sim S'}\mathbb{E}\left[|f(A(S);z)-f(A(S');z)|\right]\leq\epsilon.$$

Definition 3 (On-Average Model Stability (Lei & Ying, 2020))**.** We say a randomized optimizer A is ℓ_1 on-average model ϵ-stable if

$$\mathbb{E}_{S,S',A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|A(S)-A(S^{(i)})\|_2\Big]\leq\epsilon.$$

We say A is ℓ_2 on-average model ϵ-stable if

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

The following lemma provides a connection between the generalization gap and on-average model stability.

Lemma 1 ((Lei & Ying, 2020))**.** *Let S, S' and $S^{(i)}$ be constructed as in Definition 2, and let $\gamma>0$.*

(a) Suppose for any z, the function $\mathbf{w}\mapsto f(\mathbf{w};z)$ is convex. If A is ℓ_1 on-average model ϵ-stable and $\sup_z\|\nabla f(A(S);z)\|_2\leq G$ for any S, then $|\mathbb{E}_{S,A}[F_S(A(S))-F(A(S))]|\leq G\epsilon$.

(b) Suppose for any z, the function $\mathbf{w}\mapsto f(\mathbf{w};z)$ is nonnegative and L-smooth. If A is ℓ_2 on-average model ϵ-stable, then the following inequality holds

$$\mathbb{E}_{S,A}[F(A(S))-F_S(A(S))]\leq\frac{L}{\gamma}\mathbb{E}_{S,A}[F_S(A(S))]+\frac{L+\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,S',A}[\|A(S^{(i)})-A(S)\|_2^2].$$

4.1 Lookahead Optimizer

The Lookahead optimizer (Zhang et al., 2019), detailed in Algorithm 1, employs a two-loop structure: an inner loop to update fast weights, and an outer loop to update slow weights. In the inner loop, a standard optimizer $\mathcal{A}$ (e.g. SGD or Adam) starts from the previous slow weight model $\mathbf{w}_{t-1}$ and updates fast weights $\mathbf{v}_{k,t}$ with appropriate inner step sizes $\eta_{\tau,t}$ for k iterations. In the t-th iteration of the outer loop, the fast weight model $\mathbf{v}_{k,t}$ is then used to update the slow weight model via a linear interpolation

$$\mathbf{w}_t=(1-\alpha)\mathbf{w}_{t-1}+\alpha\mathbf{v}_{k,t}\tag{4.1}$$

where $\alpha\in(0,1)$ is the outer step size.

Algorithm 1 Lookahead Optimizer

1: **Inputs:** Data set $\mathcal{S}$, initial model $\mathbf{w}_0$, base optimizer $\mathcal{A}$, fast-weight step number k and learning rates $\{\{\eta_{\tau,t}\}_{\tau=0}^{k-1}\}_{t=1}^{T}$, slow-weight step number T and learning rate $\alpha\in(0,1)$.
2: **for** $t=1,2,\ldots,T$ **do**
3: $\quad\mathbf{v}_{0,t}=\mathbf{w}_{t-1}$
4: $\quad$**for** $\tau=1,2,\ldots,k$ **do**
5: $\quad\quad\mathbf{v}_{\tau,t}=\mathcal{A}(\mathbf{v}_{\tau-1,t},\eta_{\tau-1,t},\mathcal{S})$
6: $\quad$**end for**
7: $\quad\mathbf{w}_t=(1-\alpha)\mathbf{w}_{t-1}+\alpha\mathbf{v}_{k,t}$
8: **end for**
9: **Outputs:** Slow model $\mathbf{w}_T$

We use minibatch SGD as the standard optimizer $\mathcal{A}$, which is widely used in deep learning. The inner loop is then reformulated as in Algorithm 2. At the τ'th iteration, SGD collects a minibatch $\mathcal{B}_{\tau,t}$ by randomly drawing $|\mathcal{B}_{\tau,t}|$ data points from $\mathcal{S}$ independently, where $|\cdot|$ denotes the cardinality. Then it updates $\{\mathbf{v}_{\tau,t}\}_{\tau=1}^{k}$ by

$$\mathbf{v}_{\tau,t}=\mathbf{v}_{\tau-1,t}-\frac{\eta_{\tau-1,t}}{|\mathcal{B}_{\tau,t}|}\sum_{z\in\mathcal{B}_{\tau,t}}\nabla f(\mathbf{v}_{\tau-1,t};z),$$

where $\eta_{\tau,t}$ is a positive step size.

Algorithm 2 Stochastic Gradient Descent (SGD)

1: **Inputs:** Data set $\mathcal{S}$, learning rates $\{\eta_{\tau,t}\}_{\tau=0}^{k-1}$, initial model $\mathbf{v}_{0,t}$,
2: **for** $\tau = 1, 2, \ldots, k$ **do**
3: $\quad \mathbf{v}_{\tau,t} = \mathbf{v}_{\tau-1,t} - \frac{\eta_{\tau-1,t}}{|\mathcal{B}_{\tau,t}|} \sum_{z \in \mathcal{B}_{\tau,t}} \nabla f(\mathbf{v}_{\tau-1,t}; z)$
4: **end for**
5: **Outputs:** Fast model $\mathbf{v}_{k,t}$

5 Generalization Analysis of Lookahead Algorithm

In this section, we discuss the stability performance of Lookahead on convex and strongly convex problems. While previous work has shown that Lookahead achieves lower excess risk error compared to its vanilla inner optimizer when choosing $\mathcal{A}$ as SGD (Zhou et al., 2021), existing analysis of its generalization and optimization error suffer from two key limitations. First, they hinge on a restrictive Lipschitzness condition on the loss function. Second, they cannot imply optimistic rates to show the benefit of low-noise condition to get fast rates. In the following sections, we will analyze the stability bound of Lookahead via the ℓ_2 on-average model stability. This approach notably allows us to derive generalization bounds for Lookahead without requiring the Lipschitzness condition (Lei et al., 2025). Furthermore, by carefully selecting the algorithm's hyperparameters, we establish optimal excess risk bounds.

5.1 Convex Case

We first investigate stability bounds of Lookahead under convex condition, where Eq. (5.1) considers the ℓ_1 on-average stability and Eq. (5.2) considers the ℓ_2 on-average stability. The proof will be given in Appendix A.1.

Theorem 2 (Stability Bound of Lookahead: Convex Case)**.** *Suppose the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex, nonnegative and L-smooth for all $z \in \mathcal{Z}$. Let $\{\mathbf{v}_{\tau,t}\}$ and $\{\mathbf{w}_t\}$, $\{\mathbf{v}_{\tau,t}^{(i)}\}$ and $\{\mathbf{w}_t^{(i)}\}$ be produced based on S and $S^{(i)}$ respectively with $\eta_{\tau,t} \leq \frac{1}{L}$. We have*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \leq \alpha \sum_{h=1}^{t+1}\sum_{j=1}^{k-1} \frac{2\eta_{j,h}\sqrt{2L\mathbb{E}\left[F_S\left(\mathbf{v}_{j,h}\right)\right]}}{n} \tag{5.1}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq \left(\frac{16\alpha^2 L}{nb} + \frac{16\alpha^2 L(t+1)k}{n^2}\right)\sum_{h=1}^{t+1}\sum_{j=1}^{k-1}\eta_{j,h}^2\mathbb{E}\left[F_S\left(\mathbf{v}_{j,h}\right)\right]. \tag{5.2}$$

Remark 1 (Comparison with existing stability bounds for Lookahead)**.** For L-smooth, G-Lipschitz and convex problems, a similar ℓ_1-stability bound was derived in (Zhou et al., 2021) as shown below

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \leq \frac{2\alpha\eta GkT}{n}.$$

This bound grows linearly with kT, is independent of the mini-batch size b, and involves the global Lipschitz constant G. Our analysis removes the global G-Lipschitz requirement and thus avoids the G factor. A notable feature of our bound is its dependence on the empirical risk, $\mathbb{E}[F_S(\mathbf{v}_{j,h})]$, rather than the global Lipschitz constant G in (Zhou et al., 2021). Since the objective of the inner-loop optimizer is precisely to minimize F_S, we expect this term to decrease as training progresses. Consequently, our stability bounds become progressively tighter throughout the optimization process (Kuzborskij & Lampert, 2018; Lei & Ying, 2020). Furthermore, the bound in Eq. (5.2) provides clear intuition about the role of Lookahead's hyperparameters:

- **Batch Size (b):** The term $1/nb$ shows that increasing the minibatch size improves stability. As a comparison, the stability analysis in (Zhou et al., 2021) does not show the effect of the batch size since their stability bound is independent of b.

- **Inner Loop Iteration Number (k):** The bound increases with k, suggesting that running the inner loop for too many steps can degrade stability, likely due to the fast weights overfitting to the training set S.
- **Outer Loop Step Size (α):** Stability is proportional to α. A smaller α dampens the influence of the potentially unstable fast weights, leading to a more stable trajectory for the slow weights. This shows a clear advantage of the Lookahead mechanism in improving the stability and generalization.

We get the generalization bound via plugging the stability bounds in Theorem 2 into Lemma 1. Together with the optimization bound in Lemma 9, we have the following excess risk bound. The proof is given in Appendix A.2. We denote $A \lesssim B$ if there exists a universal constant $C > 0$ such that $A \leq CB$. We denote $A \gtrsim B$ if there exists a universal constant C such that $A \geq CB$. We denote $A \asymp B$ if $A \lesssim B$ and $A \gtrsim B$.

Theorem 3 (Excess Risk Bound of Lookahead: Convex Case)**.** *Let the assumptions of Theorem 2 hold and $R = Tk$. Then for $\bar{\mathbf{v}}_R = \frac{1}{Tk}\sum_{t=1}^{T}\sum_{\tau=0}^{k-1}\mathbf{v}_{\tau,t}$ and $\gamma > 0$, we have*

$$\mathbb{E}\left[F(\overline{\mathbf{v}}_R)\right] - F(\mathbf{w}^*) \lesssim \frac{L\eta F(\mathbf{w}^*)}{b} + \frac{1}{\alpha\eta R} + \frac{F(\mathbf{w}^*) + L\eta/b + 1/(\alpha\eta R)}{\gamma} + L(L+\gamma)\alpha^2\eta^2\left(\frac{1}{nb} + \frac{R}{n^2}\right)\left(RF(\mathbf{w}^*) + \frac{RL\eta}{b} + \frac{1}{\alpha\eta}\right). \quad (5.3)$$

Since there are terms directly proportional to $F(\mathbf{w}^*)$, the excess risk bound will be tighter when the optimal risk $F(\mathbf{w}^*)$ is small, which is common in many machine learning problems where a model can fit the data well. Excess risk bounds with this feature are called optimistic bounds (Srebro et al., 2010). The terms involving $F(\mathbf{w}^*)$ are directly related to gradient noise, as the variance of stochastic gradients can often be bounded by the function's value at the optimum.

Remark 2 (Comparison with Minibatch SGD)**.** The excess risk bound for Lookahead in Theorem 3 shares a fundamental structure with the bound for Minibatch SGD as in (Lei et al., 2025). Both are optimistic bounds that explicitly depend on the optimal risk. This similarity is expected, as both analyses aim to control generalization gap by plugging stability bounds into Lemma 3.1, then adding optimization error terms. Although the structure is similar, the specific coefficients and dependencies on parameters such as α and the structure of the variance term differ due to the unique dynamics of the Lookahead optimizer compared to standard SGD.

We now develop an explicit excess risk bound for Lookahead by choosing step sizes and number of iterations. The proof is given in Appendix A.2.

