Generalization Analysis of Biased Stochastic Gradient Methods for Minimax Problems*

Shuang Zeng[1] **Yunwen Lei**[1†] **Yiming Ying**[2]
[1]Department of Mathematics, The University of Hong Kong, Hong Kong, China
[2] School of Mathematics and Statistics, The University of Sydney, Sydney, Australia
zengsh9@connect.hku.hk leiyw@hku.hk yiming.ying@sydney.edu.au

Abstract

Minimax optimization refers to a class of optimization problems with some variables to minimize and other variables to maximize. Biased stochastic gradient methods have shown practical success in solving minimax problems to either improve robustness, enhance communication efficiency, or decrease computational costs. These successes motivate a lot of theoretical works to study the convergence of biased stochastic gradient methods, while their generalization analysis remains untouched. In this paper, we present the first framework to study the stability and generalization of biased stochastic gradient methods for minimax problems. We establish a connection between stability and generalization for minimax problems by relaxing the existing bounded gradient assumption to a bounded second moment condition. We then introduce a generalized Lipschitz-type condition on bias and gradient estimators, and derive a general stability bound to clarify the connection among bias, gradient estimators, and stability. We apply our general analysis to Zeroth-order and Clipped stochastic gradient descent ascent (SGDA), and derive stability bounds that match those of SGDA under appropriate smoothing/clipping parameters. We combine stability and convergence analyses together, and derive optimal excess risk bounds of order $1/\sqrt{n}$, where n is the sample size.

1 Introduction

Many machine learning problems can be formulated as minimax optimization problems, where the decision variables are categorized into a group of variables to minimize and another group of variables to maximize. Typical problems include adversarial training with minimization over model parameters and maximization over perturbations [37], and GANs with minimization over generators and maximization over discriminator's ability to distinguish fake from real data [21]. Beyond adversarial scenarios, minimax problems also appear in robust optimization [12], and AUC maximization [55].

A popular class of optimization methods to solve minimax problems is the gradient-based method. A representative method is the gradient descent ascent (GDA), where we conduct a gradient descent over primal variables and a gradient ascent over dual variables. To further relieve the computational cost of the full gradient, stochastic gradient methods have been proposed by approximating the true gradient with gradient estimates based on random sampling [42], leading to various efficient stochastic methods such as SGDA [38], Variance-Reduced SGDA [42], Optimistic Gradient Descent Ascent (OGDA) [13], and Stochastic Recursive gradiEnt Descent Ascent (SREDA) [36]. Oftentimes, these methods build unbiased gradient estimators whose conditional expectations recover the true

*The work of Yunwen Lei was supported by the Hong Kong Research Grants Council under the GRF projects 17305425 and 17302624. The work of Yiming Ying was supported by Australian Research Council Discovery Project (DP250101359).
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

gradient. This unbiasedness simplifies both the convergence analysis [32, 33, 36, 38, 54] and generalization analysis [18, 30, 58–60]. However, in practical applications, people often build biased gradient estimators to either improve robustness, enhance communication efficiency, or decrease computational costs. These biased gradient estimators lead to the development of biased stochastic gradient methods for minimax problems. Below we present two representative examples:

Zeroth-order SGDA is a computationally efficient method for applications where explicit gradients are computationally infeasible, expensive, or impossible to compute. This is the case when generating adversarial examples with black-box neural networks to test the adversarial robustness [11], and when solving automated machine learning tasks where the gradients w.r.t pipeline configuration parameters is infeasible to compute [52].

Clipped-SGDA is effective for deep learning tasks in recent years since training deep learning models often encounters problems with heavy-tailed noises. Due to the adaptivity in preventing the algorithm from using arbitrarily large gradients [50], it has found very successful applications in natural language processing tasks, like GPT models [9]. Additionally, gradient clipping helps preserve privacy in differentially private machine learning [1].

Due to the broad applications of biased stochastic gradient methods, efforts devoted to their theoretical analysis have gradually increased. Early work focused on stochastic gradient methods for minimization problems, e.g., the convergence analysis from the optimization perspective [16, 24] and the generalization analysis from the learning perspective [41]. Recently, some studies have extended these discussions on minimization problems to minimax problems, e.g., the convergence of Zeroth-order SGDA [6, 53] and Clipped-SGDA [22]. However, these discussions focus on the behavior of these biased methods on training examples, ignoring an important problem in machine learning–generalization. This gap motivates our core question:

Can we develop a general framework to study the generalization of biased stochastic gradient methods for minimax problems?

In this paper, we answer the above question by developing a general generalization analysis of biased stochastic gradient methods for convex-concave and smooth minimax problems by leveraging the algorithmic stability concept. Our contributions are summarized as follows.

• We develop a quantitative connection between stability and generalization for minimax problems. We also introduce generalized Lipschitz-type conditions on the bias and gradient estimators, based on which we develop rather general stability bounds of biased stochastic gradient methods for minimax problems, showing how the bias, gradient estimators and step size affect the generalization performance.

• We show the generalized Lipschitz-type conditions hold for both Zeroth-order SGDA and Clipped-SGDA. Then, we derive the first stability and generalization bounds for Zeroth-order SGDA and Clipped-SGDA, both of which match bounds of SGDA under mild conditions on the smoothing/clipping parameter.

• Our stability analysis relaxes the bounded gradient assumption [30] to a bounded second moment condition. Furthermore, for both Zeroth-order SGDA and Clipped-SGDA, we combine the stability and convergence analysis to give optimal excess risk bounds of order $O(1/\sqrt{n})$, where n is the sample size.

2 Related Work

Stability and Generalization. Algorithmic stability measures how robust an algorithm is to small perturbations in training data. Bousquet and Elisseeff [7] introduced the concept of uniform stability, which is widely used to develop high-probability generalization bounds [8, 19]. Additionally, uniform stability was weakened to study generalization bounds in expectation, including the on-average stability [47]. Another pivotal advance came from Hardt et al. [23], which pioneered the stability analysis of SGD. This work motivates many stability analyses of stochastic optimization algorithms [4, 26, 29, 31, 51, 59]. Some work studied the stability and generalization of biased stochastic gradient methods such as zeroth-order SGD and clipped-SGD [34, 41, 56].

For minimax problems, Zhang et al. [58] used argument stability to study generalization of empirical saddle points for strongly-convex-strongly-concave problems. Farnia and Ozdaglar [18] used uniform stability to study generalization of several minimax optimization algorithms such as GDA, GDmax,

and proximal point method. Lei et al. [30] gave high-probability bounds of SGDA for convex-concave problems. Zhang et al. [59] introduced uniform primal (argument) stability to study the generalization behavior of SGDA for nonconvex-concave problems.

Biased Stochastic Gradient Methods for Minimax Problems. Now we turn to analyses of biased stochastic gradient methods for minimax problems. First, biased stochastic gradient methods for minimization have been well studied. The works [3, 14, 16, 24, 49] developed a general framework to study biased stochastic gradient methods for minimization problems, which covers many popular methods such as Zeroth-order SGD, distributed optimization with biased compression operators, and SGD with delayed gradients. Second, (unbiased) stochastic gradient methods for minimax problems have also been well studied. For example, the convergence analysis of SGDA and its variants has been widely studied for convex-concave [38], strongly-convex-strongly-concave [42], nonconvex-strongly-concave [32, 36], nonconvex-concave [54] and nonconvex-nonconcave problems [33, 54]. These convergence analyses were extended to minimax gradient methods with biased gradient estimates, including Zeroth-order SGDA [6, 53] and Clipped-SGDA [22]. To our knowledge, there are no existing generalization analyses of these biased gradient methods for minimax problems.

3 Problem Formulation

Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z}$. For minimax learning, we aim to learn two parameters $\mathbf{w} \in \mathcal{W}$ and $\mathbf{v} \in \mathcal{V}$, where $\mathcal{W}$ and $\mathcal{V}$ are parameter spaces in $\mathbb{R}^d$. The performance of a pair $(\mathbf{w}, \mathbf{v})$ on an example z can be quantified by $f(\mathbf{w}, \mathbf{v}; z)$, where $f : \mathcal{W} \times \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$ is a loss function. We aim to find the best model $(\mathbf{w}^*, \mathbf{v}^*)$ by solving the minimax optimization problem:

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

where $\mathbb{E}_z[\cdot]$ means the expectation with respect to (w.r.t.) z. In practice, we do not know $\mathbb{P}$ but instead have access to a dataset $S = \{z_1, \ldots, z_n\}$ independently drawn from $\mathbb{P}$. Then, we approximate F by an empirical risk $F_S(\mathbf{w}, \mathbf{v}) := \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}, \mathbf{v}; z_i)$. We apply a (randomized) algorithm A to the dataset S and get a model $A(S) := (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \in \mathcal{W} \times \mathcal{V}$ as an approximate solution of the problem (3.1). While an algorithm A often finds a model $A(S)$ with a good performance on training, it may not generalize well to a test dataset due to the limited number of training examples. This leads to an important concept called the generalization gap as the difference between testing and training.

Since minimax problems involve both minimization and maximization, there are several performance measures to quantify the behavior of $A(S)$, including the strong primal-dual risk, weak primal-dual risk, and primal risk [30, 58]. We focus on the weak primal-dual risk, which is widely used in the literature [5, 18, 30, 58]. Let $\mathbb{E}[\cdot]$ denote the expectation w.r.t. both the algorithm A and dataset S. Denote by $B_R = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq R\}$ a ball of radius R.

Definition 1 (Weak Primal-Dual Risk)**.** Let $\mathbf{w}, \mathbf{v}$ be two random variables, which can be the output model for applying the algorithm A on the dataset S. The weak Primal-Dual (PD) population risk of a model $(\mathbf{w}, \mathbf{v})$ is defined as

$$\triangle^w(\mathbf{w}, \mathbf{v}) = \sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}_{\mathbf{w}}\big[F(\mathbf{w}, \mathbf{v}')\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}_{\mathbf{v}}\big[F(\mathbf{w}', \mathbf{v})\big],$$

which is a deterministic quantity that depends only on the distributions of $\mathbf{w}$ and $\mathbf{v}$ (we abuse the notation $\triangle^w(\mathbf{w}, \mathbf{v})$ to indicate the dependency on the distribution of $\mathbf{w}, \mathbf{v}$ rather than the models).

A standard approach to estimate population risk is to decompose it into a generalization gap and an optimization error [30, 61]

$$\Delta^w(\mathbf{w}, \mathbf{v}) = (\Delta^w(\mathbf{w}, \mathbf{v}) - \Delta_S^w(\mathbf{w}, \mathbf{v})) + \Delta_S^w(\mathbf{w}, \mathbf{v}),$$

where we refer to

$$\triangle_S^w(\mathbf{w}, \mathbf{v}) = \sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}\big[F_S(\mathbf{w}, \mathbf{v}')\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}\big[F_S(\mathbf{w}', \mathbf{v})\big]$$

as the weak PD optimization error of the model $(\mathbf{w}, \mathbf{v})$, and refer to $\triangle^w(\mathbf{w}, \mathbf{v}) - \triangle_S^w(\mathbf{w}, \mathbf{v})$ as the weak PD generalization gap. The generalization gap quantifies the difference between training and testing for the output model, while the optimization error measures the training suboptimality

of the output model as compared to the best model. We will use algorithmic stability to study the generalization gap and use optimization theory to study the optimization error.

We now introduce some definitions and assumptions. Denote by $\|\cdot\|_2$ the Euclidean norm, and $\langle\cdot,\cdot\rangle$ the dot product in $\mathbb{R}^d$.

Definition 2. A function $g : \mathcal{W} \mapsto \mathbb{R}$ is said to be convex if for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $g(\mathbf{w}) \geq g(\mathbf{w}') + \langle \mathbf{w} - \mathbf{w}', \nabla g(\mathbf{w}')\rangle$, where ∇ is the gradient operator. We say g is concave if $-g$ is convex.

Definition 3 (Convexity-concavity)**.** Let $g : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$. We say g is convex-concave if for any $\mathbf{v} \in \mathcal{V}$, the map $\mathbf{w} \mapsto g(\mathbf{w}, \mathbf{v})$ is convex and for any $\mathbf{w} \in \mathcal{W}$, the map $\mathbf{v} \mapsto g(\mathbf{w}, \mathbf{v})$ is concave.

Our analysis relies on an L-smoothness assumption and a bounded second moment assumption. A related assumption is the bounded gradient (also called the Lipschitzness) assumption, meaning that $\|\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq G$ and $\|\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq G$ for all $z \in \mathcal{Z}, \mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$. This assumption was widely used in the generalization analysis of minimax optimization problems [5, 18, 30, 59]. We relax this assumption to a bounded second moment assumption.

Assumption 1. Let $L > 0$. For any z, we assume the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is L-smooth, i.e., the following inequality holds for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}, \mathbf{v}, \mathbf{v}' \in \mathcal{V}$ and $z \in \mathcal{Z}$

$$\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}'; z) \\ \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}'; z) \end{pmatrix} \right\|_2 \leq L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2.$$

Assumption 2. For any $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$, we impose the following bounded second moment condition

$$\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(A_{\mathbf{w}}(S), \mathbf{v}; z_r)\|_2^2] \leq G_{S,A}^2, \quad \mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\mathbf{w}, A_{\mathbf{v}}(S); z_r)\|_2^2] \leq G_{S,A}^2,$$

where r follows the uniform distribution over $[n]$, $G_{S,A} > 0$ is a constant depending on the dataset S and algorithm A, and $\mathbb{E}[G_{S,A}^2] < \infty$.

4 Stability & Generalization of Biased SGDA

4.1 Connecting Stability and Generalization

Algorithmic stability is a fundamental concept for studying the generalization gap. We say two datasets are neighboring datasets if they differ by a single example.

Definition 4. Let $S = \{z_1, \ldots, z_n\}, S' = \{z_1', \ldots, z_n'\}$ be drawn independently from $\mathbb{P}$. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$ as the neighboring dataset by replacing the i-th element of S with z_i'.

Our first result is a connection between generalization gap and algorithmic stability. We use the argument stability [30] to quantify the sensitivity of an algorithm w.r.t. the perturbation of the dataset.

Definition 5 (Algorithmic Stability)**.** Let S and $S^{(i)}$ be defined in Definition 4. Let A be a randomized algorithm and $\epsilon > 0$. We say A is ϵ-argument-stable in expectation if

$$\mathbb{E}_{S,S',A}\left[\frac{1}{n}\sum_{i=1}^{n} \left\| \begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S^{(i)}) \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S^{(i)}) \end{pmatrix} \right\|_2^2\right] \leq \epsilon^2.$$

Theorem 1. *Let A be a randomized algorithm and $\epsilon > 0$. If A is ϵ-argument-stable in expectation and Assumptions 1, 2 hold, then the weak PD generalization error of $(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ satisfies*

$$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle_S^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq L\epsilon^2/2 + \sqrt{2\mathbb{E}[G_{S,A}^2]}\epsilon.$$

Remark 1. To our best knowledge, all existing generalization analyses of minimax problems impose a G-Lipschitzness assumption [5, 18, 30, 59], which is also used in convergence analyses of biased-SGDA. By using this assumption, the work [30] showed that

$$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle_S^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq G\epsilon. \tag{4.1}$$

In comparison, we relax this bounded gradient assumption to the bounded second moment assumption (Assumption 2). Instead, we also impose a smoothness assumption to achieve this relaxation from the Lipschitzness assumption to the bounded second moment assumption. This additional assumption is not restrictive since stability analysis often imposes a smoothness assumption to get effective stability bounds [23]. We are mostly interested in the case with $\epsilon \lesssim 1$[3]. Then, the bound in Theorem 1 is dominated by the second term $O(\sqrt{\mathbb{E}[G_{S,A}^2]}\epsilon)$, which is of the same order as in Eq. (4.1).

[3]We say $A \lesssim B$ if there exists a universal constant $C > 0$ such that $A \leq BC$. We denote $A \asymp B$ if $A \lesssim B$ and $B \lesssim A$.

4.2 Biased-SGDA: Convex-Concave Case

We are interested in a general class of SGDA with biased gradient estimates $g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ for convex-concave minimax optimization problems (i.e., $\mathbb{E}[g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})|\mathbf{w}_t, \mathbf{v}_t, S]$ may not be equal to $\nabla F_S(\mathbf{w}_t, \mathbf{v}_t)$), where $g : \mathcal{W} \times \mathcal{V} \times \mathbb{R}^d \to \mathbb{R}^{2d}$, $\{(\mathbf{w}_t, \mathbf{v}_t)\}_t$ is a model sequence and i_t is drawn from the uniform distribution over $[n] := \{1, \dots, n\}$, for any $t \in [T]$. We consider a generalized bias, where the bias is measured w.r.t. surrogate loss functions $\tilde{f}_1, \tilde{f}_2 : \mathcal{W} \times \mathcal{V} \times \mathcal{Z} \mapsto \mathbb{R}$. The intuition is that while $g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ may be a biased estimator of $\nabla F_S(\mathbf{w}_t, \mathbf{v}_t)$, it can be an unbiased estimator of $\nabla \widetilde{F}_S(\mathbf{w}_t, \mathbf{v}_t) = (\nabla_{\mathbf{w}} \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t), \nabla_{\mathbf{v}} \widetilde{F}_{2,S}(\mathbf{w}_t, \mathbf{v}_t))^\top$, where we define $\widetilde{F}_{1,S}, \widetilde{F}_{2,S}$ as follows

$$\widetilde{F}_{j,S}(\mathbf{w}, \mathbf{v}) := \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_j(\mathbf{w}, \mathbf{v}; z_i), \quad j = 1, 2. \tag{4.2}$$

A closely related construction can be found in the work [56], where they use the surrogate loss to analyze the stability of biased SGD for minimization problems. We extend this idea to the minimax setting through two surrogate losses $\tilde{f}_1(\mathbf{w}, \mathbf{v}; z)$ and $\tilde{f}_2(\mathbf{w}, \mathbf{v}; z)$, where $\tilde{f}_1$ and $\tilde{f}_2$ can be different. As we will show later, this flexibility works well for Zeroth-order SGDA to get 0 bias, where the smoothing parameters can be different in the primal and dual variables.

Definition 6 (Biased-SGDA)**.** Let $\mathbf{w}_1 = 0 \in \mathcal{W}$ and $\mathbf{v}_1 = 0 \in \mathcal{V}$ be the initial point. Let $\text{Proj}_{\mathcal{W}}(\cdot)$ and $\text{Proj}_{\mathcal{V}}(\cdot)$ denote the projection onto $\mathcal{W}$ and $\mathcal{V}$, respectively. Let $\{\eta_t\}_t$ be a positive stepsize sequence. At each iteration, we randomly draw i_t from the uniform distribution over $[n] := \{1, 2, \dots, n\}$ and update

$$\begin{cases} \mathbf{w}_{t+1} = \text{Proj}_{\mathcal{W}}(\mathbf{w}_t - \eta_t g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})), \\ \mathbf{v}_{t+1} = \text{Proj}_{\mathcal{V}}(\mathbf{v}_t + \eta_t g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})), \end{cases} \tag{4.3}$$

where $g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) = (g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}), g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}))^\top$ is a (biased) estimator of $\nabla \widetilde{F}_S(\mathbf{w}_t, \mathbf{v}_t) = (\nabla_{\mathbf{w}} \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t), \nabla_{\mathbf{v}} \widetilde{F}_{2,S}(\mathbf{w}_t, \mathbf{v}_t))^\top$.

Define $z_{i_t}^{(i)} = z_{i_t}$ if $i_t \neq i$ and $z_{i_t}^{(i)} = z_i'$ if $i_t = i$. We apply the same algorithm to $S^{(i)}$ and update

$$\begin{cases} \mathbf{w}_{t+1}^{(i)} = \text{Proj}_{\mathcal{W}}(\mathbf{w}_t^{(i)} - \eta_t g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)})), \\ \mathbf{v}_{t+1}^{(i)} = \text{Proj}_{\mathcal{V}}(\mathbf{v}_t^{(i)} + \eta_t g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)})). \end{cases} \tag{4.4}$$

Then, we define the bias $b_t = (b_{t,\mathbf{w}}, b_{t,\mathbf{v}})$ as follows

$$\begin{cases} b_t = \mathbb{E}_t[g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})] - \nabla \widetilde{F}_S(\mathbf{w}_t, \mathbf{v}_t), \\ b_t^{(i)} = \mathbb{E}_t[g(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)})] - \nabla \widetilde{F}_{S^{(i)}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}), \end{cases} \tag{4.5}$$

where $\mathbb{E}_t := \mathbb{E}\big[\cdot \,\big|\mathbf{w}_t, \mathbf{w}_t^{(j)}, \mathbf{v}_t, \mathbf{v}_t^{(j)}, S, S', j \in [n]\big]$.

Our main result is a general stability bound of Biased-SGDA for convex-concave problems. To this aim, we impose some *generalized Lipschitz-type* assumptions as shown in Eqs. (4.6), (4.7) and (4.8). Eq. (4.6) and Eq. (4.7) control the change rate of the gradient estimate and bias versus the change of the arguments, respectively. Eq. (4.8) controls the change rate of the surrogate function's gradient. As we will show, these generalized Lipschitz-type assumptions hold for various Biased-SGDA methods.

Theorem 2. *Assume for all z, the map $(\mathbf{w}, \mathbf{v}) \mapsto \tilde{f}_j(\mathbf{w}, \mathbf{v}; z)$ is convex-concave, $j = 1, 2$. Let A be Biased-SGDA* (4.3) *with t iterations. Assume there exist $A, \bar{A} \geq 0$, $B, \bar{B} \geq 0$ and $\widetilde{G} \geq 0$ such that*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix}\right\|_2^2\right] \leq \mathbb{E}\left[\frac{A}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix}\right\|_2^2 + B\right], \tag{4.6}$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} b_{t,\mathbf{w}} - b_{t,\mathbf{w}}^{(i)} \\ b_{t,\mathbf{v}}^{(i)} - b_{t,\mathbf{v}} \end{pmatrix}\right\|_2^2\right] \leq \mathbb{E}\left[\frac{\bar{A}}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix}\right\|_2^2 + \bar{B}\right], \tag{4.7}$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} \nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') \\ \nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}_t, \mathbf{v}_t; z_i) \end{pmatrix}\right\|_2^2\right] \leq \widetilde{G}^2. \tag{4.8}$$

If $\eta_t = \eta$ with $A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta \leq 1/T$, then A is $\big(2\eta\sqrt{BT} + 4\big(\sqrt{\bar{B}} + \frac{\widetilde{G}}{n}\big)\eta T\big)$-argument-stable in expectation.

Remark 2. A notable property of Theorem 2 is that it applies to a rather general biased gradient descent ascent method, with the bias defined w.r.t. a surrogate function. Note that we do not need to compute the surrogate function in the implementation of the algorithm. To study the stability of any specific algorithm, one just needs to verify the generalized Lipschitzness condition, and plug the corresponding $A, B, \bar{A}, \bar{B}, \widetilde{G}$ into Theorem 2 to immediately get the stability bounds. The assumption $A\eta^2 + 2\bar{A} \leq \frac{1}{T}$ holds if we choose a sufficiently small stepsize $\eta \lesssim 1/\sqrt{AT}$ and $\bar{A} \lesssim 1/T$, which, as we will see, is the case for SGDA, Zeroth-order SGDA, and Clipped-SGDA. It is clear that the stability bound involves both B and $\bar{B}$. The factor in front of $\sqrt{B}$ is $\eta\sqrt{T}$, while the factor in front of $\sqrt{\bar{B}}$ is ηT. Therefore, the bias has larger impacts (as compared to gradient estimates) on the stability. To get meaningful stability bounds, we require small B and $\bar{B}$, which is the case for Zeroth-order SGDA and Clipped-SGDA. In particular, as we will show, we can set $\bar{B} = 0$ for zeroth-order SGDA by developing an appropriate surrogate function.

5 Applications

This section shows the effectiveness of our general stability bound in Section 4.2 by applying it to SGDA, Zeroth-order SGDA, and Clipped-SGDA. Before presenting our stability bounds of these specific algorithms, we consider another bounded second moment assumption, which is stronger than Assumption 2 (see by substituting $\mathbf{w} = A_{\mathbf{w}}(S)$, $\mathbf{v} = A_{\mathbf{v}}(S)$). Assumption 3 is still weaker than the Lipschitzness assumption: for $f(w; z) = zw$ with $z \sim \mathcal{N}(0, 1)$, one may take $G_S^2 = \frac{1}{n}\sum_{i=1}^n z_i^2$ with $\mathbb{E}[G_S^2] = 1$, while $\|\nabla f\|$ is unbounded.

Assumption 3. Let r follow from the uniform distribution over $[n]$. Suppose for any $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$, there exists a data-dependent constant G_S such that ($\mathbb{E}[G_S^2] < \infty$)

$$\max\left\{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z_r)\|_2^2], \mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z_r)\|_2^2]\right\} \leq G_S^2.$$

Define $\bar{\mathbf{w}}_T = \sum_{t=1}^T(\eta_t \mathbf{w}_t)/\sum_{t=1}^T \eta_t$, $\bar{\mathbf{v}}_T = \sum_{t=1}^T(\eta_t \mathbf{v}_t)/\sum_{t=1}^T \eta_t$.

Table 1: Excess risk bounds of Zeroth-order SGDA and Clipped-SGDA for convex-concave and smooth problems. For brevity, we assume $\eta_t = \eta$ and indicate the dependency on n, L, G, R. For Zeroth-order SGDA, we take $K \asymp d$ and a sufficiently small μ, i.e., $\mu = (\mu_1^2 + \mu_2^2)^{\frac{1}{2}} \lesssim \frac{\sqrt{\mathbb{E}[G_S^2]}}{L\sqrt{n}}$. We take $\tau_1, \tau_2 \asymp L^{\frac{1}{2}} G T^{\frac{1}{2}}$ for Clipped-SGDA under Assumption 3 with $G_S \leq G$ for some $G > 0$.