Corollary 4. *Let the assumptions of Theorem 3 hold.*

1. *If $F(\mathbf{w}^*) \geq 1/n$, we can take $\eta = \frac{b}{\sqrt{nF(\mathbf{w}^*)}}$, $R \asymp \frac{n}{b}$, $\gamma = \sqrt{nF(\mathbf{w}^*)} \geq 1$, and $b \leq \sqrt{nF(\mathbf{w}^*)}/(2L)$ to derive $\mathbb{E}[F(\bar{\mathbf{v}}_R)] - F(\mathbf{w}^*) \lesssim \frac{LF(\mathbf{w}^*)^{1/2}}{\sqrt{n}} + \frac{L^2}{n}$.*
2. *If $F(\mathbf{w}^*) < 1/n$, we can take $\eta = \frac{1}{2L}$, $R \asymp n$, and $\gamma = 1$ to derive $\mathbb{E}[F(\bar{\mathbf{v}}_R)] - F(\mathbf{w}^*) \lesssim \frac{L}{n} + F(\mathbf{w}^*)$.*

Remark 3. Corollary 4 distinguishes between two key regimes based on the magnitude of the optimal risk $F(\mathbf{w}^*)$ relative to the sample size n.

1. $F(\mathbf{w}^*) \geq 1/n$: Our analysis shows that the algorithm achieves an excess risk bound of $O(\frac{1}{\sqrt{n}})$. Crucially, the number of required iterations R is on the order of n/b, demonstrating a linear speedup (Cotter et al., 2011a). This means that by increasing the minibatch size b, one can use a proportionally larger learning rate η and achieve the same error bound with fewer iterations. This acceleration is a direct benefit of variance reduction from larger batch sizes.
2. $F(\mathbf{w}^*) < 1/n$: Now the required number of iterations R scales with n, irrespective of the batch size b. In this case, the linear speedup vanishes. The optimal learning rate becomes

constant, and increasing the batch size does not reduce the number of iterations needed to reach the desired error threshold. This suggests a small stochastic gradient noise, which means variance is no longer the main limitation of the learning process.

Remark 4 (Comparison with Existing Excess Risk Bounds with Lookahead)**.** The work (Zhou et al., 2021) gave the following excess risk bound for Lookahead under convexity and G-Lipschitz continuity assumption

$$\mathbb{E}[F(\bar{\mathbf{v}}_R)] - F(\mathbf{w}^*) \leq \frac{1}{2\alpha\eta kT}\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}^*\|^2] + \frac{\eta G^2}{2} + \frac{\alpha\eta G^2 kT}{n}.$$

By setting $\eta \asymp 1/\sqrt{n}$ and choosing $\alpha Tk \asymp n$, all three terms can be made to be of the order $O(1/\sqrt{n})$. This leads to an optimized excess risk bound of order $G^2/\sqrt{n}$, which is standard for stochastic convex optimization under a Lipschitz assumption. However, it is not adaptive and can be suboptimal in many practical scenarios. In the case of $F(\mathbf{w}^*) \geq 1/n$, our bound is of order $\frac{L\sqrt{F(\mathbf{w}^*)}}{\sqrt{n}}$. As the optimal risk $F(\mathbf{w}^*)$ decreases, our bound becomes tighter. For problems where $L\sqrt{F(\mathbf{w}^*)} \ll G^2$, our bound is substantially sharper than the generic $O(G^2/\sqrt{n})$ rate. In the case of $F(\mathbf{w}^*) < 1/n$, our analysis reveals a much faster convergence rate of $\lesssim \frac{L}{n}$. This is a linear convergence rate with respect to the sample size n. Achieving an $O(1/n)$ rate is a major acceleration compared to the standard $O(1/\sqrt{n})$ rate. It shows that Lookahead can effectively leverage low-noise conditions to converge significantly faster, a behavior that the existing bound fails to capture. Furthermore, our analysis shows a linear speedup on the batch size, while the discussions in (Zhou et al., 2021) do not show the benefit of considering minibatch in both generalization and optimization.

5.2 Strongly Convex Case

We now consider strongly convex problems. The following theorem provides stability bounds for Lookahead. The proof is given in Appendix A.3.

Theorem 5 (Stability Bound of Lookahead: Strongly Convex Case)**.** *Suppose the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is μ-strongly convex, nonnegative and L-smooth for all $z \in \mathcal{Z}$. Let $\{\mathbf{v}_{\tau,t}\}$ and $\{\mathbf{w}_t\}$, $\{\mathbf{v}_{\tau,t}^{(i)}\}$ and $\{\mathbf{w}_t^{(i)}\}$ be produced based on S and $S^{(i)}$ respectively with $\frac{2\ln 2}{k\mu} \leq \eta_{\tau,t} \leq \frac{1}{L}$. We have*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \leq \frac{2\alpha\sqrt{2L}}{n}\sum_{t'=1}^{t+1}(1-\frac{\alpha}{2})^{t+1-t'}\sum_{j=0}^{k-1}\eta_{j,t'}\sqrt{\mathbb{E}\left[F_S\left(\mathbf{v}_{j,t'}\right)\right]}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big) \tag{5.4}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq \sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\Big(\frac{16\alpha^2\eta_{j,t'}^2}{nb} + \frac{32\left(t+1\right)\alpha^2\eta_{j,t'}}{n^2\mu}\Big)\mathbb{E}\left[F_S\left(\mathbf{v}_{j,t'}\right)\right]\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)^2. \tag{5.5}$$

Eq. (5.4) provides an ℓ_1-on-average stability bound. A key feature of this bound is its dependence on the empirical risk, $\sqrt{\mathbb{E}[F_S(\mathbf{v}_{j,t'})]}$. This indicates that the stability of the Lookahead algorithm improves as it finds iterates with smaller empirical risks. Eq. (5.5) provides an ℓ_2-on-average stability bound. This bound explicitly shows the benefit of minibatching. The term $\frac{16\alpha^2\eta_{j,t'}}{nb}$ demonstrates that increasing the batch size b directly improves the stability bound by reducing the variance introduced by the stochastic gradients. This is a crucial property for large-scale learning, confirming that larger batches contribute to a more stable training process for the Lookahead algorithm.

Theorem 6 (Excess Risk Bound of Lookahead: Strongly Convex Case)**.** *Let assumptions in Theorem 5 hold and let $\eta = \frac{b\mu}{2L^2(b+1)}$, $k = \frac{2L}{\alpha\mu}$ and $T \asymp \log(\mu n)$, we have*

$$\mathbb{E}[F(\mathbf{w}_T)] - F(\mathbf{w}^*) \lesssim \frac{1}{n\mu} + \big(\frac{1}{nL}+1\big)\mathbb{E}[F_S(\mathbf{w}_S)] + \big(\frac{1}{n^2}+\frac{L}{n}\big)\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_S\|^2]. \tag{5.6}$$

Remark 5 (Comparison with Existing Excess Risk Bound with Lookahead)**.** Compared with the existing Lookahead bound in the work (Zhou et al., 2021), which yields a sum of terms of order

$O(1/(\lambda^2((t+1)k)^{2\alpha})) + O(G/(n\lambda))$ and therefore requires tk to scale polynomially with n to reach the $O(1/n)$ regime, our Theorem 6 delivers a fast-rate excess risk of order $1/(n\mu)$ with only $T \asymp \log(\mu n)$ iterations. Moreover, our bound is adaptive: it tightens with the data through $(1/(nL)+1)\mathbb{E}[F_S(\mathbf{w}_S)]$ and through $(1/n^2 + L/n)\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_S\|^2]$, becoming much smaller under interpolation, which is not captured by the existing result. Finally, the stepsize η scales with the minibatch b, implying linear speedup in b, while prior analyses do not show such minibatch gains.

6 Conclusion

In this work, we investigate the stability and generalization properties of the Lookahead optimizer, a widely used algorithm for large-scale machine learning problems. While many discussions focus on its optimization benefits, we provide a rigorous analysis from the perspective of statistical learning theory. We develop on-average stability bounds for both convex and strongly convex problems, and we show how stability can be improved by small training errors, leading to optimistic bounds that depend on the empirical risk rather than a restrictive, global Lipschitz constant.

Our stability analysis implies optimal excess population risk bounds for both settings. Specifically, we demonstrate that Lookahead achieves the standard $O(1/\sqrt{n})$ rate for convex problems and the optimal $O(1/(n\mu))$ rate for strongly convex problems. A key finding is the adaptivity of Lookahead in the convex case, which achieves its rate without prior knowledge of the optimal risk $F(\mathbf{w}^*)$, a practical advantage over standard Minibatch SGD.

There are several limitations to our current work which open avenues for future research. A primary limitation is that our analysis is confined to convex and strongly convex loss functions. Given the prevalence of non-convex optimization in modern deep learning, extending our stability analysis to the non-convex setting is a crucial next step. Furthermore, while we establish the optimal statistical rate for the strongly convex case, our analysis does not demonstrate a linear speedup with respect to the batch size, a property observed in Minibatch SGD. Investigating whether different hyperparameter schedules could unlock such a speedup for Lookahead would be of significant interest. We plan to address these limitations in our future research.

References

Steve A Adeshina and Adeyinka P Adedigba. Bag of tricks for improving deep learning performance on multimodal image classification. *Bioengineering*, 9:312, 2022.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pp. 161–168, 2008.

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

Ha-Eun Byun, Boeun Kim, and Jay H Lee. Multi-step lookahead bayesian optimization with active learning using reinforcement learning and its application to data-driven batch-to-batch optimization. *Computers & Chemical Engineering*, 167:107987, 2022.

Cong Chen, Huili Zhang, and Yinfeng Xu. Online machine minimization with lookahead. *Journal of combinatorial optimization*, 43:1149–1172, 2022a.

Cong Chen, Huili Zhang, and Yinfeng Xu. Online machine minimization with lookahead. *Journal of Combinatorial Optimization*, 43(5):1149–1172, 2022b.

Hailong Chen, Mei Du, Yingyu Zhang, and Chang Yang. Research on disease prediction method based on r-lookahead-lstm. *Computational Intelligence and Neuroscience*, 2022:8431912, 2022c.

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

Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems*, volume 24, 2011a.

Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems*, 24, 2011b.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *The Journal of Machine Learning Research*, 13:165–202, 2012.

Qichen Deng and Bruno F. Santos. Lookahead approximate dynamic programming for stochastic aircraft maintenance check scheduling optimization. *European Journal of Operational Research*, 299:814–833, 2022.

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.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2820–2829, 2018.

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

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.

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 661–670, 2014.

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

Sayan Nag. Lookahead optimizer improves the performance of convolutional autoencoders for reconstruction of natural images. *arXiv preprint arXiv:2012.05694*, 2020.

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

Denys Pushkin and Luis Barba. Multilayer lookahead: a nested version of lookahead. *arXiv preprint arXiv:2110.14254*, 2021.

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*, 34:8609–8621, 2021.

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

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

Ohad Shamir and Nathan Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 850–857, 2014.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pp. 2199–2207, 2010.

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

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.

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, 2025a.

Xu Wang, Bo Zhang, Xintong Du, Huailin Chen, Tianwen Zhu, and Chundu Wu. Self-adjusting look-ahead distance of precision path tracking for high-clearance sprayers in field navigation. *Agronomy*, 15:1433, 2025b.

Anna Winnicki, Joseph Lubars, Michael Livesay, and R Srikant. The role of lookahead and approximate policy evaluation in reinforcement learning with linear value function approximation. *Operations Research*, 73: 139–156, 2025.

Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*, 2020.

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*, 2024.