Algorithm	Reference	Step size	Iteration number	Excess risk bound
Zeroth-order SGDA	Remark 7	$\eta \asymp 1/(L\sqrt{T})$	$T \asymp n$	$\frac{\mathbb{E}[G_S^2]/L + LR^2}{\sqrt{n}}$
Clipped-SGDA	Remark 9	$\eta \asymp 1/(L\sqrt{T})$	$T \asymp n$	$\frac{G^2/L + LR^2 + G^2/L^2}{\sqrt{n}}$

5.1 SGDA

We first consider SGDA with no bias, and our aim is to verify the effectiveness of our general result by recovering the existing results on the classical SGDA in Lei et al. [30]. For SGDA, $\mathbf{w}_{t+1}, \mathbf{v}_{t+1}$ are produced by Eq. (4.3) with

$$\begin{cases} g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) = \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}), \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) = \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}). \end{cases} \tag{5.1}$$

It is clear $g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ is an unbiased estimate of $\nabla F_S(\mathbf{w}_t, \mathbf{v}_t)$ and then we have $b_t = 0$ by setting $\tilde{f}_1 = \tilde{f}_2 = f$.

Theorem 3 (Stability of SGDA)**.** *Assume $\forall z, (\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave . Let Assumption 1, 3 hold. If $L^2\eta^2 \leq \frac{1}{T}$, then SGDA is $\frac{4\sqrt{2\mathbb{E}[G_S^2]}}{\sqrt{n}}\eta\big(\sqrt{T} + 2T/\sqrt{n}\big)$-argument-stable in expectation.*

Remark 3. Under a $\bar{G}$-Lipschitzness assumption (i.e., $\|\nabla_{\mathbf{w}} f(\mathbf{w},\mathbf{v};z)\|_2 \le \bar{G}, \|\nabla_{\mathbf{v}} f(\mathbf{w},\mathbf{v};z)\|_2 \le \bar{G}$ for all z), Lei et al. [30] shows that SGDA is $\sqrt{8e(1+\frac{\bar{G}}{n})}\bar{G}\exp\left(\frac{1}{2}L^2\eta^2 T\right)\frac{\eta\sqrt{T}}{\sqrt{n}}$-argument-stable in expectation. In comparison, we relax the Lipschitzness assumption in Lei et al. [30] to the bounded second moment assumption (Assumption 3). If $n \asymp T$ and $L^2\eta^2 \le 1/T$, both our stability bound and the existing bound [30] are of order $\eta\sqrt{T/n}$ (ignoring the dependency on $G_S, \bar{G}$). Then, under a milder assumption, our general analysis for biased stochastic gradient methods recovers the existing stability bounds developed explicitly for SGDA.

5.2 Zeroth-order SGDA

We now consider Zeroth-order SGDA [6, 53], which is a computationally efficient method for applications where explicit gradient is computationally infeasible, expensive, or impossible to compute. For example, this is the case for generating adversarial examples with black-box deep neural networks to test the adversarial robustness [11], and also for solving automated machine learning tasks where the gradients w.r.t pipeline configuration parameters are infeasible to compute [52]. The basic idea is to approximate the first-order gradients based only on function values.

In particular, for Zeroth-order SGDA, $\mathbf{w}_{t+1}, \mathbf{v}_{t+1}$ are produced by Eq. (4.3) with ($\mathbf{u}_{1,t,l}, \mathbf{u}_{2,t,l} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ denotes the uniform distribution over the unit sphere in $\mathbb{R}^d$)

$$\begin{cases} g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) = \frac{1}{K}\sum_{l=1}^{K}\frac{f(\mathbf{w}_t+\mu_1\mathbf{u}_{1,t,l},\mathbf{v}_t;z_{i_t})-f(\mathbf{w}_t-\mu_1\mathbf{u}_{1,t,l},\mathbf{v}_t;z_{i_t})}{2\mu_1/d}\mathbf{u}_{1,t,l}, \\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) = \frac{1}{K}\sum_{l=1}^{K}\frac{f(\mathbf{w}_t,\mathbf{v}_t+\mu_2\mathbf{u}_{2,t,l};z_{i_t})-f(\mathbf{w}_t,\mathbf{v}_t-\mu_2\mathbf{u}_{2,t,l};z_{i_t})}{2\mu_2/d}\mathbf{u}_{2,t,l}, \end{cases} \tag{5.2}$$

where $\mu_1, \mu_2 \in \mathbb{R}^+$ are smoothing parameters determining the step size in approximating the gradient of f. For simplicity, we omit the dependency on $\mathbf{u}_{1,t,l}$ and $\mathbf{u}_{2,t,l}$ in the notation of $g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})$ and $g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})$. We also denote $\mu = (\mu_1^2+\mu_2^2)^{1/2}$ and assume $d \ge 2$. The following lemma justifies the effectiveness of the surrogate function. The first statement show that the surrogate functions preserve the smoothness, convexity-concavity, and bounded second moment condition (Assumption 3). While $g_{\mathbf{w}}(\mathbf{w},\mathbf{v};z)$ and $g_{\mathbf{v}}(\mathbf{w},\mathbf{v};z)$ are biased estimates of $\nabla_{\mathbf{w}}F_S(\mathbf{w},\mathbf{v})$ and $\nabla_{\mathbf{v}}F_S(\mathbf{w},\mathbf{v})$, the second statement shows that they are unbiased gradient estimates of the surrogate functions. Then, the bias defined in Eq. (4.5) is 0. This shows the benefit of surrogate functions in the stability analyses. Proofs are given in Section B.3.2.

Lemma 4. *Consider the surrogate functions*

$$\tilde{f}_1(\mathbf{w},\mathbf{v};z) = \mathbb{E}_{\mathbf{u}}\big[f(\mathbf{w}+\mu_1\mathbf{u},\mathbf{v};z)\big], \quad \tilde{f}_2(\mathbf{w},\mathbf{v};z) = \mathbb{E}_{\mathbf{u}}\big[f(\mathbf{w},\mathbf{v}+\mu_2\mathbf{u};z)\big], \tag{5.3}$$

where $\mathbf{u} \sim \mathrm{Unif}(\mathbb{B}^d)$ *with* $\mathrm{Unif}(\mathbb{B}^d)$ *denoting the uniform distribution over the unit ball in* $\mathbb{R}^d$*, and we omit* μ_1, μ_2 *in* $\tilde{f}_1, \tilde{f}_2$ *for brevity.*

- *Assume for any* z, $(\mathbf{w},\mathbf{v}) \mapsto f(\mathbf{w},\mathbf{v};z)$ *is convex-concave, smooth, and satisfies Assumption 3. Then for any* z, $(\mathbf{w},\mathbf{v}) \mapsto \tilde{f}_1(\mathbf{w},\mathbf{v};z)$ *and* $(\mathbf{w},\mathbf{v}) \mapsto \tilde{f}_2(\mathbf{w},\mathbf{v};z)$ *are convex-concave, smooth, and satisfy Assumption 3.*
- *Let* g *be defined in Eq.* (5.2)*. Then, for any* $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$, $\mathbb{E}_{\mathbf{u}}[g_{\mathbf{w}}(\mathbf{w},\mathbf{v};z)] = \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w},\mathbf{v};z)$ *and* $\mathbb{E}_{\mathbf{u}}[g_{\mathbf{v}}(\mathbf{w},\mathbf{v};z)] = \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w},\mathbf{v};z)$.

To apply Theorem 2, we will also show that Eq. (4.6) and Eq. (4.8) hold with appropriate A, B and $\widetilde{G}$, which are constructed in Lemma 21 and Lemma 18, proved in Section B.3.2. We then apply Lemma 21 and Lemma 18 to Theorem 2 to derive the stability bounds for Zeroth-order SGDA.

Theorem 5 (Stability of Zeroth-order SGDA)**.** *Assume for all* z*, the map* $(\mathbf{w},\mathbf{v}) \mapsto f(\mathbf{w},\mathbf{v};z)$ *is convex-concave. Let Assumptions 1, 3 hold. If* $L^2(1+\frac{4d}{K})\eta^2 \le \frac{1}{T}$*, Zeroth-order SGDA is* ϵ*-argument-stable in expectation with*

$$\epsilon \lesssim \Big(\frac{\sqrt{\mathbb{E}[G_S^2]}}{\sqrt{n}}\Big(1+\frac{d^{\frac{1}{2}}}{K^{\frac{1}{2}}}\Big)\Big)\eta\sqrt{T} + \frac{\sqrt{\mathbb{E}[G_S^2]}}{n}\eta T.$$

Remark 4 (Comparison with SGDA). If we choose $K \asymp d$, then the conditions in Theorem 5 hold with the standard choice $\eta \lesssim 1/(L\sqrt{T})$. The stability bound of Zeroth-order SGDA is of the same order as the stability bound for SGDA in Theorem 3 in this case, which is also verified in our experiments (Section A).

Remark 5 (Parameter choices). The works [34, 41] study Zeroth-order SGD with a fast-decaying step size $\eta_t \lesssim 1/t$ and require the smoothing parameter $\mu \lesssim 1/(dn)$. More recently, [56] improves their results with a milder condition $\eta \lesssim 1/\sqrt{T}$, but still requires $\mu \lesssim 1/(d\sqrt{n})$. In comparison, we conduct the first stability analysis for Zeroth-order SGDA under the same mild condition $\eta \lesssim 1/\sqrt{T}$, and do not need any requirement for μ. We give the detailed comparison in Table 2.

Table 2: Stability Analysis for Zeroth-order Methods, G denotes the bounded gradient assumption and G_S denotes the bounded second moment assumption (Assumption 3), L denotes the smoothness assumption (Liu et al. [34], Nikolakakis et al. [41], Zeng and Lei [56] consider Zeroth-order SGD). We ignore all the factors independent of d, K, n, T, η.

Analysis	η	μ	Assumption	Bounds
Stability in [41]	$O(\frac{1}{t})$	$O(\frac{1}{d^{3/2}n})$	L, G	$O(\frac{T}{n})$
Stability in [34]	$O(\frac{1}{t})$	$O(\frac{1}{dn})$	L, G	$O(\frac{T}{n})$
Stability in [56]	$O(\frac{1}{\sqrt{T}})$	$O(\frac{1}{d\sqrt{n}})$	L	$O(\frac{(1+\sqrt{d/K})\eta\sqrt{T}}{\sqrt{n}} + \frac{\eta T}{n})$
Our stability	$O(\frac{1}{\sqrt{T}})$	No requirement	L, G_S	$O(\frac{(1+\sqrt{d/K})\eta\sqrt{T}}{\sqrt{n}} + \frac{\eta T}{n})$

We now bound optimization errors for Zeroth-order SGDA with $\mathcal{W} = \mathcal{V} = B_R$.

Lemma 6 (Optimization of Zeroth-order SGDA)**.** *Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Let Assumption 3 hold. If we choose the gradient estimator in Eq.* (5.2)*, then*

$$\mathbb{E}[\sup_{\mathbf{v}\in B_R} F_S(\bar{\mathbf{w}}_T, \mathbf{v}) - \inf_{\mathbf{w}\in B_R} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)] \lesssim \frac{R^2}{\eta T} + (1 + \frac{d}{K})\mathbb{E}[G_S^2]\eta + \mu\sqrt{\mathbb{E}[G_S^2]}.$$

Remark 6 (Comparison). Our optimization error recovers the results in existing works on the convergence analysis of Zeroth-order methods [6, 17, 39, 48]. Furthermore, we show that when $K \asymp d$, an optimization error of order $1/\sqrt{T}$ is achievable under dimension-independent choices of $\eta \lesssim 1/\sqrt{T}, \mu \lesssim 1/\sqrt{T}$, while the existing works require dimension-dependent parameters η, μ.

We now combine Theorem 5 and Lemma 6 to derive excess risk bounds for projected Zeroth-order SGDA on B_R.

Theorem 7 (Excess risk of Zeroth-order SGDA)**.** *Let assumptions and definitions in Theorem 1, Theorem 5 and Lemma 6 hold. Assume*

$$\frac{1}{L\sqrt{n}} + \frac{\eta T}{n} \lesssim \frac{1}{L}. \tag{5.4}$$

For $R > 0$, let $\triangle^{w,R}(\mathbf{w}, \mathbf{v}) := \sup_{\mathbf{v}'\in B_R} \mathbb{E}[F(\mathbf{w}, \mathbf{v}')] - \inf_{\mathbf{w}'\in B_R} \mathbb{E}[F(\mathbf{w}', \mathbf{v})]$. Then,

$$\Delta^{w,R}(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \lesssim \frac{\mathbb{E}[G_S^2]/L}{\sqrt{n}} + \frac{\mathbb{E}[G_S^2]}{n}\eta T + \frac{R^2}{\eta T} + \frac{\mathbb{E}[G_S^2]}{L^2\eta T} + \mu\sqrt{\mathbb{E}[G_S^2]}. \tag{5.5}$$

Remark 7 (Minimax-optimal excess risk). With our choice of $\mu \lesssim \frac{\sqrt{\mathbb{E}[G_S^2]}}{L\sqrt{n}}$, Eq. (5.5) implies

$$\Delta^{w,R}(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \lesssim \frac{\mathbb{E}[G_S^2]/L}{\sqrt{n}} + \frac{\mathbb{E}[G_S^2]}{n}\eta T + \frac{R^2}{\eta T} + \frac{\mathbb{E}[G_S^2]}{L^2\eta T}.$$

By choosing $\eta \asymp 1/(L\sqrt{T})$ with a sufficiently small universal constant and $T \asymp n$, this further yields

$$\Delta^{w,R}(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \lesssim \frac{\mathbb{E}[G_S^2]/L + LR^2}{\sqrt{n}}.$$

That is, our analysis gives an excess risk bound of order $1/\sqrt{n}$, which has the minimax-optimal dependence on the sample size n. To our knowledge, this gives the first optimal risk bounds for Zeroth-order SGDA. It is clear that Eq. (5.4) holds under our choices of η and T. Moreover, if we take $K \asymp d$, then our parameter choice shows $L^2(1 + \frac{d}{K})\eta^2 \lesssim \frac{1}{T}$ which is consistent with the condition $L^2(1 + \frac{4d}{K})\eta^2 \leq \frac{1}{T}$ required in Theorem 5.

Table 3: Excess Risk Analysis for Zeroth-order Methods, G_S denotes the bounded second moment assumption, and L denotes the smoothness assumption (Zeng and Lei [56] consider Zeroth-order SGD).

Analysis	η	μ	Assumption	Bounds
Excess risk in [56]	$O\left(\frac{\sqrt{K/d}}{\sqrt{n}}\right)$	$O(\frac{1}{d\sqrt{n}})$	L	$O(n^{-1/2})$
Our excess risk bound	$O\left(\frac{\sqrt{K/d}}{\sqrt{n}}\right)$	$O(\frac{1}{\sqrt{n}})$	L, G_S	$O(n^{-1/2})$

5.3 Clipped-SGDA

Gradient clipping is known to alleviate the problem of exploding gradients in recurrent neural networks. Clipped-SGDA [22] applies the clipping operator $\mathrm{clip}(\mathbf{v}, \tau) := \mathbf{v} \min(1, \tau/\|\mathbf{v}\|_2)$ with the clipping parameter τ to the gradient estimator. Specifically, for Clipped-SGDA, $\mathbf{w}_{t+1}, \mathbf{v}_{t+1}$ are produced by Eq. (4.3) with

$$\begin{cases} g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) = \mathrm{clip}(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}), \tau_1), \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) = \mathrm{clip}(\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}), \tau_2), \end{cases} \tag{5.6}$$

where $\tau_1, \tau_2 > 0$ are clipping parameters.

By clipping gradients, Clipped-SGDA improves the robustness and has found successful applications, e.g., training GPT models [9] and differentially private algorithms [1]. To apply our general results to Clipped-SGDA, we verify Eq. (4.6) and Eq. (4.7) in Lemma 25 and Lemma 30, proved in Section B.3.3. We always assume $\tilde{f}_1 = \tilde{f}_2 = f$, by which we can easily show that Eq. (4.8) holds with $\widetilde{G} = 2\sqrt{2\mathbb{E}[G_S^2]}$. We then apply Lemma 25 and Lemma 30 to Theorem 2, and get the stability bounds for Clipped-SGDA.

Theorem 8 (Stability of Clippeed-SGDA)**.** *Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Let Assumptions 1, 3 hold with $G_S \leq G$ for some $G > 0$. For η, τ_1, τ_2 satisfying $L^2\eta^2 + 4\sqrt{2}LG^2(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2})\eta \leq \frac{1}{T}$, Clipped-SGDA is ϵ-argument-stable in expectation with $\epsilon \lesssim G\eta T/n + G\eta\sqrt{T}/\sqrt{n}$.*

Remark 8. If we choose $\tau_1, \tau_2 \gtrsim GT^{\frac{1}{4}}$ and $\eta \lesssim 1/(L\sqrt{T})$, then the conditions in Theorem 8 hold. It is clear that the stability bound of Clipped-SGDA is of the same order as the stability bound for SGDA in Theorem 3, which is also verified in our experiments (Section A).

We now control optimization errors for Clipped-SGDA. Analogous to Lemma 6, we consider projected Clipped-SGDA on B_R.

Lemma 9 (Optimization error for Clipped-SGDA)**.** *Assume for all z, the map $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Let Assumptions 1, 3 hold with $G_S \leq G$ for some $G > 0$. Assume $G \lesssim \tau_1$ and $G \lesssim \tau_2$. If we choose gradient estimator in Eq.* (5.6)*, then*

$$\mathbb{E}\big[\sup_{\mathbf{v} \in B_R} F_S(\bar{\mathbf{w}}_T, \mathbf{v}) - \inf_{\mathbf{w} \in B_R} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big] \lesssim \frac{R^2}{\eta T} + G^2\eta + G^4(\tau_1^{-2} + \tau_2^{-2})\eta T.$$

We now combine Theorem 8 and Lemma 9 to get the following risk bounds for projected Clipped-SGDA on B_R. Recall $\Delta^{w,R}$ as defined in Theorem 7.

Theorem 10. *Let assumptions and definitions in Theorem 1, Theorem 8 and Lemma 9 hold. Then,*

$$\Delta^{w,R}(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \lesssim \frac{LG^2\eta^2T^2}{n^2} + \frac{LG^2\eta^2T}{n} + \frac{G^2\eta T}{n} + \frac{G^2\eta\sqrt{T}}{\sqrt{n}} + \frac{R^2}{\eta T} + G^2\eta + G^4(\tau_1^{-2} + \tau_2^{-2})\eta T.$$

Remark 9 (Minimax optimal excess risk). Setting $\tau_1, \tau_2 \asymp L^{\frac{1}{2}}GT^{\frac{1}{2}}$ similarly to Clipped-SGD [40]), Theorem 10 implies

$$\Delta^{w,R}(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \lesssim \frac{LG^2\eta^2T^2}{n^2} + \frac{LG^2\eta^2T}{n} + \frac{G^2\eta T}{n} + \frac{G^2\eta\sqrt{T}}{\sqrt{n}} + \frac{R^2}{\eta T} + G^2\eta + \frac{G^2}{L}\eta T.$$

By choosing $\eta \asymp 1/(L\sqrt{T})$ and $T \asymp n$, this further yields

$$\Delta^{w,R}(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \lesssim (G^2/L + R^2 L + G^2/L^2)/\sqrt{n}.$$

That is, our analysis gives excess risk rates of order $1/\sqrt{n}$. This shows our generalization analysis of general biased gradient methods implies optimal excess risk bounds for Clipped-SGDA. With our choices of τ_1, τ_2, η, it is clear that $L^2\eta^2 + LG^2\big(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2}\big)\eta \lesssim 1/T$, which is consistent with condition $L^2\eta^2 + 4\sqrt{2}LG^2(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2})\eta \leq \frac{1}{T}$ in Theorem 8.

6 Future work

There are some interesting problems for further study. First, it is interesting to apply our framework to other biased minimax optimization methods such as decentralized SGDA [61]. Second, as we only consider the convex-concave problem, it is interesting to extend our analysis to strongly-convex-strongly-concave [42], nonconvex-strongly-concave [32, 36] and [32, 36, 43] nonconvex-concave problems [54]. Furthermore, we can extend our work to minimax problems with neural networks, which satisfy the ρ-weak-convexity-weak-concavity assumption [30], with the parameter ρ decaying as the order of $1/\sqrt{m}$ (m is the network width) [44].

Impact Statements

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

We are grateful to the anonymous reviewers and area chairs for their insightful and constructive comments. The work of Yunwen Lei was supported by the Hong Kong Research Grants Council under the GRF projects 17305425 and 17302624. The work of Yiming Ying is supported by Australian Research Council Discovery Project (DP250101359).

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[3] A. Ajalloeian and S. U. Stich. On the convergence of SGD with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

[4] R. Bassily, V. Feldman, C. Guzmán, and K. Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, pages 4381–4391, 2020.

[5] R. Bassily, C. Guzmán, and M. Menart. Differentially private algorithms for the stochastic saddle point problem with optimal rates for the strong gap. In *Annual Conference on Learning Theory*, pages 2482–2508. PMLR, 2023.

[6] A. Beznosikov, A. Sadiev, and A. Gasnikov. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 105–119. Springer, 2020.

[7] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2 (Mar):499–526, 2002.

[8] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[10] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[11] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[12] R. S. Chen, B. Lucier, Y. Singer, and V. Syrgkanis. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[13] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[14] Y. Demidovich, G. Malinovsky, I. Sokolov, and P. Richtárik. A guide through the zoo of biased sgd. In *Advances in Neural Information Processing Systems*, volume 36, pages 23158–23171, 2023.

[15] J. Dolbeault, M. J. Esteban, M. Kowalczyk, and M. Loss. Sharp interpolation inequalities on the sphere: new methods and consequences. *Chinese Annals of Mathematics, Series B*, 34(1):99–112, 2013.

[16] D. Driggs, J. Liang, and C.-B. Schönlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022.

[17] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5): 2788–2806, 2015.

[18] F. Farnia and A. Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pages 3174–3185. PMLR, 2021.

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

[20] X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[22] E. Gorbunov, M. Danilova, D. Dobre, P. Dvurechenskii, A. Gasnikov, and G. Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. In *Advances in Neural Information Processing Systems*, volume 35, pages 31319–31332, 2022.

[23] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[24] Y. Hu, S. Zhang, X. Chen, and N. He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 2759–2770, 2020.

[25] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[26] I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[28] M. Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

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

[30] Y. Lei, Z. Yang, T. Yang, and Y. Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.

[31] Y. Lei, Z. Wang, and X. Yuan. Learning theory of stochastic regularized dual averaging. *IEEE Transactions on Information Theory*, 72(10):7709–7730, 2026.

[32] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[33] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021.

[34] X. Liu, H. Zhang, B. Gu, and H. Chen. General stability analysis for zeroth-order optimization algorithms. In *International Conference on Learning Representations*, 2024.

[35] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, December 2015.

[36] L. Luo, H. Ye, Z. Huang, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 20566–20577, 2020.

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[38] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[39] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[40] T. D. Nguyen, T. H. Nguyen, A. Ene, and H. L. Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

[41] K. Nikolakakis, F. Haddadpour, D. Kalogerias, and A. Karbasi. Black-box generalization: Stability of zeroth-order learning. volume 35, pages 31525–31541, 2022.

[42] B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[43] H. Rafique, M. Liu, Q. Lin, and T. Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

[44] D. Richards and I. Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 34, pages 8609–8621, 2021.

[45] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[46] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[47] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[48] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

[49] S. Surendran, A. Fermanian, A. Godichon-Baggioni, and S. Le Corff. Non-asymptotic analysis of biased adaptive stochastic approximation. In *Advances in Neural Information Processing Systems*, volume 37, pages 12897–12943, 2024.

[50] M. Tucat, A. Mukherjee, P. Sen, M. Sun, and O. Rivasplata. Regularized gradient clipping provably trains wide and deep neural networks. *Transactions on Machine Learning Research*, 2025, 2025.

[51] S. Vary, T. Farghly, I. Kuzborskij, and P. Rebeschini. On-Average Stability of Multipass Preconditioned SGD and Effective Dimension. In *Conference on Learning Theory*, pages 6458–6495, 2026.

[52] C. Wang, Q. Wu, M. Weimer, and E. Zhu. Flaml: A fast and lightweight automl library. *Proceedings of Machine Learning and Systems*, 3:434–447, 2021.

[53] Z. Xu, Z.-Q. Wang, J.-L. Wang, and Y.-H. Dai. Zeroth-order alternating gradient descent ascent algorithms for a class of nonconvex-nonconcave minimax problems. *Journal of Machine Learning Research*, 24(313): 1–25, 2023.

[54] J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 1153–1165, 2020.

[55] Y. Ying, L. Wen, and S. Lyu. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[56] S. Zeng and Y. Lei. Stochastic gradient methods: Bias, stability and generalization. *Journal of Machine Learning Research*, 27(6):1–55, 2026.

[57] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pages 15383–15393, 2020.

[58] J. Zhang, M. Hong, M. Wang, and S. Zhang. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR, 2021.

[59] S. Zhang, Y. Hu, L. Zhang, and N. He. Generalization bounds of nonconvex-(strongly)-concave stochastic minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 694–702. PMLR, 2024.

[60] B. Zhu, S. Li, and Y. Liu. Towards sharper risk bounds for minimax problems. In *International Joint Conference on Artificial Intelligence*, pages 5698–5706, 2024.