Jiangtao Zhang, Shunyu Liu, Jie Song, Tongtian Zhu, Zhengqi Xu, and Mingli Song. Lookaround optimizer: k steps around, 1 step average. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yunfeng Zhang, Shukai Li, Yin Yuan, and Lixing Yang. Multi-step look ahead deep reinforcement learning approach for automatic train regulation of urban rail transit lines with energy-saving. *Engineering Applications of Artificial Intelligence*, 145:110181, 2025.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and E Weinan. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 2020, 2020.

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

Xuan Zuo, Hui-Yan Li, Shan Gao, Pu Zhang, and Wan-Ru Du. Nala: a nesterov accelerated look-ahead optimizer for deep learning. *PeerJ Computer Science*, 10:e2167, 2024.

A Proof of Results in Section 5

A.1 Proof of Theorem 2

Our proof of Theorem 2 relies on the following two lemmas. Lemma 7 shows the self-bounding property for nonnegative and smooth functions, meaning that the norm of gradients can be bounded by function values. Lemma 8 establishes the co-coercivity of smooth and convex functions, as well as the non-expansiveness of the gradient operator $\mathbf{w} \mapsto \mathbf{w} - \eta \nabla f(\mathbf{w}; z)$.

Lemma 7 (Self-Bounding Property (Srebro et al., 2010))**.** *Assume for all z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative and L-smooth. Then*

$$\|\nabla f(\mathbf{w}; z)\|_2^2 \leq 2Lf(\mathbf{w}; z).$$

Lemma 8 ((Hardt et al., 2016))**.** *Assume for all $z \in Z$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex and L-smooth. Then for $\eta \leq 2/L$ we have*

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

Furthermore, if $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is μ-strongly convex and $\eta \leq 1/L$ then

$$\begin{aligned}
\|(\mathbf{w} - \eta \nabla f(\mathbf{w}; z)) - (\mathbf{w}' - \eta \nabla f(\mathbf{w}'; z))\|_2 &\leq (1 - \eta\mu/2)\|\mathbf{w} - \mathbf{w}'\|_2, \\
\|(\mathbf{w} - \eta \nabla f(\mathbf{w}; z)) - (\mathbf{w}' - \eta \nabla f(\mathbf{w}'; z))\|_2^2 &\leq (1 - \eta\mu)\|\mathbf{w} - \mathbf{w}'\|_2^2.
\end{aligned}$$

We can now prove Theorem 2. For simplicity, we define $J_{\tau,t} = \{i_{\tau,t}^{(1)}, \ldots, i_{\tau,t}^{(b)}\}$, where $i_{\tau,t}^{(j)} \sim \text{Unif}([n])$ is the j-th index sampled to compute a stochastic gradient for minibatch SGD, i.e., $\mathcal{B}_{\tau,t} = \{z_{i_{\tau,t}^{(1)}}, \ldots, z_{i_{\tau,t}^{(b)}}\}$.

Proof. To begin with, define

$$A_{\tau,t}^{(m)} = |\{j : i_{\tau,t}^{(j)} = m\}|,$$

that is, $A_{\tau,t}^{(m)}$ represents the number of indices equal to m in the batch of t-th outer loop iteration, and τ-th inner loop iteration. Then we can reformulate the Lookahead update as

$$\begin{aligned}
\mathbf{w}_{t+1} &= (1-\alpha)\mathbf{w}_t + \alpha \mathbf{v}_{k,t+1} \\
&= (1-\alpha)\mathbf{w}_t + \alpha\Big(\mathbf{v}_{k-1,t+1} - \frac{\eta_{k-1,t+1}}{b}\sum_{m=1}^{n} A_{k-1,t+1}^{(m)} \nabla f(\mathbf{v}_{k-1,t+1}; z_m)\Big), \\
\mathbf{w}_{t+1}^{(i)} &= (1-\alpha)\mathbf{w}_t^{(i)} + \alpha\Big(\mathbf{v}_{k-1,t+1}^{(i)} - \frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n} A_{k-1,t+1}^{(m)} \nabla f(\mathbf{v}_{k-1,t+1}^{(i)}; z_m) \\
&\quad - \frac{A_{k,t+1}^{(i)}\eta_{k-1,t+1}}{b}\nabla f(\mathbf{v}_{k-1,t+1}^{(i)}; z_i')\Big),
\end{aligned} \tag{A.1}$$

from which we know

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2 &\le \big(1-\alpha\big)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\alpha\|\mathbf{v}_{k,t+1}-\mathbf{v}_{k,t+1}^{(i)}\|_2\\ &\le (1-\alpha)\,\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\alpha\Big\|\mathbf{v}_{k-1,t+1}-\frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n}A_{k-1,t+1}^{(m)}\nabla f(\mathbf{v}_{k-1,t+1};z_m)\\ &\quad-\frac{A_{k-1,t+1}^{(i)}\eta_{k-1,t+1}}{b}\nabla f\left(\mathbf{v}_{k-1,t+1};z_i\right)-\mathbf{v}_{k-1,t+1}^{(i)}+\frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n}A_{k-1,t+1}^{(m)}\nabla f(\mathbf{v}_{k-1,t+1}^{(i)};z_m)\\ &\quad+\frac{A_{k-1,t+1}^{(i)}\eta_{k-1,t+1}}{b}\nabla f(\mathbf{v}_{k-1,t+1}^{(i)};z_i')\Big\|_2.\end{aligned}$$

Define $\mathfrak{C}_{k-1,t+1}^{(i)} = \|\nabla f\left(\mathbf{v}_{k-1,t+1};z_i\right)-\nabla f(\mathbf{v}_{k-1,t+1}^{(i)};z_i')\|_2$. By assumption, f is L-smooth and $\sum_{m:m\neq i}^{n}A_{k-1,t+1}^{(m)}\le b$, from which we know $\mathbf{v}\mapsto\frac{1}{b}\sum_{m:m\neq i}^{n}A_{k-1,t+1}^{(m)}f\left(\mathbf{v};z_m\right)$ is L-smooth. Since by assumption $\eta_{k-1,t+1}\le\frac{1}{L}$, by Lemma 8 we have

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\\ &\le(1-\alpha)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\frac{\alpha A_{k-1,t+1}^{(i)}\eta_{k-1,t+1}}{b}\|\nabla f\left(\mathbf{v}_{k-1,t+1};z_i\right)-\nabla f(\mathbf{v}_{k-1,t+1}^{(i)};z_i')\|_2\\ &+\alpha\Big\|\mathbf{v}_{k-1,t+1}-\frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n}A_{k-1,t+1}^{(m)}\nabla f(\mathbf{v}_{k-1,t+1};z_m)-\Big(\mathbf{v}_{k-1,t+1}^{(i)}-\frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n}A_{k-1,t+1}^{(m)}\nabla f(\mathbf{v}_{k-1,t+1}^{(i)};z_m)\Big)\Big\|_2\\ &\le(1-\alpha)\,\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\frac{\alpha\eta_{k-1,t+1}A_{k-1,t+1}^{(i)}\mathfrak{C}_{k-1,t+1}^{(i)}}{b}+\alpha\|\mathbf{v}_{k-1,t+1}-\mathbf{v}_{k-1,t+1}^{(i)}\|_2.\end{aligned}\tag{A.2}$$

Note the above inequality actually shows a recurrent relationship on $\|\mathbf{v}_{k,t+1}-\mathbf{v}_{k,t+1}^{(i)}\|_2$ and $\|\mathbf{v}_{k-1,t+1}-\mathbf{v}_{k-1,t+1}^{(i)}\|_2$. By iteration on inner-loop, we have

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2&\le(1-\alpha)\,\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\frac{\alpha}{b}\sum_{j=0}^{k-1}\eta_{j,t+1}A_{j,t+1}^{(i)}\mathfrak{C}_{j,t+1}^{(i)}+\alpha\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\\ &=\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\frac{\alpha}{b}\sum_{j=0}^{k-1}\eta_{j,t+1}A_{j,t+1}^{(i)}\mathfrak{C}_{j,t+1}^{(i)},\end{aligned}$$

where we have used that $\mathbf{v}_{0,t+1}=\mathbf{w}_t$. By iteration on outer-loop, we have

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\le\frac{\alpha}{b}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}A_{j,h}^{(i)}\mathfrak{C}_{j,h}^{(i)}.\tag{A.3}$$

By definition of $A_{k,t}^{(m)}$, it is a random variable following the binomial distribution $B(b,\frac{1}{n})$, it then follows that

$$\mathbb{E}\big[A_{k,t}^{(m)}\big]=\frac{b}{n},\quad \mathrm{Var}\left(A_{k,m}^{(t)}\right)=\frac{b}{n}\big(1-\frac{1}{n}\big)\le\frac{b}{n}.\tag{A.4}$$

Furthermore, by Lemma 7, we know

$$\mathfrak{C}_{j,h}^{(i)}\le\|\nabla f(\mathbf{v}_{j,h};z_i)\|_2+\|\nabla f(\mathbf{v}_{j,h}^{(i)};z_i')\|_2\le\sqrt{2Lf\left(\mathbf{v}_{j,h};z_i\right)}+\sqrt{2Lf(\mathbf{v}_{j,h}^{(i)};z_i')}.\tag{A.5}$$

Since (x_i,y_i) and (x_i',y_i') are symmetric, we know $\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i\right)\right]=\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i'\right)\right]$. This, together with Eq (A.5), further implies that

$$\mathbb{E}\big[\mathfrak{C}_{j,h}^{(i)}\big]\le2\mathbb{E}\Big[\sqrt{2Lf\big(\mathbf{v}_{j,h};z_i\big)}\Big].\tag{A.6}$$

By combining (A.3) and (A.4), we have

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big]&\le\frac{\alpha}{b}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathbb{E}\big[A_{j,h}^{(i)}\mathfrak{C}_{j,h}^{(i)}\big]=\frac{\alpha}{b}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathbb{E}\big[\mathbb{E}_{J_{j,h}}\big[A_{j,h}^{(i)}\big]\mathfrak{C}_{j,h}^{(i)}\big]\\ &=\frac{\alpha}{n}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathbb{E}\big[\mathfrak{C}_{j,h}^{(i)}\big]\le\frac{2\alpha}{n}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathbb{E}\Big[\sqrt{2Lf(\mathbf{v}_{j,h};z_i)}\Big],\end{aligned}\tag{A.7}$$

where we used (A.6) in the last inequality. By the concavity of $x \mapsto \sqrt{x}$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big] &\le \frac{2\alpha}{n}\sum_{i=1}^{n}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\frac{\eta_{j,h}}{n}\mathbb{E}\Big[\sqrt{2Lf(\mathbf{v}_{j,h};z_i)}\Big]\\ &\le \alpha\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\frac{2\eta_{j,h}}{n}\sqrt{\frac{2L}{n}\sum_{i=1}^{n}\mathbb{E}\left[f(\mathbf{v}_{j,h};z_i)\right]}\\ &= \alpha\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\frac{2\eta_{j,h}\sqrt{2L\mathbb{E}\left[F_S(\mathbf{v}_{j,h})\right]}}{n}. \end{aligned} \tag{A.8}$$

This established the stated ℓ_1-stability (5.1).
To study the ℓ_2-stability, we apply the following expectation-variance decomposition to Eq. (A.3).