[61] M. Zhu, L. Shen, B. Du, and D. Tao. Stability and generalization of the decentralized stochastic gradient descent ascent algorithm. In *Advances in Neural Information Processing Systems*, volume 36, pages 31717–31751, 2023.

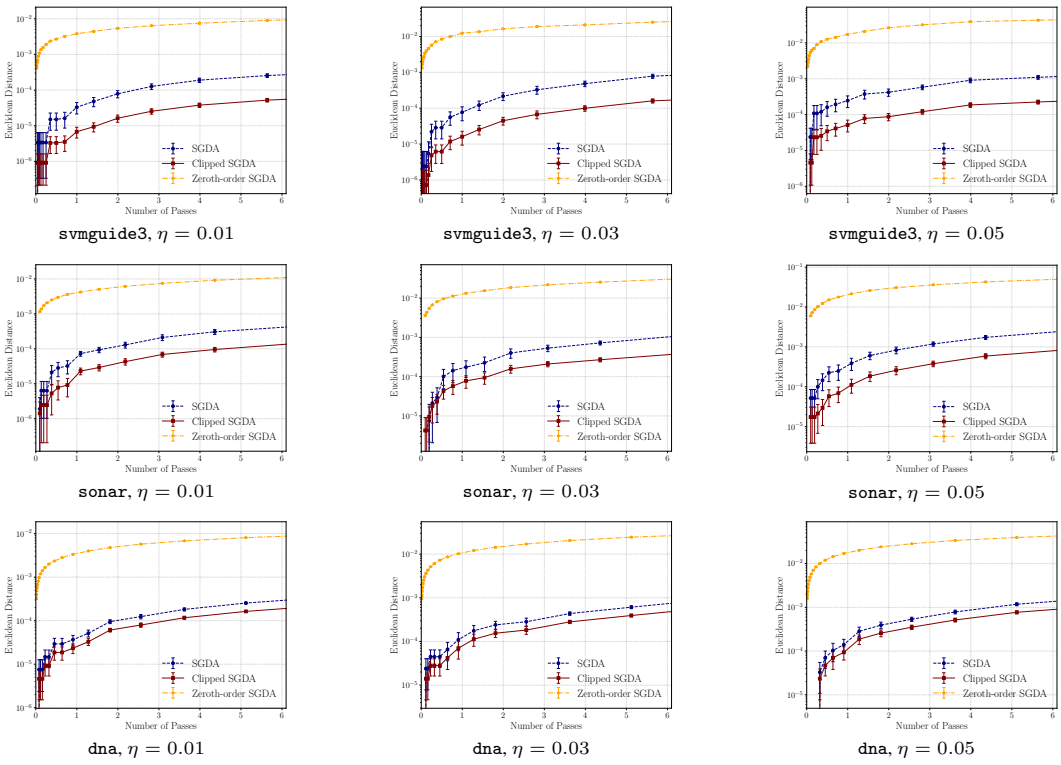
A Experiments

In this section, we report experimental results to verify our theoretical findings. Our code is available here. Our experimental setup follows the work [30]. Specifically, we consider the Area Under ROC Curve (AUC) maximization problem, which can be formulated as a minimax problem [55]:

$$\min_{(\mathbf{w},a,b)\in\mathbb{R}^{d+2}} \max_{\alpha\in\mathbb{R}} \mathbb{E}[f(\mathbf{w},a,b,\alpha;z)],$$

where $f(\mathbf{w},a,b,\alpha;z) = (h(\mathbf{w};x)-a)^2\mathbb{I}_{[y=1]}/p + (h(\mathbf{w};x)-b)^2\mathbb{I}_{[y=-1]}/(1-p) + 2(1-\alpha)h(\mathbf{w};x)(\mathbb{I}_{[y=-1]}/(1-p) - \mathbb{I}_{[y=1]}/p) - \alpha^2$, $h(\mathbf{w};x)$ denotes a scoring function, $p=\mathbb{P}[y=1]$, and $\mathbb{I}_{[\cdot]}$ is the indicator function. We use three datasets from Chang and Lin [10]: `svmguide3`, `sonar`, and `DNA`. For each training dataset S, we build a neighboring dataset S' by changing its last example. We then apply SGDA, Zeroth-order SGDA, and Clipped SGDA to both S and S' to get model sequences $\{(\mathbf{w}_t,\mathbf{v}_t)\}$ and $\{(\mathbf{w}'_t,\mathbf{v}'_t)\}$. We then evaluate the stability and generalization error of these algorithms along the training process by the Euclidean distance $\Delta_t = (\|\mathbf{w}_t-\mathbf{w}'_t\|_2^2 + \|\mathbf{v}_t-\mathbf{v}'_t\|_2^2)^{\frac{1}{2}}$ and Test AUC (AUC on the test dataset).

For each method applied to each data, we use $\eta_t = \eta/\sqrt{T}$ with $\eta\in\{0.01, 0.03, 0.05\}$, repeat the experiments 25 times to ensure reliability, and show results with mean and standard deviation. We choose $\mu_1=\mu_2=0.0001, K=d$ (21 for `svmguide3`, 60 for `sonar` and 180 for `DNA`) defined in Eq. (5.2) for Zeroth-order SGDA, and $\tau_1=\tau_2=0.5$`(svmguide3)`$, 0.1$`(sonar,DNA)` defined in Eq. (5.6) for Clipped-SGDA. We use the number of passes to denote the iteration number divided by the training sample size. As shown in Figure 1, Δ_t of SGDA, Zeroth-order SGDA, and Clipped-SGDA increase along the training process and appear to be comparable (note that Δ_t of Zeroth-order SGDA is slightly larger due to the scalar d in the definition of gradient estimators of Zeroth-order SGDA in Eq. (5.2)), which verifies our theoretical findings in Theorem 5 and Theorem 8. As shown in Figure 2, the generalization errors of SGDA, Zeroth-order SGDA, and Clipped-SGDA also appear to be comparable, which verifies our theoretical findings in Theorem 7 and Theorem 10.

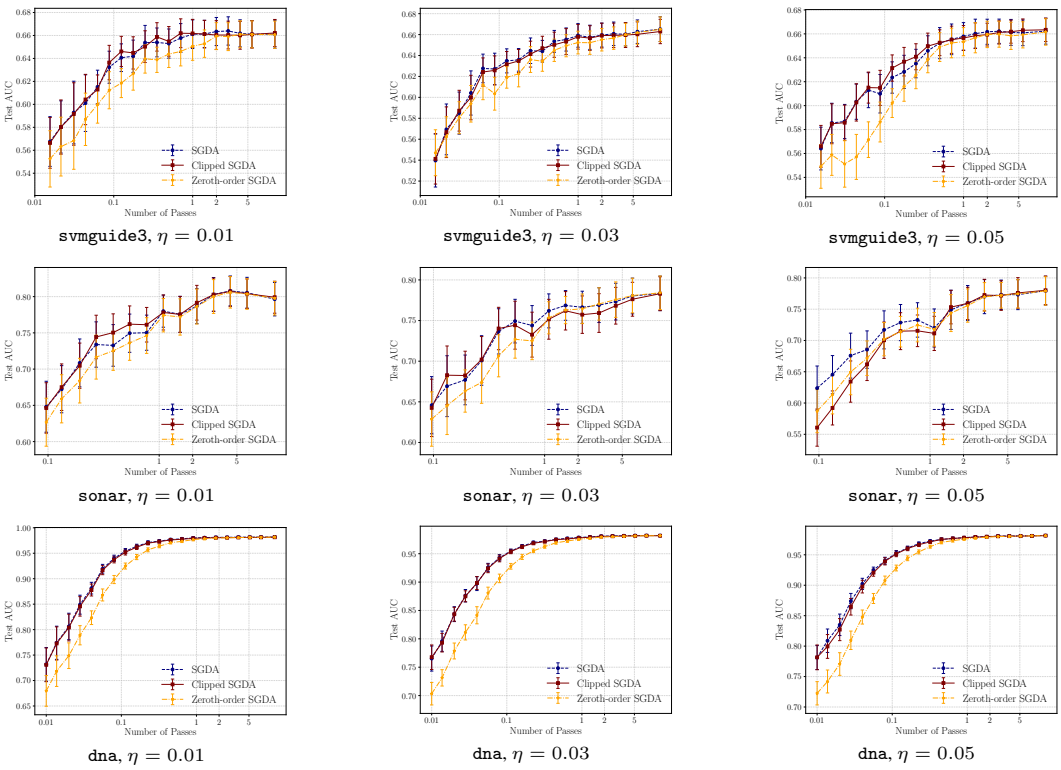


`svmguide3`, $\eta=0.01$ | `svmguide3`, $\eta=0.03$ | `svmguide3`, $\eta=0.05$

`sonar`, $\eta=0.01$ | `sonar`, $\eta=0.03$ | `sonar`, $\eta=0.05$

`dna`, $\eta=0.01$ | `dna`, $\eta=0.03$ | `dna`, $\eta=0.05$

Figure 1: Δ_t versus the number of passes for different datasets (`svmguide3`, `sonar`, `DNA`) and $\eta\in\{0.01, 0.03, 0.05\}$.



svmguide3, $\eta = 0.01$ svmguide3, $\eta = 0.03$ svmguide3, $\eta = 0.05$

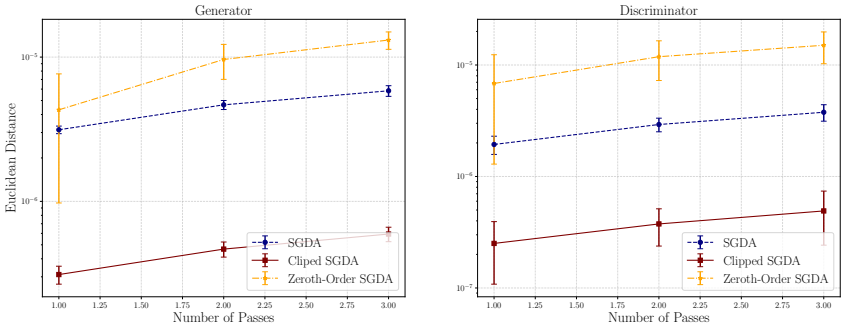
sonar, $\eta = 0.01$ sonar, $\eta = 0.03$ sonar, $\eta = 0.05$

dna, $\eta = 0.01$ dna, $\eta = 0.03$ dna, $\eta = 0.05$

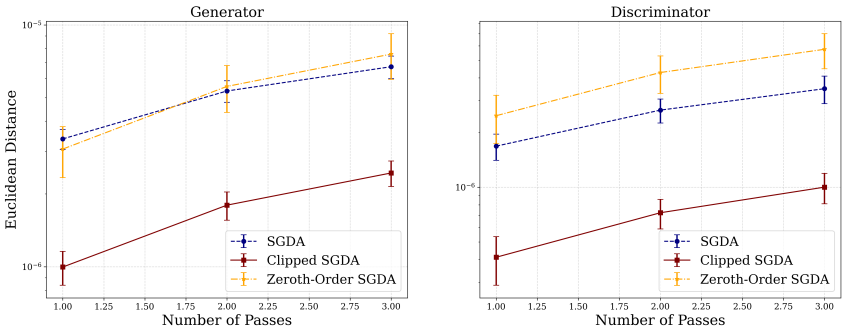
Figure 2: Test AUC versus # of passes for different datasets (`svmguide3`, `sonar`, `dna`) and $\eta \in \{0.01, 0.03, 0.05\}$.

We also consider the stability of biased stochastic gradient methods for a nonconvex-nonconcave problem: GAN [21] on the MNIST [27], CIFAR-10 [25], and CelebA [35] datasets. We conduct the experiments with one NVIDIA GeForce RTX 3090 24GB GPU. We use $\eta_t = 0.0002$, choose $\mu_1 = \mu_2 = 0.0001, K \asymp d$ for Zeroth-order SGDA, and clipping parameter $\tau_1 = \tau_2 = 1$ in Eq. (5.6) for Clipped-SGDA. As shown in Figure 3, Δ_t of SGDA, Zeroth-order SGDA, and Clipped-SGDA increase along the training process and also appear to be comparable. Interestingly, the stability of Clipped-SGDA outperforms that of SGDA in this case.