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2 \le \frac{\alpha}{b}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\big(A_{j,h}^{(i)}-\frac{b}{n}\big)\mathfrak{e}_{j,h}^{(i)} + \frac{\alpha}{n}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathfrak{e}_{j,h}^{(i)}. \tag{A.9}$$

Taking square on both sides, then applying expectation with respect to S and $J_{k,t}$ for $t\in[T]$ and $k\in[k]$, we have

$$\begin{aligned}&\mathbb{E}\Big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\Big]\\ &\le \frac{2\alpha^2}{b^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\Big(A_{j,h}^{(i)}-\frac{b}{n}\Big)\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big] + \frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big]\\ &= \frac{2\alpha^2}{b^2}\mathbb{E}\Big[\sum_{h,h'=1}^{t+1}\sum_{j,j'=0}^{k-1}\eta_{j,h}\eta_{j',h'}\Big(A_{j,h}^{(i)}-\frac{b}{n}\Big)\Big(A_{j',h'}^{(i)}-\frac{b}{n}\Big)\mathfrak{e}_{j,h}^{(i)}\mathfrak{e}_{j',h'}^{(i)}\Big] + \frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big], \end{aligned} \tag{A.10}$$

where we have used $(a+b)^2\le 2(a^2+b^2)$. Note that if $(h,j)\neq(h',j')$, then (we can assume $h<h'$, $j<j'$ without loss of generality)

$$\begin{aligned}\mathbb{E}\Big[\Big(A_{j,h}^{(i)}-\frac{b}{n}\Big)\Big(A_{j',h'}^{(i)}-\frac{b}{n}\Big)\mathfrak{e}_{j,h}^{(i)}\mathfrak{e}_{j',h'}^{(i)}\Big] &= \mathbb{E}\mathbb{E}_{J_{j',h'}}\Big[\Big(A_{j,h}^{(i)}-\frac{b}{n}\Big)\Big(A_{j',h'}^{(i)}-\frac{b}{n}\Big)\mathfrak{e}_{j,h}^{(i)}\mathfrak{e}_{j',h'}^{(i)}\Big]\\ &= \mathbb{E}\Big[\Big(A_{j,h}^{(i)}-\frac{b}{n}\Big)\mathbb{E}_{J_{j',h'}}\Big[A_{j',h'}^{(i)}-\frac{b}{n}\Big]\mathfrak{e}_{j,h}^{(i)}\mathfrak{e}_{j',h'}^{(i)}\Big] = 0, \end{aligned} \tag{A.11}$$

where we notice $A_{j,h}^{(i)}$, $\mathfrak{e}_{j,h}^{(i)}$, and $\mathfrak{e}_{j',h'}^{(i)}$ are independent of $J_{j',h'}$. It then follows that

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big] &\le \frac{2\alpha^2}{b^2}\mathbb{E}\Big[\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\Big(A_{j,h}^{(i)}-\frac{b}{n}\Big)^2\Big(\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big] + \frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big]\\ &= \frac{2\alpha^2}{b^2}\mathbb{E}\Big[\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\,\mathrm{Var}\Big(A_{j,h}^{(i)}\Big)\Big(\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big] + \frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big]\\ &\le \frac{2\alpha^2}{nb}\mathbb{E}\Big[\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\Big(\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big] + \frac{8\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\|\nabla f(\mathbf{v}_{j,h};z_i)\|_2\Big)^2\Big], \end{aligned}$$

where we used $\mathrm{Var}(A_{j,h}^{(i)})=\frac{b}{n}(1-\frac{1}{n})\le\frac{b}{n}$ in the second inequality and used the fact that

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\mathfrak{e}_{j,h}^{(i)}\Big)^2\Big] &\le 2\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\|\nabla f(\mathbf{v}_{j,h};z_i)\|_2\Big)^2\Big] + 2\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\|\nabla f(\mathbf{v}_{j,h}^{(i)};z_i')\|_2\Big)^2\Big]\\ &= 4\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\|\nabla f\left(\mathbf{v}_{j,h};z_i\right)\|_2\Big)^2\Big]. \end{aligned}$$

We also notice that

$$\begin{aligned}\mathbb{E}\big[\big(\mathfrak{e}_{j,h}^{(i)}\big)^2\big] &\le 2\mathbb{E}\big[\|\nabla f(\mathbf{v}_{j,h},z_i)\|_2^2\big] + 2\mathbb{E}\big[\|\nabla f(\mathbf{v}_{j,h}^{(i)};z_i')\|_2^2\big]\\ &\le 4L\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i\right)+f\left(\mathbf{v}_{j,h}^{(i)};z_i'\right)\right] = 8L\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i\right)\right]. \end{aligned} \tag{A.12}$$

It then follows that

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq \frac{16\alpha^2 L}{nb}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i\right)\right] + \frac{8\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\|\nabla f\left(\mathbf{v}_{j,h};z_i\right)\|_2\Big)^2\Big]. \tag{A.13}$$

By taking an average over all $i \in [n]$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\right] \\
&\leq \frac{16\alpha^2 L}{n^2 b}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\sum_{i=1}^{n}\eta_{j,h}^2\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i\right)\right] + \frac{8\alpha^2}{n^3}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}\|\nabla f\left(\mathbf{v}_{j,h};z_i\right)\|_2\Big)^2\Big] \\
&\leq \frac{16\alpha^2 L}{nb}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\mathbb{E}\left[F_S\left(\mathbf{v}_{j,h}\right)\right] + \frac{8(t+1)k\alpha^2}{n^3}\sum_{i=1}^{n}\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\mathbb{E}\left[\|\nabla f\left(\mathbf{v}_{j,h};z_i\right)\|_2^2\right] \\
&\leq \Big(\frac{16\alpha^2 L}{nb}+\frac{16\alpha^2 L(t+1)k}{n^2}\Big)\sum_{h=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,h}^2\mathbb{E}\left[F_S\left(\mathbf{v}_{j,h}\right)\right],
\end{aligned} \tag{A.14}$$

where the second inequality holds by applying Cauchy-Schwarz inequality, and the third inequality follows from self-bounding property. The proof is completed. □

A.2 Proof of Theorem 3

We first introduce the optimization error bound for Lookahead in the convex case.

Lemma 9 (Optimization Errors of Lookahead: Convex Case)**.** *Suppose the assumptions in Theorem 2 hold, and further assume that $\eta < \frac{b}{L(b+1)}$, then the following inequality holds*

$$\mathbb{E}\left[F_S\left(\overline{\mathbf{v}}_R\right)-F_S\left(\mathbf{w}^*\right)\right] \leq \frac{b\mathbb{E}\left[\|\mathbf{w}_0-\mathbf{w}_S\|^2\right]}{2\alpha\eta kT\big(b-L\eta(b+1)\big)} + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b-L\eta(b+1)}, \tag{A.15}$$

where $\bar{\mathbf{v}}_R = \frac{1}{Tk}\sum_{t=1}^{T}\sum_{\tau=0}^{k-1}\mathbf{v}_{\tau,t}$.

We need the following property for the L-smooth and convex functions for the proof.

Lemma 10 ((Woodworth et al., 2020))**.** *For any L-smooth and convex F, and any x, and y,*

$$\|\nabla F\left(x\right)-\nabla F\left(y\right)\|^2 \leq L\langle\nabla F\left(x\right)-\nabla F\left(y\right),x-y\rangle,$$

and

$$\|\nabla F\left(x\right)-\nabla F\left(y\right)\|^2 \leq 2L\left(F\left(x\right)-F\left(y\right)-\langle\nabla F\left(y\right),x-y\rangle\right).$$

Proof of Lemma 9. Since $F_S(\mathbf{w}_S) \leq F_S(\mathbf{w}^*)$, an upper bound for $F_S(\overline{\mathbf{v}}_R) - F_S(\mathbf{w}_S)$ is also an upper bound for $F_S(\overline{\mathbf{v}}_R) - F_S(\mathbf{w}^*)$. For the proof below, we assume that the learning rate is constant, that is, $\eta_{\tau,t} = \eta$. We denote $B_{k,t} = \{z_{i_{k,t}^{(1)}},\ldots,z_{i_{k,t}^{(b)}}\}$ and $f\big(\mathbf{v};B_{k,t}\big) = \frac{1}{b}\sum_{j=1}^{b}f(\mathbf{v};z_{i_{k,t}^{(j)}})$. We can hence reformulate the minibatch SGD update as

$$\mathbf{v}_{\tau+1,t} = \mathbf{v}_{\tau,t} - \eta\nabla f\big(\mathbf{v}_{\tau,t};B_{\tau,t}\big).$$

We first notice that

$$\begin{aligned}
\mathbb{E}\big[\|\nabla f\big(\mathbf{v}_{\tau,t};B_{\tau,t}\big)\|^2\big] &= \mathbb{E}\big[\|\nabla f\big(\mathbf{v}_{\tau,t};B_{\tau,t}\big)-\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big] + \mathbb{E}\big[\|\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big] \\
&= \frac{1}{b}\mathbb{E}\big[\|\nabla f\big(\mathbf{v}_{\tau,t};z_{i_{\tau,t}^{(1)}}\big)-\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big] + \mathbb{E}\big[\|\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big] \\
&= \frac{\mathbb{E}\big[\|\nabla f\big(\mathbf{v}_{\tau,t};z_{i_{\tau,t}^{(1)}}\big)\|^2\big]}{b} - \frac{\mathbb{E}\big[\|\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big]}{b} + \mathbb{E}\big[\|\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big] \\
&\leq \frac{2L\mathbb{E}\big[F_S(\mathbf{v}_{\tau,t})\big]}{b} + \mathbb{E}\big[\|\nabla F_S\left(\mathbf{v}_{\tau,t}\right)\|^2\big] \\
&\leq \frac{2L\mathbb{E}\big[F_S(\mathbf{v}_{\tau,t})\big]}{b} + 2L\mathbb{E}[F_S(\mathbf{v}_{\tau,t}) - F_S(\mathbf{w}_S)],
\end{aligned} \tag{A.16}$$

where the last inequality follows from Lemma 10, where we set $y = \mathbf{w}_S$. We then analyze the single step in the inner-loop,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{v}_{\tau+1,t} - \mathbf{w}_S\|^2\right] &= \mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \eta\nabla f\big(\mathbf{v}_{\tau,t}; B_{\tau,t}\big) - \mathbf{w}_S\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2 - 2\eta\langle \mathbf{v}_{\tau,t} - \mathbf{w}_S, \nabla f\big(\mathbf{v}_{\tau,t}; B_{\tau,t}\big)\rangle + \eta^2\|\nabla f\big(\mathbf{v}_{\tau,t}; B_{\tau,t}\big)\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2\right] - 2\eta\mathbb{E}\left[\langle \mathbf{v}_{\tau,t} - \mathbf{w}_S, \nabla F_S\left(\mathbf{v}_{\tau,t}\right)\rangle\right] + \eta^2\mathbb{E}\left[\|\nabla f\big(\mathbf{v}_{\tau,t}; B_{\tau,t}\big)\|^2\right].\end{aligned} \tag{A.17}$$

By convexity, we have $\langle \mathbf{v}_{\tau,t} - \mathbf{w}_S, \nabla F_S\left(\mathbf{v}_{\tau,t}\right)\rangle \geq F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)$. Substituting this and the above result, we get

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{v}_{\tau+1,t} - \mathbf{w}_S\|^2\right] &\leq \mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2\right] - 2\eta\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] + \eta^2\Big(\frac{2L\mathbb{E}\big[F_S(\mathbf{v}_{\tau,t})\big]}{b} + 2L\mathbb{E}\big[F_S(\mathbf{v}_{\tau,t}) - F_S(\mathbf{w}_S)\big]\Big) \\ &= \mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2\right] - \big(2\eta - \frac{2L\eta^2(b+1)}{b}\big)\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] + \frac{2L\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b}.\end{aligned}$$