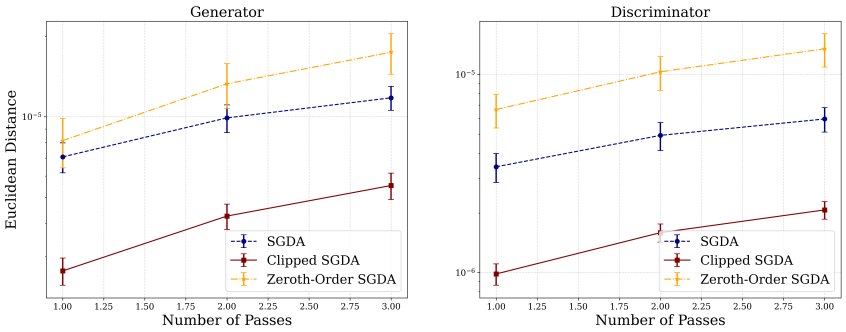
We further study the parameter sensitivity to the smoothing parameter μ and clipping threshold τ. Figure 4 shows that the stability of Zeroth-order SGDA is independent of the choice of μ, and that the generalization error of it matches that of SGDA when $\mu \lesssim \frac{\sqrt{\mathbb{E}[G_S^2]}}{L\sqrt{n}}$, verifying our theoretical findings in Remark 5 and Remark 7. Figure 5 shows that the stability and generalization error of Clipped SGDA match that of SGDA when $\tau_1, \tau_2 \gtrsim GT^{\frac{1}{4}}$ and $\eta \lesssim 1/\sqrt{T}$, verifying our theoretical findings in Remark 8 and Remark 9.



(a) MNIST



(b) CIFAR-10



(c) CelebA

Figure 3: Δ_t versus # of passes. Left: generator, right: discriminator. $\eta = 0.0002$, $K = 300$, $\mu_1 = \mu_2 = 0.0001, \tau_1 = \tau_2 = 1$.

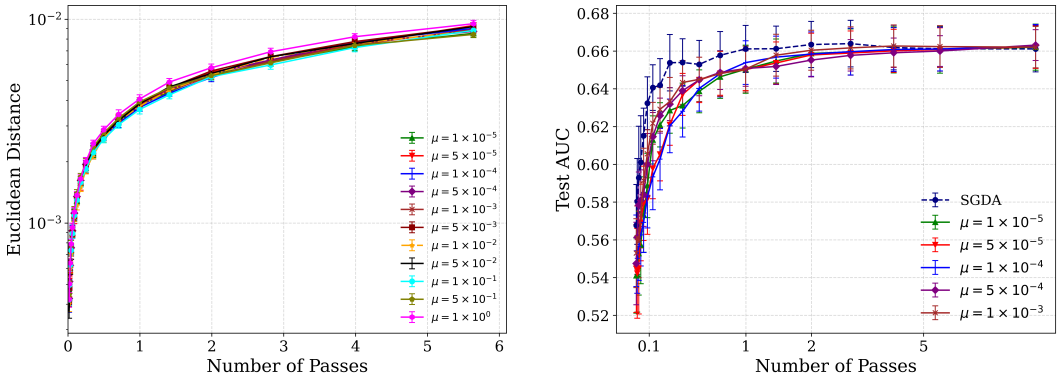
B Proof

B.1 Proof on Connecting Stability and Generalization

In this section, we present the proof of our results on connecting stability and generalization.

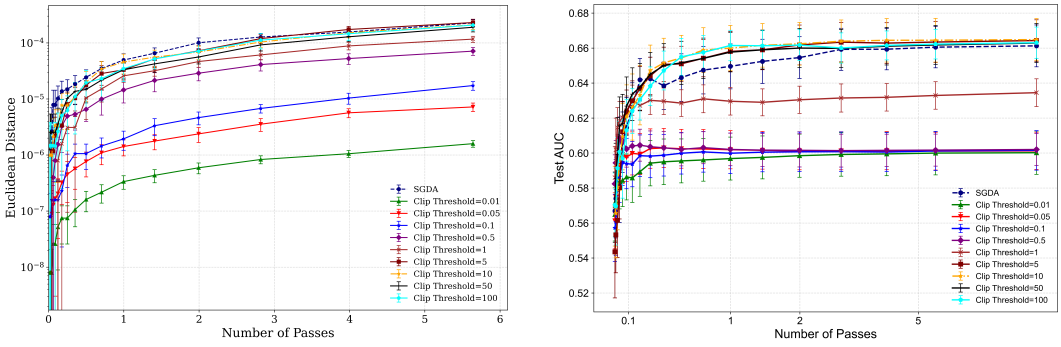
Proof of Theorem 1. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be two datasets drawn from the same distribution. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$. For any function $g, \widetilde{G}$, we have the basic inequalities

$$\begin{aligned} \sup_{\mathbf{w}} g(\mathbf{w}) - \sup_{\mathbf{w}} \widetilde{G}(\mathbf{w}) &\leq \sup_{\mathbf{w}} \big(g(\mathbf{w}) - \widetilde{G}(\mathbf{w})\big) \\ \inf_{\mathbf{w}} g(\mathbf{w}) - \inf_{\mathbf{w}} \widetilde{G}(\mathbf{w}) &\leq \sup_{\mathbf{w}} \big(g(\mathbf{w}) - \widetilde{G}(\mathbf{w})\big). \end{aligned} \tag{B.1}$$



(a) Euclidean distance Δ_t (b) Test AUC

Figure 4: Parameter sensitivity of Zeroth-order SGDA w.r.t. the smoothing parameter μ on `svmguide3`: (a) Δ_t versus # of passes; (b) Test AUC versus # of passes. $\eta = 0.01$.



(a) Euclidean distance Δ_t (b) Test AUC

Figure 5: Parameter sensitivity of Clipped-SGDA w.r.t. the clipping threshold τ on `svmguide3`: (a) Δ_t versus # of passes; (b) Test AUC versus # of passes. $\eta = 0.01$.

It follows from (B.1) that

$$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle^w_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \sup_{\mathbf{v}' \in \mathcal{V}} \mathbb{E}[F(A_{\mathbf{w}}(S), \mathbf{v}') - F_S(A_{\mathbf{w}}(S), \mathbf{v}')] + \sup_{\mathbf{w}' \in \mathcal{W}} \mathbb{E}[F_S(\mathbf{w}', A_{\mathbf{v}}(S)) - F(\mathbf{w}', A_{\mathbf{v}}(S))]. \quad \text{(B.2)}$$

According to the symmetry between z_i and z_i' we know

$$\begin{aligned}\mathbb{E}[F(A_{\mathbf{w}}(S), \mathbf{v}') - F_S(A_{\mathbf{w}}(S), \mathbf{v}')] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[F(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}')] - \mathbb{E}[F_S(A_{\mathbf{w}}(S), \mathbf{v}')] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}'; z_i) - f(A_{\mathbf{w}}(S), \mathbf{v}'; z_i)\big],\end{aligned}$$

where the second identity holds since z_i is not used to train $A_{\mathbf{w}}(S^{(i)})$. In a similar way, we can prove

$$\mathbb{E}[F_S(\mathbf{w}', A_{\mathbf{v}}(S)) - F(\mathbf{w}', A_{\mathbf{v}}(S))] = \frac{1}{n}\sum_{i=1}^n \big[f(\mathbf{w}', A_{\mathbf{v}}(S^{(i)}); z_i) - f(\mathbf{w}', A_{\mathbf{v}}(S); z_i)\big].$$

As a combination of the above three inequalities we get

$$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle^w_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \sup_{\mathbf{v}' \in \mathcal{V}} \Big[\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}'; z_i) - f(A_{\mathbf{w}}(S), \mathbf{v}'; z_i)\big]\Big] + \sup_{\mathbf{w}' \in \mathcal{W}} \Big[\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w}', A_{\mathbf{v}}(S^{(i)}); z_i) - f(\mathbf{w}', A_{\mathbf{v}}(S); z_i)\big]\Big]. \quad \text{(B.3)}$$

According to Assumption 1, we know

$$f(A_{\mathbf{w}}(S^{(i)}),\mathbf{v}';z_i)-f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\leq\left\langle\begin{pmatrix}A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\\ \mathbf{v}'-\mathbf{v}'\end{pmatrix},\begin{pmatrix}\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\\ \nabla_{\mathbf{v}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\end{pmatrix}\right\rangle+$$
$$\frac{L}{2}\left\|\begin{pmatrix}A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\\ \mathbf{v}'-\mathbf{v}'\end{pmatrix}\right\|_2^2=\langle A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S),\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\rangle+\frac{L}{2}\|A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\|_2^2.$$

According to Schwartz's inequality we know for any $\gamma>0$,

$$\begin{aligned}\langle A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S),\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\rangle&\leq\|A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\|_2\|\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\|_2\\&\leq\frac{\gamma}{2}\|A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\|_2^2+\frac{1}{2\gamma}\|\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\|_2^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[f(A_{\mathbf{w}}(S^{(i)}),\mathbf{v}';z_i)-f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)]\leq\frac{L+\gamma}{2}\mathbb{E}\big[\|A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\|_2^2\big]\\+\frac{1}{2\gamma}\mathbb{E}\big[\|\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\|_2^2\big].\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}\big[f(\mathbf{w}',A_{\mathbf{v}}(S^{(i)});z_i)-f(\mathbf{w}',A_{\mathbf{v}}(S);z_i)\big]\leq\frac{L+\gamma}{2}\mathbb{E}\big[\|A_{\mathbf{v}}(S^{(i)})-A_{\mathbf{v}}(S)\|_2^2\big]\\+\frac{1}{2\gamma}\mathbb{E}\big[\|\nabla_{\mathbf{v}}f(\mathbf{w}',A_{\mathbf{v}}(S);z_i)\|_2^2\big].\end{aligned}$$

Combining the above two inequalities with Eq. (B.3), we have

$$\begin{aligned}&\triangle^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))-\triangle_S^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))\leq\frac{L+\gamma}{2n}\sum_{i=1}^n\mathbb{E}\left[\left\|\begin{pmatrix}A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\\ A_{\mathbf{v}}(S^{(i)})-A_{\mathbf{v}}(S)\end{pmatrix}\right\|_2^2\right]\\&+\frac{1}{2\gamma}\sup_{\mathbf{v}'\in\mathcal{V}}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n\|\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\|_2^2\Big]+\frac{1}{2\gamma}\sup_{\mathbf{w}'\in\mathcal{W}}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n\|\nabla_{\mathbf{v}}f(\mathbf{w}',A_{\mathbf{v}}(S);z_i)\|_2^2\Big].\end{aligned}$$

Using the fact that (Assumption 2)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\|\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';z_i)\|_2^2=\mathbb{E}_r[\|\nabla_{\mathbf{w}}f(A_{\mathbf{w}}(S),\mathbf{v}';S_r)\|_2^2]\leq G_{S,A}^2,\\\frac{1}{n}\sum_{i=1}^n\|\nabla_{\mathbf{v}}f(\mathbf{w}',A_{\mathbf{v}}(S);z_i)\|_2^2=\mathbb{E}_r[\|\nabla_{\mathbf{v}}f(\mathbf{w}',A_{\mathbf{v}}(S);S_r)\|_2^2]\leq G_{S,A}^2,\end{aligned}$$

we further get

$$\triangle^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))-\triangle_S^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))\leq\frac{L+\gamma}{2n}\sum_{i=1}^n\mathbb{E}\left[\left\|\begin{pmatrix}A_{\mathbf{w}}(S^{(i)})-A_{\mathbf{w}}(S)\\ A_{\mathbf{v}}(S^{(i)})-A_{\mathbf{v}}(S)\end{pmatrix}\right\|_2^2\right]+\frac{\mathbb{E}[G_{S,A}^2]}{\gamma}.$$

Since A is ϵ-argument-stable in expectation, we know

$$\triangle^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))-\triangle_S^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))\leq\frac{L+\gamma}{2}\epsilon^2+\frac{\mathbb{E}[G_{S,A}^2]}{\gamma}.$$

If we choose $\gamma=\sqrt{2\mathbb{E}[G_{S,A}^2]}/\epsilon$, then

$$\triangle^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))-\triangle_S^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))\leq L\epsilon^2/2+\sqrt{2\mathbb{E}[G_{S,A}^2]}\epsilon.$$

This completes the proof. □

B.2 Proofs on General Stability Bounds

In this section, we present the proof for our general stability bounds of Biased-SGDA under the convex-concave case.

To prove Theorem 2, we require the following standard lemma [45], which shows the monotonicity of the gradient map for convex-concave functions.