It then follows that

$$2\eta\Big(1-\frac{L\eta(b+1)}{b}\Big)\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] \leq \mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2 - \|\mathbf{v}_{\tau+1,t} - \mathbf{w}_S\|^2\right] + \frac{2L\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b}.$$

Recall the assumption of $\eta \leq \frac{b}{L(b+1)}$, we can divide by $2\eta(1 - \frac{L\eta(b+1)}{b})$ and get

$$\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] \leq \frac{b}{2\eta\big(b - L\eta(b+1)\big)}\mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2 - \|\mathbf{v}_{\tau+1,t} - \mathbf{w}_S\|^2\right] + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b - L\eta(b+1)}.$$

We take an average of the above inequality from $\tau = 0$ to $k-1$, and get

$$\begin{aligned}\frac{1}{k}\sum_{\tau=0}^{k-1}\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] &\leq \frac{b}{2\eta k\big(b - L\eta(b+1)\big)}\sum_{\tau=0}^{k-1}\mathbb{E}\left[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2 - \|\mathbf{v}_{\tau+1,t} - \mathbf{w}_S\|^2\right] + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b - L\eta(b+1)} \\ &= \frac{b}{2\eta k\big(b - L\eta(b+1)\big)}\mathbb{E}\left[\|\mathbf{v}_{0,t} - \mathbf{w}_S\|^2 - \|\mathbf{v}_{k,t} - \mathbf{w}_S\|^2\right] + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b - L\eta(b+1)}.\end{aligned} \tag{A.18}$$

By the slow updating rule of Lookahead, we know $(1-\alpha)(\mathbf{w}_{t-1} - \mathbf{w}^*) = (\mathbf{w}_t - \mathbf{w}^*) - \alpha(\mathbf{v}_{k,t} - \mathbf{w}^*)$ and get

$$\|\mathbf{v}_{0,t} - \mathbf{w}_S\|^2 - \|\mathbf{v}_{k,t} - \mathbf{w}_S\|^2 = \|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2 - \|\mathbf{v}_{k,t} - \mathbf{w}_S\|^2 \leq \frac{1}{\alpha}\left(\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2 - \|\mathbf{w}_t - \mathbf{w}_S\|^2\right).$$

Substituting this into (A.18), we have

$$\frac{1}{k}\sum_{\tau=0}^{k-1}\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] \leq \frac{b}{2\alpha\eta k\big(b - L\eta(b+1)\big)}\mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2 - \|\mathbf{w}_t - \mathbf{w}_S\|^2\right] + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b - L\eta(b+1)}.$$

We take an average of the above inequality and get

$$\begin{aligned}\frac{1}{kT}\sum_{t=1}^{T}\sum_{\tau=0}^{k-1}\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] &\leq \frac{b}{2\alpha\eta kT\big(b - L\eta(b+1)\big)}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2 - \|\mathbf{w}_t - \mathbf{w}_S\|^2\right] + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b - L\eta(b+1)} \\ &\leq \frac{b\mathbb{E}\left[\|\mathbf{w}_0 - \mathbf{w}_S\|^2\right]}{2\alpha\eta kT\big(b - L\eta(b+1)\big)} + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}_S\big)\big]}{b - L\eta(b+1)} \\ &\leq \frac{b\mathbb{E}\left[\|\mathbf{w}_0 - \mathbf{w}_S\|^2\right]}{2\alpha\eta kT\big(b - L\eta(b+1)\big)} + \frac{L\eta\mathbb{E}\big[F_S\big(\mathbf{w}^*\big)\big]}{b - L\eta(b+1)}.\end{aligned} \tag{A.19}$$

We complete the proof by applying the Jensen's inequality. □

Proof of Theorem 3. By Lemma 1 (part (b)) and (5.2), we have (note our stability bounds also apply to $\bar{\mathbf{v}}_R$ due to the convexity of norm)

$$\mathbb{E}\left[F(\overline{\mathbf{v}}_R) - F_S(\overline{\mathbf{v}}_R)\right] \leq \frac{L}{\gamma}\mathbb{E}[F_S(\overline{\mathbf{v}}_R)] + (L+\gamma)\left(\frac{8\alpha^2 L}{nb} + \frac{8\alpha^2 LTk}{n^2}\right)\sum_{h=1}^{T}\sum_{j=0}^{k-1}\eta_{j,h}^2\mathbb{E}\left[F_S\left(\mathbf{v}_{j,h}\right)\right]. \tag{A.20}$$

By (A.19) we know that

$$\frac{1}{kT}\sum_{t=1}^{T}\sum_{\tau=0}^{k-1}\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right)\right] \lesssim F(\mathbf{w}^*) + \frac{L\eta F(\mathbf{w}^*)}{b} + \frac{1}{\alpha\eta kT}. \tag{A.21}$$

Let $R = Tk$. We combine the above inequalities and get

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{v}}_R) - F_S(\overline{\mathbf{v}}_R)\right] \lesssim\ & \frac{L(F(\mathbf{w}^*) + L\eta F(\mathbf{w}^*)/b + 1/(\alpha\eta R))}{\gamma} \\ & + L(L+\gamma)\alpha^2\eta^2\left(\frac{1}{nb} + \frac{R}{n^2}\right)\left(RF(\mathbf{w}^*) + RL\eta F(\mathbf{w}^*)/b + 1/(\alpha\eta)\right). \end{aligned}\tag{A.22}$$

We plug (A.22) and the optimization error bound (A.15) back into (3.1) and get

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{v}}_R)\right] - F(\mathbf{w}^*) \lesssim\ & \frac{L\eta F(\mathbf{w}^*)}{b} + \frac{1}{\alpha\eta R} + \frac{F(\mathbf{w}^*) + L\eta F(\mathbf{w}^*)/b + 1/(\alpha\eta R)}{\gamma} + \\ & L(L+\gamma)\alpha^2\eta^2\left(\frac{1}{nb} + \frac{R}{n^2}\right)\left(RF(\mathbf{w}^*) + RL\eta F(\mathbf{w}^*)/b + 1/(\alpha\eta)\right).\end{aligned}$$

The proof is completed. □

Proof of Corollary 4. We first consider the case $F(\mathbf{w}^*) \geq \frac{1}{n}$. Fix any constant $\alpha \in (0, 1]$, we choose $\eta = \frac{b}{\sqrt{nF(\mathbf{w}^*)}}$, $R \asymp \frac{n}{b}$, and $\gamma = \sqrt{nF(\mathbf{w}^*)} \geq 1$. Note the assumption $b \leq \sqrt{nF(\mathbf{w}^*)}/(2L)$ ensures that $\eta \leq 1/(2L)$. Then Eq. (5.3) implies

$$\begin{aligned}\mathbb{E}\left[F(\overline{\mathbf{v}}_R) - F(\mathbf{w}^*)\right] &\lesssim \frac{LF(\mathbf{w}^*)}{\sqrt{nF(\mathbf{w}^*)}} + \frac{F(\mathbf{w}^*)^{\frac{1}{2}}}{\sqrt{n}} + \frac{(nF(\mathbf{w}^*))^{\frac{1}{2}} + L + 1}{n} \\ &\quad + \frac{2L}{n^2F(\mathbf{w}^*)}\big(L + (nF(\mathbf{w}^*))^{\frac{1}{2}}\big)\big(nF(\mathbf{w}^*) + (L+1)(nF(\mathbf{w}^*))^{\frac{1}{2}}\big) \\ &\lesssim \frac{LF(\mathbf{w}^*)^{1/2}}{\sqrt{n}} + \frac{L^2}{n}.\end{aligned}$$

We now consider the case $F(\mathbf{w}^*) < \frac{1}{n}$. We fix $\alpha \in (0, 1]$ as a constant, and choose $\eta = \frac{1}{2L}$, $R \asymp n$, and $\gamma = 1$. Then Eq. (5.3) implies

$$\mathbb{E}\left[F(\overline{\mathbf{v}}_R) - F(\mathbf{w}^*)\right] \lesssim F(\mathbf{w}^*) + \frac{L}{n} + \frac{L+1}{4nL}\big(nF(\mathbf{w}^*) + 2L\big) \lesssim \frac{L}{n} + F(\mathbf{w}^*).$$

The proof is completed. □

A.3 Proof of Theorem 5

Proof. Recalling from Eq. (A.1) the refined Lookahead updating rule, we have

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2 \\ &\leq (1-\alpha)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + \alpha\Big\|\mathbf{v}_{k-1,t+1} - \frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n} A_{k-1,t+1}^{(m)}\nabla f(\mathbf{v}_{k-1,t+1}; z_m) - \mathbf{v}_{k-1,t+1}^{(i)} \\ &+ \frac{\eta_{k-1,t+1}}{b}\sum_{m:m\neq i}^{n} A_{k-1,t+1}^{(m)}\nabla f(\mathbf{v}_{k-1,t+1}^{(i)}; z_m)\Big\|_2 + \frac{\alpha A_{k-1,t+1}^{(i)}\eta_{k-1,t+1}}{b}\|\nabla f\left(\mathbf{v}_{k-1,t+1}; z_i\right) - \nabla f(\mathbf{v}_{k-1,t+1}^{(i)}; z_i')\|_2.\end{aligned}$$

Since f is smooth and $\sum_{m:m\neq i}^{n} A_{k-1,t+1}^{(m)} \leq b$, therefore $\mathbf{v} \mapsto \frac{1}{b}\sum_{m:m\neq i}^{n} A_{k-1,t+1}^{(m)} f(\mathbf{v}; z_m)$ is L-smooth. It follows from Lemma 8 and the assumption $\eta_{k-1,t+1} \leq \frac{1}{L}$ that

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2 \leq\ & (1-\alpha)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + \\ & \frac{\alpha\eta_{k-1,t+1}A_{k-1,t+1}^{(i)}\mathfrak{C}_{k-1,t+1}^{(i)}}{b} + \alpha\left(1 - \frac{\mu\eta_{k-1,t+1}}{2}\right)\|\mathbf{v}_{k-1,t+1} - \mathbf{v}_{k-1,t+1}^{(i)}\|_2.\end{aligned}\tag{A.23}$$

We take the expectation on both sides and get

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big] \le (1-\alpha)\,\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\big] + \frac{2\alpha\eta_{k-1,t+1}\sqrt{2L\mathbb{E}\left[f\left(\mathbf{v}_{k-1,t+1};z_i\right)\right]}}{n} + \alpha\left(1-\frac{\mu\eta_{k-1,t+1}}{2}\right)\mathbb{E}\big[\|\mathbf{v}_{k-1,t+1}-\mathbf{v}_{k-1,t+1}^{(i)}\|_2\big],$$

where we have used (A.4) and (A.6). We do the iteration on inner-loop, and get

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big] &\le (1-\alpha)\,\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\big] + \frac{2\alpha\sqrt{2L}}{n}\sum_{j=0}^{k-1}\eta_{j,t+1}\sqrt{\mathbb{E}\left[f\left(\mathbf{v}_{j,t+1};z_i\right)\right]}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t+1}}{2}\right)\\ &\quad+\alpha\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\big]\prod_{j=0}^{k-1}\left(1-\frac{\mu\eta_{j,t+1}}{2}\right)\\ &\le (1-\frac{\alpha}{2})\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\big] + \frac{2\alpha\sqrt{2L}}{n}\sum_{j=0}^{k-1}\eta_{j,t+1}\sqrt{\mathbb{E}\left[f\left(\mathbf{v}_{j,t+1};z_i\right)\right]}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t+1}}{2}\right),\end{aligned}$$

where we have used the following inequality due to the the assumption $\eta_{j,t+1}\ge\frac{2\ln 2}{k\mu}$

$$\prod_{j=0}^{k-1}\left(1-\frac{\mu\eta_{j,t+1}}{2}\right)\le\exp\Big(-\sum_{j=0}^{k}\frac{\mu\eta_{j,t+1}}{2}\Big)\le\exp\Big(-k\frac{\mu 2\log 2}{2k\mu}\Big)=\frac{1}{2}. \tag{A.24}$$