Lemma 11 ([45]). *Let f be a convex-concave function. Then*

$$\left\langle \begin{pmatrix} \mathbf{w}-\mathbf{w}' \\ \mathbf{v}-\mathbf{v}' \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w},\mathbf{v}) - \nabla_{\mathbf{w}} f(\mathbf{w}',\mathbf{v}') \\ \nabla_{\mathbf{v}} f(\mathbf{w}',\mathbf{v}') - \nabla_{\mathbf{v}} f(\mathbf{w},\mathbf{v}) \end{pmatrix} \right\rangle \geq 0. \tag{B.4}$$

Proof of Theorem 2. Let $S = \{z_1,\ldots,z_n\}$ and $S^{(i)} = \{z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n\}$. Let $\{\mathbf{w}_t,\mathbf{v}_t\}$ and $\{\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}\}$ be the sequence produced by (4.3) w.r.t. S and $S^{(i)}$, respectively. Note that the projection step is nonexpansive, according to Eq. (4.3), Eq. (4.4), we know

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)} \\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}^{(i)} \end{pmatrix} \right\|_2^2 \leq & \left\| \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2^2 + \eta^2 \left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \end{pmatrix} \right\|_2^2 \\ & - 2\eta \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \\ g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \end{pmatrix} \right\rangle. \end{aligned} \tag{B.5}$$

Recall the definition of the bias as in Eq. (4.5). Taking a conditional expectation implies

$$\begin{aligned} & \mathbb{E}_t\left[\left\| \begin{pmatrix} \mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)} \\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}^{(i)} \end{pmatrix} \right\|_2^2 \right] \\ & \leq \left\| \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2^2 + \eta^2 \mathbb{E}_t\left[\left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \end{pmatrix} \right\|_2^2 \right] \\ & - 2\eta \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}}\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t) + b_{t,\mathbf{w}} - \nabla_{\mathbf{w}}\widetilde{F}_{1,S^{(i)}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}) - b_{t,\mathbf{w}}^{(i)} \\ \nabla_{\mathbf{v}}\widetilde{F}_{2,S^{(i)}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}) + b_{t,\mathbf{v}}^{(i)} - \nabla_{\mathbf{v}}\widetilde{F}_{2,S}(\mathbf{w}_t,\mathbf{v}_t) - b_{t,\mathbf{v}} \end{pmatrix} \right\rangle. \end{aligned} \tag{B.6}$$

We know

$$\begin{aligned} & - 2\eta \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}}\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t) - \nabla_{\mathbf{w}}\widetilde{F}_{1,S^{(i)}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}) \\ \nabla_{\mathbf{v}}\widetilde{F}_{2,S^{(i)}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}) - \nabla_{\mathbf{v}}\widetilde{F}_{2,S}(\mathbf{w}_t,\mathbf{v}_t) \end{pmatrix} \right\rangle \\ & = -\frac{2\eta}{n} \sum_{j:j\neq i} \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_j) - \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j) \\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j) - \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_j) \end{pmatrix} \right\rangle \\ & - \frac{2\eta}{n} \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_i) - \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i') \\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i') - \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_i) \end{pmatrix} \right\rangle. \end{aligned}$$

Noting that $\tilde{f}_1, \tilde{f}_2$ are convex w.r.t. $\mathbf{w}$ and concave w.r.t $\mathbf{v}$, it then follows from Lemma 11 and Schwartz's inequality that

$$\begin{aligned} & - 2\eta \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}}\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t) - \nabla_{\mathbf{w}}\widetilde{F}_{1,S^{(i)}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}) \\ \nabla_{\mathbf{v}}\widetilde{F}_{2,S^{(i)}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)}) - \nabla_{\mathbf{v}}\widetilde{F}_{2,S}(\mathbf{w}_t,\mathbf{v}_t) \end{pmatrix} \right\rangle \\ & \leq \frac{2\eta}{n} \left\| \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2 \left\| \begin{pmatrix} \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_i) - \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i') \\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i') - \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_i) \end{pmatrix} \right\|_2. \end{aligned} \tag{B.7}$$

Using Schwarz's inequality, we also know

$$- 2\eta \left\langle \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix}, \begin{pmatrix} b_{t,\mathbf{w}}-b_{t,\mathbf{w}}^{(i)} \\ b_{t,\mathbf{v}}^{(i)}-b_{t,\mathbf{v}} \end{pmatrix} \right\rangle \leq 2\eta \left\| \begin{pmatrix} \mathbf{w}_t-\mathbf{w}_t^{(i)} \\ \mathbf{v}_t-\mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2 \left\| \begin{pmatrix} b_{t,\mathbf{w}}-b_{t,\mathbf{w}}^{(i)} \\ b_{t,\mathbf{v}}^{(i)}-b_{t,\mathbf{v}} \end{pmatrix} \right\|_2. \tag{B.8}$$

Plugging Eq. (B.7) and Eq. (B.8) into Eq. (B.6), we know

$$
\begin{aligned}
&\mathbb{E}_t\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}^{(i)}\end{pmatrix}\right\|_2^2\right]\\
&\le\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2+\eta^2\mathbb{E}_t\left[\left\|\begin{pmatrix}g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})-g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right\|_2^2\right]\\
&+\frac{2\eta}{n}\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2\left\|\begin{pmatrix}\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_i)-\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')\\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')-\nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_i)\end{pmatrix}\right\|_2\\
&+2\eta\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2\left\|\begin{pmatrix}b_{t,\mathbf{w}}-b_{t,\mathbf{w}}^{(i)}\\ b_{t,\mathbf{v}}^{(i)}-b_{t,\mathbf{v}}\end{pmatrix}\right\|_2.
\end{aligned}
$$

Taking an average of $i\in[n]$, and using the fact that $\sum_{i=1}^n a_ib_i\le(\sum_{i=1}^n a_i^2)^{\frac{1}{2}}(\sum_{i=1}^n b_i^2)^{\frac{1}{2}}$, we get

$$
\begin{aligned}
&\mathbb{E}_t\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}^{(i)}\end{pmatrix}\right\|_2^2\right]\\
&\le\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2+\eta^2\mathbb{E}_t\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})-g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right\|_2^2\right]\\
&+\frac{2\eta}{n}\left(\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2\right)^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_i)-\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')\\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')-\nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_i)\end{pmatrix}\right\|_2^2\right)^{\frac{1}{2}}\\
&+2\eta\left(\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2\right)^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}b_{t,\mathbf{w}}-b_{t,\mathbf{w}}^{(i)}\\ b_{t,\mathbf{v}}^{(i)}-b_{t,\mathbf{v}}\end{pmatrix}\right\|_2^2\right)^{\frac{1}{2}}.
\end{aligned}
$$

Introduce

$$
\bar{\Delta}_t=\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2\right]\right)^{\frac{1}{2}},\quad\forall t\in[T].
$$

Taking the expectation to the above inequality and using the inequality $\mathbb{E}[XY]\le(\mathbb{E}[X^2])^{\frac{1}{2}}(\mathbb{E}[Y^2])^{\frac{1}{2}}$, we have

$$
\begin{aligned}
\bar{\Delta}_{t+1}^2\le&\,\bar{\Delta}_t^2+\eta^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})-g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right\|_2^2\right]\\
&+\frac{2\eta}{n}\bar{\Delta}_t\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_i)-\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')\\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')-\nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_i)\end{pmatrix}\right\|_2^2\right]\right)^{\frac{1}{2}}\\
&+2\eta\bar{\Delta}_t\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}b_{t,\mathbf{w}}-b_{t,\mathbf{w}}^{(i)}\\ b_{t,\mathbf{v}}^{(i)}-b_{t,\mathbf{v}}\end{pmatrix}\right\|_2^2\right]\right)^{\frac{1}{2}}.
\end{aligned}
$$

Using conditions as in Eq. (4.6), Eq. (4.7) and Eq. (4.8), this further yields

$$
\bar{\Delta}_{t+1}^2\le(1+A\eta^2+2\bar{A}^{\frac{1}{2}}\eta)\bar{\Delta}_t^2+2\eta\big(\sqrt{\bar{B}}+\frac{\widetilde{G}}{n}\big)\bar{\Delta}_t+\eta^2B.
$$

Applying this inequality recursively and using the fact that $\bar{\Delta}_1=0$ imply that

$$
\begin{aligned}
\bar{\Delta}_{t+1}^2&\le\eta^2B\sum_{k=1}^t(1+A\eta^2+2\bar{A}^{\frac{1}{2}}\eta)^{t-k}+2\eta\sum_{k=1}^t\big(\sqrt{\bar{B}}+\frac{\widetilde{G}}{n}\big)\bar{\Delta}_k(1+A\eta^2+2\bar{A}^{\frac{1}{2}}\eta)^{t-k}\\
&=\eta^2B\frac{\big(1+A\eta^2+2\bar{A}^{\frac{1}{2}}\eta\big)^t-1}{A\eta^2+2\bar{A}^{\frac{1}{2}}\eta}+2\eta\sum_{k=1}^t\big(\sqrt{\bar{B}}+\frac{\widetilde{G}}{n}\big)\bar{\Delta}_k(1+A\eta^2+2\bar{A}^{\frac{1}{2}}\eta)^{t-k}.
\end{aligned}
$$

Since the above inequality applies for any $t \in [T]$, if we introduce $\bar{\Delta} = \max_{t \le T} \bar{\Delta}_t$, we further get

$$\bar{\Delta}^2 \le \eta^2 B \frac{\left(1 + A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta\right)^T - 1}{A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta} + 2\eta\big(\sqrt{\bar{B}} + \frac{\widetilde{G}}{n}\big)\frac{\left(1 + A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta\right)^T - 1}{A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta}\bar{\Delta}.$$

Note the right-hand side of the above inequality is an increasing function of $A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta$. By taking η such that $A\eta^2 + 2\bar{A}^{\frac{1}{2}}\eta \le \frac{1}{T}$ in the above inequality and using $(1 + 1/T)^T \le e$, we get

$$\bar{\Delta}^2 \le 2\eta^2 BT + 4\eta\big(\sqrt{\bar{B}} + \frac{\widetilde{G}}{n}\big)T\bar{\Delta}.$$

Solving the above quadratic inequality of $\bar{\Delta}$ implies that (if $x^2 \le ax + b$ for some $a, b \ge 0$, then $x^2 \le a^2 + 2b$)

$$\bar{\Delta}^2 \le 4B\eta^2 T + 16\big(\sqrt{\bar{B}} + \frac{\widetilde{G}}{n}\big)^2\eta^2 T^2.$$

The stated bound follows directly. □

B.3 Proofs on Applications

In this section, we present stability bounds for the specific Biased-SGDA algorithms for the convex-concave case. We first present a lemma that is often applied in our proof. This lemma shows the condition Eq. (4.8) holds with $\widetilde{G} = 2\sqrt{2\mathbb{E}[G_S^2]}$ when we choose $\tilde{f}_1 = \tilde{f}_2 = f$.

Lemma 12. *Under Assumption 3, we know*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\end{pmatrix}\right\|_2^2\right] \le 8\mathbb{E}[G_S^2].$$

Proof. By the inequality that $(a+b)^2 \le 2a^2 + 2b^2$, we know

$$\begin{aligned}
&\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\end{pmatrix}\right\|_2^2 \\
&= \left\|\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\right\|_2^2 + \left\|\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\right\|_2^2 \\
&\le 2\Big(\left\|\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\right\|_2^2 + \left\|\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\right\|_2^2 + \left\|\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\right\|_2^2 + \left\|\nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\right\|_2^2\Big).
\end{aligned}$$

Taking the expectation of the above inequality, then taking an average of $i \in [n]$, we get (r follows the uniform distribution over $[n]$)

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\end{pmatrix}\right\|_2^2\right] \\
&\le 4\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\right\|_2^2\right] + 4\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\right\|_2^2\right] \\
&= 4\mathbb{E}\big[\|\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_r)\|_2^2\big] + 4\mathbb{E}\big[\|\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_r)\|_2^2\big] \le 8\mathbb{E}[G_S^2],
\end{aligned}$$

where we have used the identity

$$\mathbb{E}[\|\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\|_2^2] = \mathbb{E}[\|\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\|_2^2], \mathbb{E}[\|\nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\|_2^2] = \mathbb{E}[\|\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_i)\|_2^2]$$

in the first step, and Assumption 3 in the last step. □

B.3.1 SGDA

In this subsection, we give the proofs for SGDA. We first prove Lemma 13 to show that Eq. (4.6), Eq. (4.7) hold for SGDA.

Lemma 13. *Assume for all z, the function $(\mathbf{w},\mathbf{v})\mapsto f(\mathbf{w},\mathbf{v};z)$ is convex-concave. Let Assumption 1, 3 hold. If we choose the gradient estimator in Eq.* (5.1) *and $\tilde{f}_1=\tilde{f}_2=f$, then Eq.* (4.6)*, Eq.* (4.7) *hold with $A=L^2, B=\frac{8\mathbb{E}[G_S^2]}{n}, \bar{A}=0, \bar{B}=0$.*

Proof of Lemma 13. Since $b_t=b_t^{(i)}=0$, it is clear that Eq. (4.7) holds with $\bar{A}=0,\bar{B}=0$. Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})-g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right\|_2^2\right]\\
&=\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j:j\neq i}\left\|\begin{pmatrix} \nabla_{\mathbf{w}}f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j)-\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_j)\\ \nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_j)-\nabla_{\mathbf{v}}f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j)\end{pmatrix}\right\|_2^2\right]\\
&+\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\left\|\begin{pmatrix} \nabla_{\mathbf{w}}f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')-\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_i)\\ \nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_i)-\nabla_{\mathbf{v}}f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')\end{pmatrix}\right\|_2^2\right]\\
&\leq L^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}_t^{(i)}\\ \mathbf{v}_t-\mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2\right]+\frac{8\mathbb{E}[G_S^2]}{n},
\end{aligned}\tag{B.9}$$

where we have used the fact that i_t follows the uniform distribution over $[n]$ for the first step, Assumption 1 and Lemma 12 for the last step. This shows Eq. (4.6) holds with $A=L^2, B=\frac{8\mathbb{E}[G_S^2]}{n}$. □

We then plug these estimates into Theorem 2 to immediately get Theorem 3. We omit the details for brevity.