By iteration on outer-loop,

$$\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big]\le\frac{2\alpha\sqrt{2L}}{n}\sum_{t'=1}^{t+1}(1-\frac{\alpha}{2})^{t+1-t'}\sum_{j=0}^{k-1}\eta_{j,t'}\sqrt{\mathbb{E}\left[f\left(\mathbf{v}_{j,t'};z_i\right)\right]}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right). \tag{A.25}$$

Taking an average over i and using the concavity of $x\mapsto\sqrt{x}$, we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big] &\le \frac{2\alpha\sqrt{2L}}{n^2}\sum_{t'=1}^{t+1}(1-\frac{\alpha}{2})^{t+1-t'}\sum_{j=0}^{k-1}\sum_{i=1}^{n}\eta_{j,t'}\sqrt{\mathbb{E}\left[f\left(\mathbf{v}_{j,t'};z_i\right)\right]}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right)\\ &\le \frac{2\alpha\sqrt{2L}}{n}\sum_{t'=1}^{t+1}(1-\frac{\alpha}{2})^{t+1-t'}\sum_{j=0}^{k-1}\eta_{j,t'}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[f\left(\mathbf{v}_{j,t'};z_i\right)\right]\Big)^{\frac{1}{2}}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right)\\ &= \frac{2\alpha\sqrt{2L}}{n}\sum_{t'=1}^{t+1}(1-\frac{\alpha}{2})^{t+1-t'}\sum_{j=0}^{k-1}\eta_{j,t'}\sqrt{\mathbb{E}\left[F_S\left(\mathbf{v}_{j,t'}\right)\right]}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right).\end{aligned}$$

This established the stated ℓ_1-stability bound (5.4).
We now prove Eq. (5.5). Recall Eq. (A.2), we do iteration on inner-loop in Eq. (A.23) and get

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\\ &\le(1-\alpha)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\frac{\alpha}{b}\sum_{j=0}^{k-1}\eta_{j,t+1}A_{j,t+1}^{(i)}\mathfrak{e}_{j,t+1}^{(i)}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t+1}}{2}\right)+\alpha\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2\prod_{j=0}^{k-1}\left(1-\frac{\mu\eta_{j,t+1}}{2}\right)\\ &\le\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\frac{\alpha}{b}\sum_{j=0}^{k-1}\eta_{j,t+1}A_{j,t+1}^{(i)}\mathfrak{e}_{j,t+1}^{(i)}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t+1}}{2}\right).\end{aligned}$$

Then we iterate on outer-loop and get

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\le\frac{\alpha}{b}\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}A_{j,t'}^{(i)}\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right)\\ &=\frac{\alpha}{b}\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\Big(A_{j,t'}^{(i)}-\frac{b}{n}\Big)\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right)+\frac{\alpha}{n}\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\left(1-\frac{\mu\eta_{j',t'}}{2}\right).\end{aligned}$$

By taking the square and the expectation on both sides, we get

$$
\begin{aligned}
&\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\\
&\le \frac{2\alpha^2}{b^2}\mathbb{E}\Big[\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}(A_{j,t'}^{(i)}-\frac{b}{n})\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)^2\Big]+\frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)^2\Big]\\
&= \frac{2\alpha^2}{b^2}\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}^2\mathbb{E}\Big[\Big(A_{j,t'}^{(i)}-\frac{b}{n}\Big)^2(\mathfrak{e}_{j,t'}^{(i)})^2\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)^2\Big]+\frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)^2\Big]\\
&\le \frac{2\alpha^2}{nb}\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}^2\mathbb{E}\Big[(\mathfrak{e}_{j,t'}^{(i)})^2\Big]\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)^2+\frac{2\alpha^2}{n^2}\mathbb{E}\Big[\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)^2\Big]
\end{aligned}
\tag{A.26}
$$

where we used (A.11) and $\mathbb{E}_{B_{j,t'}}\big[\big(A_{j,t'}^{(i)}-\frac{b}{n}\big)^2\big]\le\frac{b}{n}$. For the second term, we apply the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\mathfrak{e}_{j,t'}^{(i)}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)^2\\
&\le\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}(\mathfrak{e}_{j,t'}^{(i)})^2\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)\\
&\le\frac{2(t+1)}{\mu}\Big(\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\eta_{j,t'}(\mathfrak{e}_{j,t'}^{(i)})^2\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big),
\end{aligned}
\tag{A.27}
$$

where the following result is used in the last inequality

$$
\begin{aligned}
\sum_{j=0}^{k-1}\eta_{j,t'}\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)&=\frac{2}{\mu}\sum_{j=0}^{k-1}\Big(1-\Big(1-\frac{\mu\eta_{j,t'}}{2}\Big)\Big)\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\\
&=\frac{2}{\mu}\sum_{j=0}^{k-1}\Big(\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)-\prod_{j'=j}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)\\
&=\frac{2}{\mu}\Big(1-\prod_{j'=0}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big)\Big)\le\frac{2}{\mu}.
\end{aligned}
\tag{A.28}
$$

Combining the above discussions together, we further get

$$
\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\le\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\Big(\frac{2\alpha^2\eta_{j,t'}^2}{nb}+\frac{4\,(t+1)\,\alpha^2\eta_{j,t'}}{n^2\mu}\Big)\mathbb{E}\big[(\mathfrak{e}_{j,t'}^{(i)})^2\big]\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big).
$$

Recalling result in (A.12), $\mathbb{E}\big[(\mathfrak{e}_{j,t'}^{(i)})^2\big]\le 8L\mathbb{E}\left[f\left(\mathbf{v}_{j,h};z_i\right)\right]$, we further derive

$$
\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\le\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\Big(\frac{16\alpha^2\eta_{j,t'}^2}{nb}+\frac{32\,(t+1)\,\alpha^2\eta_{j,t'}}{n^2\mu}\Big)\mathbb{E}\left[f\left(\mathbf{v}_{j,t'};z_i\right)\right]\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big).
$$

Taking an average over $i\in[n]$, we get the stated bound

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\le\sum_{t'=1}^{t+1}\sum_{j=0}^{k-1}\Big(\frac{16\alpha^2\eta_{j,t'}^2}{nb}+\frac{32\,(t+1)\,\alpha^2\eta_{j,t'}}{n^2\mu}\Big)\mathbb{E}\left[F_S\left(\mathbf{v}_{j,t'}\right)\right]\prod_{j'=j+1}^{k-1}\Big(1-\frac{\mu\eta_{j',t'}}{2}\Big).
$$

The proof is completed. □

A.4 Proof of Theorem 6

We first state and prove the optimization error bound.

Lemma 11 (Optimization Error of Lookahead: Strongly Convex Case)**.** *Suppose the assumptions in Theorem 5 hold, by setting the learning rate $\eta = \frac{b\mu}{2L^2(b+1)}$, the optimization error of the output $\mathbf{w}_T$ of Lookahead satisfies*

$$\mathbb{E}[F_S(\mathbf{w}_T) - F_S(\mathbf{w}^*)] \leq \frac{L}{2}\, e^{-\frac{3}{4}\alpha k\mu\eta T}\, \mathbb{E}\big[\|\mathbf{w}_0 - \mathbf{w}_S\|^2\big] + \frac{L\alpha}{2}\sum_{t=0}^{T-1} e^{-\frac{3}{4}\alpha k\mu\eta t} \sum_{k'=0}^{k-1} e^{-\frac{3}{4}\mu\eta k'}\, \frac{2\eta^2 L}{b}\, \mathbb{E}[F_S(\mathbf{w}_S)]\,. \tag{A.29}$$

Furthermore, by choosing $b \lesssim n$, $k = \frac{2L}{\alpha\mu}$, and $T \asymp \log(\mu n)$, we have

$$\mathbb{E}[F_S(\mathbf{w}_T) - F_S(\mathbf{w}^*)] \lesssim \frac{L}{n}\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_S\|^2] + \mathbb{E}[F_S(\mathbf{w}_S)]. \tag{A.30}$$

Proof. Since $F_S(\mathbf{w}_S) \leq F_S(\mathbf{w}^*)$, an upper bound for $F_S(\mathbf{w}_T) - F_S(\mathbf{w}_S)$ is also an upper bound for $F_S(\mathbf{w}_T) - F_S(\mathbf{w}^*)$. Since the function $F_S(\mathbf{w})$ is μ-strongly convex and $\mathbf{w}_S$ is the optimum of $F_S(\mathbf{w})$, we have

$$\begin{aligned} F_S(\mathbf{v}_{\tau-1,t}) &\geq F_S(\mathbf{w}_S) + \langle \nabla F_S(\mathbf{w}_S),\, \mathbf{w}_S - \mathbf{v}_{\tau-1,t}\rangle + \frac{\mu}{2}\,\|\mathbf{w}_S - \mathbf{v}_{\tau-1,t}\|_2^2 \\ &= F_S(\mathbf{w}_S) + \frac{\mu}{2}\,\|\mathbf{w}_S - \mathbf{v}_{\tau-1,t}\|_2^2\,. \end{aligned}$$

Similarly, we have

$$F_S(\mathbf{w}_S) \geq F_S(\mathbf{v}_{\tau-1,t}) + \langle \nabla F_S(\mathbf{v}_{\tau-1,t})\,,\, \mathbf{w}_S - \mathbf{v}_{\tau-1,t}\rangle + \frac{\mu}{2}\,\|\mathbf{w}_S - \mathbf{v}_{\tau-1,t}\|_2^2\,.$$

It then follows that

$$\begin{aligned} \mathbb{E}[\big\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\big\|^2] &= \mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \eta\,\nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big) - \mathbf{w}_S\big\|^2\big] \\ &= \mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2 - 2\eta\big\langle \mathbf{v}_{\tau-1,t} - \mathbf{w}_S,\, \nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big)\big\rangle + \eta^2\big\|\nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big)\big\|^2\big] \\ &= \mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2 - 2\eta\big\langle \mathbf{v}_{\tau-1,t} - \mathbf{w}_S,\, \nabla F_S(\mathbf{v}_{\tau-1,t})\,\big\rangle + \eta^2\big\|\nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big)\big\|^2\big] \\ &\leq \mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2 + 2\eta\Big(F_S(\mathbf{w}_S) - F_S(\mathbf{v}_{\tau-1,t}) - \frac{\mu}{2}\big\|\mathbf{w}_S - \mathbf{v}_{\tau-1,t}\big\|_2^2\Big) + \eta^2\big\|\nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big)\big\|^2\big] \\ &\leq \mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2 + 2\eta\big(-\frac{\mu}{2}\big\|\mathbf{w}_S - \mathbf{v}_{\tau-1,t}\big\|_2^2 - \frac{\mu}{2}\big\|\mathbf{w}_S - \mathbf{v}_{\tau-1,t}\big\|_2^2\big) + \eta^2\big\|\nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big)\big\|^2\big] \\ &\leq (1 - 2\mu\eta)\mathbb{E}[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2] + \eta^2\mathbb{E}\big[\big\|\nabla f\big(\mathbf{v}_{\tau-1,t}; B_{\tau-1,t}\big)\big\|^2\big]. \end{aligned}$$

For the second term, we use the result of (A.16) and have

$$\begin{aligned} \mathbb{E}\big[\big\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\big\|^2\big] &\leq (1 - 2\mu\eta)\mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2\big] + \eta^2\frac{2L\mathbb{E}\big[F_S(\mathbf{v}_{\tau-1,t})\big]}{b} + 2L\eta^2\mathbb{E}[F_S(\mathbf{v}_{\tau-1,t}) - F_S(\mathbf{w}_S)] \\ &\leq (1 - 2\mu\eta + \eta^2L^2)\mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2\big] + \eta^2\frac{2L\mathbb{E}\big[F_S(\mathbf{v}_{\tau-1,t}) - F_S(\mathbf{w}_S)\big] + 2L\mathbb{E}[F_S(\mathbf{w}_S)]}{b} \\ &\leq (1 - 2\mu\eta + \eta^2\frac{L^2(b+1)}{b})\mathbb{E}\big[\big\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\big\|^2\big] + \eta^2\frac{2L\mathbb{E}[F_S(\mathbf{w}_S)]}{b}, \end{aligned}$$

where we have used $F_S(\mathbf{w}) - F_S(\mathbf{w}_S) \leq \frac{L}{2}\|\mathbf{w} - \mathbf{w}_S\|_2^2$. For simplicity, we define C as

$$C = \frac{L^2(b+1)}{b}.$$

The recurrence relation simplifies as

$$\mathbb{E}[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2] \leq \big(1 - 2\mu\eta + C\eta^2\big)\,\mathbb{E}[\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\|^2] + \eta^2\frac{2L\mathbb{E}[F_S(\mathbf{w}_S)]}{b}. \tag{A.31}$$

We now choose

$$\eta = \frac{\mu}{2C} = \frac{\mu b}{2L^2(b+1)}.$$

Substituting this value back into the multiplicative factor gives

$$1 - 2\mu\left(\frac{\mu}{2C}\right) + C\left(\frac{\mu}{2C}\right)^2 = 1 - \frac{\mu^2}{C} + \frac{\mu^2}{4C} = 1 - \frac{3\mu^2}{4C} = 1 - \frac{3}{2}\mu\eta.$$

With this choice, the one-step recurrence (A.31) becomes

$$\mathbb{E}[\|\mathbf{v}_{\tau,t} - \mathbf{w}_S\|^2] \leq \left(1 - \frac{3}{2}\mu\eta\right)\mathbb{E}[\|\mathbf{v}_{\tau-1,t} - \mathbf{w}_S\|^2] + \eta^2\frac{2L\mathbb{E}[F_S(\mathbf{w}_S)]}{b}.$$

By applying the previous inequality recursively for the inner loop, we have

$$\mathbb{E}[\|\mathbf{v}_{k,t} - \mathbf{w}_S\|^2] \leq \left(1 - \frac{3}{2}\mu\eta\right)^k \mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2] + \sum_{k'=0}^{k-1} \left(1 - \frac{3}{2}\mu\eta\right)^{k'} \eta^2 \frac{2L\mathbb{E}[F_S(\mathbf{w}_S)]}{b}.$$

We now substitute this result back to the outer-loop. Recall the slow weights recurrence $\mathbf{w}_t = (1-\alpha)\mathbf{w}_{t-1} + \alpha\mathbf{v}_{k,t}$,

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

Taking the expectation gives

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_S\|^2] &\leq (1-\alpha)\mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2] + \alpha\mathbb{E}[\|\mathbf{v}_{k,t} - \mathbf{w}_S\|^2] \\ &\leq (1-\alpha)\mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2] + \alpha\left(1 - \frac{3}{2}\mu\eta\right)^k \mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2] + \alpha\sum_{k'=0}^{k-1}\left(1 - \frac{3}{2}\mu\eta\right)^{k'} \frac{2L\eta^2\mathbb{E}[F_S(\mathbf{w}_S)]}{b} \\ &= \left[1 - \alpha + \alpha\left(1 - \frac{3}{2}\mu\eta\right)^k\right]\mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2] + \alpha\sum_{k'=0}^{k-1}\left(1 - \frac{3}{2}\mu\eta\right)^{k'} \frac{2L\eta^2\mathbb{E}[F_S(\mathbf{w}_S)]}{b} \\ &= \left[1 - \alpha\left(1 - \left(1 - \frac{3}{2}\mu\eta\right)^k\right)\right]\mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_S\|^2] + \alpha\sum_{k'=0}^{k-1}\left(1 - \frac{3}{2}\mu\eta\right)^{k'} \frac{2L\eta^2\mathbb{E}[F_S(\mathbf{w}_S)]}{b}.\end{aligned}$$

Let ρ be the contraction factor for the outer loop:

$$\rho = 1 - \alpha\left(1 - \left(1 - \frac{3}{2}\mu\eta\right)^k\right).$$

Since $0 < (1 - \frac{3}{2}\mu^2/C) < 1$ and $\alpha > 0$, we have $0 < \rho < 1$. Unwinding this recurrence from $t = 1$ to T:

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_S\|^2] \leq \rho^t\mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_S\|^2] + \alpha\sum_{t'=0}^{t-1}\rho^{t'}\sum_{k'=0}^{k-1}\left(1 - \frac{3}{2}\mu\eta\right)^{k'} \frac{2L\eta^2\mathbb{E}[F_S(\mathbf{w}_S)]}{b}. \tag{A.32}$$

Finally, using the L-smoothness property, $\mathbb{E}[F_S(\mathbf{w}_t) - F_S(\mathbf{w}_S)] \leq \frac{L}{2}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_S\|^2]$, we arrive at the final optimization error bound.

$$\begin{aligned}\mathbb{E}[F_S(\mathbf{w}_T) - F_S(\mathbf{w}_S)] \leq \frac{L}{2}&\left[1 - \alpha\left(1 - \left(1 - \frac{3}{2}\mu\eta\right)^k\right)\right]^T \mathbb{E}[\|\mathbf{w}_0 - \mathbf{w}_S\|^2 \\ &+ \frac{L\alpha}{2}\sum_{t'=0}^{T-1}\left[1 - \alpha\left(1 - \left(1 - \frac{3}{2}\mu\eta\right)^k\right)\right]^{t'}\sum_{k'=0}^{k-1}\left(1 - \frac{3}{2}\mu\eta\right)^{k'} \frac{2L\eta^2\mathbb{E}[F_S(\mathbf{w}_S)]}{b}.\end{aligned} \tag{A.33}$$

We use the inequalities $1 + x \leq e^x$ for all real x and $1 - e^{-x} \geq \frac{x}{1+x}$ for all $x \geq 0$ to get the following.

$$\begin{aligned}\left[1 - \alpha\big(1 - (1 - \frac{3}{2}\mu\eta)^k\big)\right]^T &\leq \exp\Big\{-\alpha\big[1 - (1 - \frac{3}{2}\mu\eta)^k\big]T\Big\} \\ &\leq \exp\Big\{-\alpha\big(1 - \exp\{-\frac{3}{2}k\mu\eta\}\big)T\Big\} \\ &\leq \exp\Big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta + 2}T\Big\}.\end{aligned}$$

Then the optimization error bound becomes

$$\begin{aligned}\mathbb{E}[F_S(\mathbf{w}_T) - F_S(\mathbf{w}_S)] \leq \frac{L}{2}&\exp\Big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta + 2}T\Big\}\mathbb{E}\left[\|\mathbf{w}_0 - \mathbf{w}_S\|^2\right] \\ &+ \frac{L\alpha}{2}\sum_{t=0}^{T-1}\exp\Big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta + 2}t\Big\}\sum_{k'=0}^{k-1}\exp\{-\frac{3}{2}\mu\eta k'\}\frac{2\eta^2 L}{b}\,\mathbb{E}[F_S(\mathbf{w}_S)]\,.\end{aligned} \tag{A.34}$$

We now choose the parameters to be $k = \frac{2L}{\mu\alpha}$, $T \asymp \log(n)$, and we fix α. Since $b \geq 1$, we have $L\eta = \frac{b\mu}{2L(b+1)} \in [1/4, 1/2)$. Then with the above k, we know

$$k\mu\eta \geq \frac{2L}{\mu\alpha}\mu\eta = \frac{2}{\alpha}L\eta \geq \frac{1}{2\alpha}.$$

Hence

$$\frac{3k\mu\eta}{3k\mu\eta+2} \geq \frac{3}{3+4\alpha}. \tag{A.35}$$

It then follows that

$$\sum_{t=0}^{T-1} \exp\Big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta+2}t\Big\} \leq \frac{1}{1-e^{-3\alpha/(3+4\alpha)}} \asymp 1.$$

Also, since $\mu\eta \leq \mu/2L \leq 1$, we can use $1-e^{-x} \geq x/2$ for $x \in (0,1]$ and get

$$\sum_{k'=0}^{k-1} e^{-\frac{3}{2}\mu\eta k'} \;=\; \frac{1-e^{-\frac{3}{2}\mu\eta k}}{1-e^{-\frac{3}{2}\mu\eta}} \;\leq\; \frac{1}{1-e^{-\mu\eta}} \;\leq\; \frac{2}{\mu\eta}.$$

Plugging these into (A.34) yields the bound for the second term

$$\begin{aligned}\frac{L\alpha}{2}\sum_{t=0}^{T-1}\exp\Big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta+2}T\Big\}\sum_{k'=0}^{k-1}\exp\{-\frac{3}{2}\mu\eta k'\}\frac{2\eta^2L}{b}\mathbb{E}\left[F_S(\mathbf{w}_S)\right] &\lesssim \frac{L\alpha}{2}\frac{2}{\mu\eta}\frac{2\eta^2L}{b}\mathbb{E}\left[F_S(\mathbf{w}_S)\right] \\ &\lesssim \frac{L^2\eta}{\mu b}\mathbb{E}\left[F_S(\mathbf{w}_S)\right].\end{aligned}$$

Since $\eta = \frac{\mu b}{2L^2(b+1)}$, this simplifies to

$$\frac{L\alpha}{2}\sum_{t=0}^{T-1}\exp\Big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta+2}t\Big\}\sum_{k'=0}^{k-1}\exp\{-\frac{3}{2}\mu\eta k'\}\frac{2\eta^2L}{b}\mathbb{E}\left[F_S(\mathbf{w}_S)\right] \lesssim \frac{1}{2(b+1)}\mathbb{E}\left[F_S(\mathbf{w}_S)\right] \lesssim \mathbb{E}\left[F_S(\mathbf{w}_S)\right]. \tag{A.36}$$

For the first term, together with (A.35), our choice of T ensures

$$\frac{L}{2}\exp\big\{-\alpha\frac{3k\mu\eta}{3k\mu\eta+2}T\big\}\mathbb{E}\left[\|\mathbf{w}_0-\mathbf{w}_S\|^2\right] \lesssim \frac{L}{n}\mathbb{E}\left[\|\mathbf{w}_0-\mathbf{w}_S\|^2\right]. \tag{A.37}$$

Combining (A.36) and (A.37) gives the final result. □

We now state and prove the generalization bound.

Lemma 12 (Generalization Gap of Lookahead: Strongly Convex Case)**.** *Suppose the assumptions in Theorem 5 hold. Let $\mathbf{w}_T$ be the final output of Lookahead optimizer. By setting the learning rate $\eta = \frac{b\mu}{2L^2(b+1)}$, we have*

$$\mathbb{E}[F(\mathbf{w}_T) - F_S(\mathbf{w}_T)] \lesssim \frac{1}{n\mu} + \frac{1}{n^2}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{nL}\mathbb{E}[F_S(\mathbf{w}_S)].$$

Proof of Lemma 12. We now assume the constant step size $\eta_{\tau,t} = \eta$. Let $\mathbf{w}_S = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. We denote $B_{k,t} = \{z_{i_{k,t}^{(1)}}, \ldots, z_{i_{k,t}^{(b)}}\}$ and $f\big(\mathbf{v}; B_{k,t}\big) = \frac{1}{b}\sum_{j=1}^{b} f(\mathbf{v}; z_{i_{k,t}^{(j)}})$. We can hence reformulate the minibatch SGD update as

$$\mathbf{v}_{\tau+1,t} = \mathbf{v}_{\tau,t} - \eta\nabla f(\mathbf{v}_{\tau,t}; B_{\tau,t}). \tag{A.38}$$

By the strong convexity of f,

$$\begin{aligned}\mathbb{E}[\|\mathbf{v}_{\tau+1,t}-\mathbf{w}_S\|_2^2] &= \mathbb{E}[\|\mathbf{v}_{\tau,t}-\eta\nabla f(\mathbf{v}_{\tau,t};B_{\tau,t})-\mathbf{w}_S\|_2^2] \\ &= \mathbb{E}[\|\mathbf{v}_{\tau,t}-\mathbf{w}_S\|_2^2] - 2\eta\mathbb{E}[\langle\mathbf{v}_{\tau,t}-\mathbf{w}_S, \nabla F_S(\mathbf{v}_{\tau,t})\rangle] + \eta^2\mathbb{E}[\|\nabla f(\mathbf{v}_{\tau,t};B_{\tau,t})\|_2^2] \\ &\leq (1-\mu\eta_{\tau,t})\mathbb{E}[\|\mathbf{v}_{\tau,t}-\mathbf{w}_S\|_2^2] - 2\eta\mathbb{E}[F_S(\mathbf{v}_{\tau,t}) - F_S(\mathbf{w}_S)] + \eta^2\mathbb{E}[\|\nabla f(\mathbf{v}_{\tau,t};B_{\tau,t})\|_2^2].\end{aligned} \tag{A.39}$$

For the last term, we bound it using (A.16) and get

$$\mathbb{E}[\|\mathbf{v}_{\tau+1,t}-\mathbf{w}_S\|_2^2] \leq (1-\mu\eta)\mathbb{E}[\|\mathbf{v}_{\tau,t}-\mathbf{w}_S\|_2^2 - \big(2\eta - \frac{2L\eta^2(b+1)}{b}\big)\mathbb{E}\left[F_S\left(\mathbf{v}_{\tau,t}\right) - F_S\left(\mathbf{w}_S\right)\right] + \frac{2L\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b}.$$

For $\eta = \frac{b\mu}{2L^2(b+1)} \leq \frac{b}{2L(b+1)}$, we have

$$\mathbb{E}[\|\mathbf{v}_{\tau+1,t}-\mathbf{w}_S\|_2^2] \leq (1-\mu\eta)\mathbb{E}[\|\mathbf{v}_{\tau,t}-\mathbf{w}_S\|_2^2] - \eta\mathbb{E}[F_S(\mathbf{v}_{\tau,t}) - F_S(\mathbf{w}_S)] + \frac{2L\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b}.$$

We multiply both sides by $\big(1-\frac{\alpha}{2}\big)^{T-t}(1-\mu\eta/2)^{k-\tau}$ and get

$$\big(1-\frac{\alpha}{2}\big)^{T-t}(1-\mu\eta/2)^{k-\tau}\mathbb{E}[\|\mathbf{v}_{\tau+1,t}-\mathbf{w}_S\|_2^2] \le \big(1-\frac{\alpha}{2}\big)^{T-t}(1-\mu\eta/2)^{k-\tau+1}\mathbb{E}[\|\mathbf{v}_{\tau,t}-\mathbf{w}_S\|_2^2] -$$
$$\big(1-\frac{\alpha}{2}\big)^{T-t}\eta(1-\mu\eta/2)^{k-\tau}\mathbb{E}[F_S(\mathbf{v}_{\tau,t})-F_S(\mathbf{w}_S)] + \frac{2L(1-\frac{\alpha}{2})^{T-t}(1-\mu\eta/2)^{k-\tau}\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b}.$$

By taking a summation of the above inequality, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k-1}\eta_{\tau,t}(1-\mu\eta/2)^{k-\tau}\mathbb{E}[F_S(\mathbf{v}_{\tau,t})-F_S(\mathbf{w}_S)] \\
&\le \sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}(1-\mu\eta/2)^{k+1}\mathbb{E}[\|\mathbf{w}_{t-1}-\mathbf{w}_S\|_2^2] + 2L\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k}\frac{(1-\mu\eta/2)^{k-\tau}\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b} \\
&\le \frac{1}{2}\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\mathbb{E}[\|\mathbf{w}_{t-1}-\mathbf{w}_S\|_2^2] + 2L\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k}\frac{(1-\mu\eta/2)^{k-\tau}\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b},
\end{aligned} \tag{A.40}$$

where we have used Eq. (A.24). We first look at the first term of Eq. (A.40). By (A.32), we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{t-1}-\mathbf{w}_S\|_2^2] &\le \rho^t\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \alpha\sum_{t'=0}^{t-2}\rho^{t'}\sum_{k'=0}^{k-1}\left(1-\frac{3}{2}\mu\eta\right)^{k'}\frac{2L\eta^2\mathbb{E}[F_S(\mathbf{w}_S)]}{b} \\
&\lesssim \frac{1}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{L}\mathbb{E}[F_S(\mathbf{w}_S)].
\end{aligned}$$

where the last inequlity follows from the result of (A.30) and the fact that $\mathbb{E}[F_S(\mathbf{w}_t)-F_S(\mathbf{w}_S)] \le \frac{L}{2}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_S\|^2] \lesssim \frac{L}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \mathbb{E}[F_S(\mathbf{w}_S)]$. Together with the summation, we have

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\mathbb{E}[\|\mathbf{w}_{t-1}-\mathbf{w}\|_2^2] &\lesssim \frac{1}{2}\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\big(\frac{1}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{L}\mathbb{E}[F_S(\mathbf{w}_S)]\big) \\
&\le \frac{1}{2}\frac{1}{1-(1-\frac{\alpha}{2})}\big(\frac{1}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{L}\mathbb{E}[F_S(\mathbf{w}_S)]\big) \\
&\lesssim \frac{1}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{L}\mathbb{E}[F_S(\mathbf{w}_S)].
\end{aligned} \tag{A.41}$$

For the second term of (A.40), by Eq. (A.28) and $\eta \le \frac{\mu}{2L^2}$,

$$\begin{aligned}
2L\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k}\frac{(1-\mu\eta/2)^{k-\tau}\eta^2\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b} &\le \frac{\mu}{\alpha L}\sum_{\tau=0}^{k}\frac{(1-\mu\eta/2)^{k-\tau}\eta\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{b} \\
&\lesssim \frac{\mathbb{E}\big[F_S\left(\mathbf{w}_S\right)\big]}{\alpha L}.
\end{aligned} \tag{A.42}$$

We fix the outer-loop learning rate α and combine Eq. (A.41) and Eq. (A.42) to obtain

$$\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k-1}\eta(1-\mu n/2)^{k-(\tau+1)}\mathbb{E}[F_S(\mathbf{v}_{\tau,t})-F_S(\mathbf{w}_S)] \lesssim \frac{1}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{L}\mathbb{E}[F_S(\mathbf{w}_S)]. \tag{A.43}$$

Recall from Eq. (5.4), we denote S_T:

$$S_T = \sum_{t'=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{j=0}^{k-1}\eta_{j,t'}\sqrt{\mathbb{E}[F_S(\mathbf{v}_{j,t'})]}(1-\mu n/2)^{k-(\tau+1)}.$$

We use the inequality $\sqrt{x} \le (1+x)/2$ for non-negative x. This gives:

$$S_T \le \frac{1}{2}\sum_{t'=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{j=0}^{k-1}\eta_{j,t'}\left(1+\mathbb{E}[F_S(\mathbf{v}_{j,t'})]\right)(1-\mu n/2)^{k-(\tau+1)}.$$

We split this into two parts,

$$S_T \le \underbrace{\frac{1}{2}\Big[\sum_{t'=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{j=0}^{k-1}\eta_{j,t'}(1-\mu n/2)^{k-(\tau+1)}\Big]}_{\text{Part A}} + \underbrace{\frac{1}{2}\Big[\sum_{t'=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{j=0}^{k-1}\eta_{j,t'}\mathbb{E}[F_S(\mathbf{v}_{j,t'})](1-\mu n/2)^{k-(\tau+1)}\Big]}_{\text{Part B}}.$$

We bound each part:

Part A: This part is bounded using the result from Eq. (A.28). The identity shows that for each outer step t', the inner sum over j is bounded by $2/\mu$. Summing over T outer steps yields:

$$\frac{1}{2}\sum_{t'=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{j=0}^{k-1}\eta_{j,t'}(1-\mu n/2)^{k-(\tau+1)} \lesssim \frac{1}{\mu}. \tag{A.44}$$

Part B: Notice that

$$\begin{aligned}
&\sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k-1}\eta(1-\mu n/2)^{k-(\tau+1)}\mathbb{E}[F_S(\mathbf{v}_{\tau,t})] \\
&= \sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k-1}\eta(1-\mu n/2)^{k-(\tau+1)}\mathbb{E}[F_S(\mathbf{v}_{\tau,t})-F_S(\mathbf{w}_S)] + \sum_{t=1}^{T}\big(1-\frac{\alpha}{2}\big)^{T-t}\sum_{\tau=0}^{k-1}\eta(1-\mu n/2)^{k-(\tau+1)}\mathbb{E}[F_S(\mathbf{w}_S)] \\
&\lesssim \frac{1}{n}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{L}\mathbb{E}[F_S(\mathbf{w}_S)].
\end{aligned} \tag{A.45}$$

Combining (A.44) and (A.45) we have:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathbf{w}_T-\mathbf{w}_T^{(i)}\|_2\right] \lesssim \frac{1}{n\mu} + \frac{1}{n^2}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{nL}\mathbb{E}[F_S(\mathbf{w}_S)]. \tag{A.46}$$

By Lemma 1 (a), (A.46) implies

$$\mathbb{E}[F(\mathbf{w}_T)-F_S(\mathbf{w}_T)] \lesssim \frac{1}{n\mu} + \frac{1}{n^2}\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2] + \frac{1}{nL}\mathbb{E}[F_S(\mathbf{w}_S)]. \tag{A.47}$$

The proof is completed. □

Proof of Theorem 6. Note that for $\alpha \le \frac{b\mu}{2\ln 2(b+1)L}$, we have

$$\eta = \frac{b\mu}{2L^2(b+1)} \ge \frac{\ln 2}{L}\alpha = \frac{2\ln 2}{\mu}\frac{\alpha\mu}{2L} = \frac{2\ln 2}{\mu k}$$

Which satisfy the required condition in theorem 5. We now combine the results of lemma 12 and lemma 11 together and get

$$\mathbb{E}[F(\mathbf{w}_T)-F(\mathbf{w}^*)] \lesssim \frac{1}{n\mu} + \big(\frac{1}{nL}+1\big)\mathbb{E}[F_S(\mathbf{w}_S)] + \big(\frac{1}{n^2}+\frac{L}{n}\big)\mathbb{E}[\|\mathbf{w}_0-\mathbf{w}_S\|^2]. \tag{A.48}$$

for $k = \frac{2L}{\alpha\mu}$, and $T \asymp \log(\mu n)$. This completes the proof. □