B.3.2 Zeroth-order SGDA

In this subsection, we give the proofs for Zeroth-order SGDA. Lemma 15 is a simple variant of Lemma 1 in [41]. We give the proof for completeness.

Lemma 14. *Let $\mathbf{u}\in\mathbb{R}^d$ from $\mathrm{Unif}(\mathbb{S}^{d-1})$, then $\mathbb{E}[\mathbf{u}\mathbf{u}^\top]=\frac{1}{d}\mathbf{I}_d$, where $\mathbf{I}_d$ denotes the d-dimensional identity matrix.*

Proof. By definition, the uniform distribution on the unit sphere is rotationally invariant. That is, for any orthogonal matrix $\mathbf{R}\in\mathbb{R}^d$, the random vectors $\mathbf{R}\mathbf{u}$ and $\mathbf{u}$ have the same distribution.
Define $\mathbf{A}=\mathbb{E}[\mathbf{u}\mathbf{u}^\top]$. We first show that $\mathbf{A}=c\mathbf{I}_d$ for some constant $c\in\mathbb{R}$.
For any orthogonal matrix $\mathbf{R}$, using linearity of expectation and rotational invariance,

$$\mathbf{R}\mathbf{A}\mathbf{R}^\top=\mathbf{R}\,\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\,\mathbf{R}^\top=\mathbb{E}[\mathbf{R}\mathbf{u}(\mathbf{R}\mathbf{u})^\top]=\mathbb{E}[\mathbf{u}\mathbf{u}^\top]=\mathbf{A}.$$

This implies that $\mathbf{A}$ commutes with all orthogonal matrices. Since any matrix that commutes with all orthogonal matrices must be a scalar multiple of the identity, we conclude that $\mathbf{A}=c\mathbf{I}_d$ for some constant c.
Now we determine c. Since

$$\mathrm{tr}(\mathbf{A})=\mathrm{tr}\left(\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\right)=\mathbb{E}\big[\,\mathrm{tr}(\mathbf{u}\mathbf{u}^\top)\big]=\mathbb{E}[\mathbf{u}^\top\mathbf{u}]=\mathbb{E}[\|\mathbf{u}\|_2^2]=1,\quad\textbf{and}\quad\mathrm{tr}(\mathbf{A})=\mathrm{tr}(c\mathbf{I}_d)=cd,$$

we know $c=\frac{1}{d}$. The result follows directly. □

Lemma 15. *Let $\mathbf{u}_l\in\mathbb{R}^d, l\in\{1,2\ldots,K\}$ be drawn independently from $\mathrm{Unif}(\mathbb{S}^{d-1})$. For every random vector $\mathbf{v}\in\mathbb{R}^d$ independent of all $\mathbf{u}_l, l\in\{1,2\ldots,K\}$, it is true that*

$$\mathbb{E}_{\mathbf{u}}\Big[\Big\|\frac{1}{K}\sum_{l=1}^K\langle\mathbf{v},\mathbf{u}_l\rangle\mathbf{u}_l\Big\|_2^2\Big]=\frac{1}{K^2}\Big(\frac{K}{d}+\frac{K(K-1)}{d^2}\Big)\|\mathbf{v}\|_2^2=\frac{K+d-1}{Kd^2}\|\mathbf{v}\|_2^2,$$

where $\mathbb{E}_{\mathbf{u}}[\cdot]$ denotes the conditional expectation w.r.t. $\mathbf{u}_1,\ldots,\mathbf{u}_K$.

Proof. For any fixed $\mathbf{v} \in \mathbb{R}^d$, we have

$$\Big\| \frac{1}{K} \sum_{l=1}^{K} \langle \mathbf{v}, \mathbf{u}_l \rangle \mathbf{u}_l \Big\|_2^2 = \frac{1}{K^2} \Big[\sum_{l=1}^{K} \mathbb{E}[(\langle \mathbf{v}, \mathbf{u}_l \rangle)^2] + \sum_{l \neq m} \mathbb{E}[\langle \mathbf{v}, \mathbf{u}_l \rangle \langle \mathbf{u}_l, \mathbf{u}_m \rangle \langle \mathbf{v}, \mathbf{u}_m \rangle] \Big]. \tag{B.10}$$

For the first term in Eq. (B.10), we have

$$\mathbb{E}[(\langle \mathbf{v}, \mathbf{u}_l \rangle)^2] = \mathbf{v}^\top \mathbb{E}[\mathbf{u}_l \mathbf{u}_l^\top] \mathbf{v} = \frac{1}{d} \|\mathbf{v}\|_2^2,$$

where we use the fact that $\mathbb{E}[\mathbf{u}_l \mathbf{u}_l^\top] = \frac{1}{d} \mathbf{I}_d$ from Lemma 14.
For the second term in Eq. (B.10), due to the independency of $\mathbf{u}_l$ and $\mathbf{u}_m$, we know

$$\mathbb{E}[\langle \mathbf{v}, \mathbf{u}_l \rangle \langle \mathbf{u}_l, \mathbf{u}_m \rangle \langle \mathbf{v}, \mathbf{u}_m \rangle] = \mathbb{E}[(\langle \mathbf{v}, \mathbf{u}_l \rangle \mathbf{u}_l)^\top] \cdot \mathbb{E}[\langle \mathbf{v}, \mathbf{u}_m \rangle \mathbf{u}_m] = (\frac{1}{d} \mathbf{v})^\top \cdot (\frac{1}{d} \mathbf{v}) = \frac{1}{d^2} \|\mathbf{v}\|_2^2,$$

where we use the fact that $\mathbb{E}[(\langle \mathbf{v}, \mathbf{u} \rangle \mathbf{u})] = \frac{1}{d} \mathbf{v}$ for any $\mathbf{u}$ from $\mathrm{Unif}(\mathbb{S}^{d-1})$. We plug the above two identities back into Eq. (B.10), and derive

$$\mathbb{E}_{\mathbf{u}} \Big[\Big\| \frac{1}{K} \sum_{l=1}^{K} \langle \mathbf{v}, \mathbf{u}_l \rangle \mathbf{u}_l \Big\|_2^2 \Big] = \frac{1}{K^2} \Big(\frac{K}{d} + \frac{K(K-1)}{d^2} \Big) \|\mathbf{v}\|_2^2 = \frac{K + d - 1}{K d^2} \|\mathbf{v}\|_2^2.$$

The proof is completed. □

The following lemma is a direct application of Lemma 15, which is useful for our main results.

Lemma 16. *Let $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^d$ be drawn independently from $\mathrm{Unif}(\mathbb{S}^{d-1})$. For every random vector $\mathbf{v}_1 \in \mathbb{R}^d$ independent of all $\mathbf{u}_1$ and $\mathbf{v}_2 \in \mathbb{R}^d$ independent of all $\mathbf{u}_2$, it is true that*

$$\mathbb{E}_{\mathbf{u}} \Big[\Big\| \begin{pmatrix} \mathbf{v}_1^\top \mathbf{u}_1 \mathbf{u}_1 \\ \mathbf{v}_2^\top \mathbf{u}_2 \mathbf{u}_2 \end{pmatrix} \Big\|_2^2 \Big] = \frac{1}{d} \Big\| \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{pmatrix} \Big\|_2^2,$$

where $\mathbb{E}_{\mathbf{u}}[\cdot]$ denotes the conditional expectation w.r.t. $\mathbf{u}_1, \mathbf{u}_2$.

Proof. By Lemma 15 with $K = 1$, we derive

$$\mathbb{E}_{\mathbf{u}} \Big[\Big\| \begin{pmatrix} \mathbf{v}_1^\top \mathbf{u}_1 \mathbf{u}_1 \\ \mathbf{v}_2^\top \mathbf{u}_2 \mathbf{u}_2 \end{pmatrix} \Big\|_2^2 \Big] = \mathbb{E}_{\mathbf{u}_1} \big[\|\mathbf{v}_1^\top \mathbf{u}_1 \mathbf{u}_1\|_2^2 \big] + \mathbb{E}_{\mathbf{u}_2} \big[\|\mathbf{v}_2^\top \mathbf{u}_2 \mathbf{u}_2\|_2^2 \big] = \frac{1}{d} \|\mathbf{v}_1\|_2^2 + \frac{1}{d} \|\mathbf{v}_2\|_2^2.$$

The stated bound follows directly. □

We now prove an important lemma which shows that $\tilde{f}_1, \tilde{f}_2$ preserve the smoothness, bounded second moment condition, convexity, and concavity of f, and that g is an unbiased estimator.

Proof of Lemma 4. **1.** Since for any z, $\mathbf{w} \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex, from Lemma 4.1 in Gao et al. [20], we know $\mathbf{w} \mapsto \tilde{f}_1(\mathbf{w}, \mathbf{v}; z)$ is convex. Since for any z, $\mathbf{v} \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is concave, we know $\mathbf{v} \mapsto \tilde{f}_1(\mathbf{w}, \mathbf{v}; z)$ is concave. Then, $(\mathbf{w}, \mathbf{v}) \mapsto \tilde{f}_1(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Similarly, $(\mathbf{w}, \mathbf{v}) \mapsto \tilde{f}_2(\mathbf{w}, \mathbf{v}; z)$ is convex-concave.
Recall the surrogate functions

$$\tilde{f}_1(\mathbf{w}, \mathbf{v}; z) = \mathbb{E}_{\mathbf{u} \sim \mathrm{Unif}(\mathbb{B}^d)} \big[f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) \big] = \frac{1}{\kappa} \int_{\mathbb{B}^d} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) d\mathbf{u},$$

$$\tilde{f}_2(\mathbf{w}, \mathbf{v}; z) = \mathbb{E}_{\mathbf{u} \sim \mathrm{Unif}(\mathbb{B}^d)} \big[f(\mathbf{w}, \mathbf{v} + \mu_2 \mathbf{u}; z) \big] = \frac{1}{\kappa} \int_{\mathbb{B}^d} f(\mathbf{w}, \mathbf{v} + \mu_2 \mathbf{u}; z) d\mathbf{u},$$

where κ is the volume of the ball $\mathbb{B}^d$, we know

$$\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; z) = \frac{1}{\kappa} \int_{\mathbb{B}^d} \nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) d\mathbf{u}, \quad \nabla_{\mathbf{w}} \tilde{f}_2(\mathbf{w}, \mathbf{v}; z) = \frac{1}{\kappa} \int_{\mathbb{B}^d} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v} + \mu_2 \mathbf{u}; z) d\mathbf{u},$$

$$\nabla_{\mathbf{v}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; z) = \frac{1}{\kappa} \int_{\mathbb{B}^d} \nabla_{\mathbf{v}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) d\mathbf{u}, \quad \nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}, \mathbf{v}; z) = \frac{1}{\kappa} \int_{\mathbb{B}^d} \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v} + \mu_2 \mathbf{u}; z) d\mathbf{u}.$$

Then for any $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ and $\mathbf{v}, \mathbf{v}' \in \mathcal{V}$,

$$\begin{aligned}
&\left\| \begin{pmatrix} \nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}', \mathbf{v}'; z) \\ \nabla_{\mathbf{v}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; z) - \nabla_{\mathbf{v}} \tilde{f}_1(\mathbf{w}', \mathbf{v}'; z) \end{pmatrix} \right\|_2 \\
&= \left\| \begin{pmatrix} \frac{1}{\kappa} \int_{\mathbb{B}^d} (\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - \nabla_{\mathbf{w}} f(\mathbf{w}' + \mu_1 \mathbf{u}, \mathbf{v}'; z)) d\mathbf{u} \\ \frac{1}{\kappa} \int_{\mathbb{B}^d} (\nabla_{\mathbf{v}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - \nabla_{\mathbf{v}} f(\mathbf{w}' + \mu_1 \mathbf{u}, \mathbf{v}'; z)) d\mathbf{u} \end{pmatrix} \right\|_2 \\
&= \left\| \frac{1}{\kappa} \int_{\mathbb{B}^d} \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - \nabla_{\mathbf{w}} f(\mathbf{w}' + \mu_1 \mathbf{u}, \mathbf{v}'; z) \\ \nabla_{\mathbf{v}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - \nabla_{\mathbf{v}} f(\mathbf{w}' + \mu_1 \mathbf{u}, \mathbf{v}'; z) \end{pmatrix} d\mathbf{u} \right\|_2 \\
&\leq \frac{1}{\kappa} \int_{\mathbb{B}^d} \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - \nabla_{\mathbf{w}} f(\mathbf{w}' + \mu_1 \mathbf{u}, \mathbf{v}'; z) \\ \nabla_{\mathbf{v}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - \nabla_{\mathbf{v}} f(\mathbf{w}' + \mu_1 \mathbf{u}, \mathbf{v}'; z) \end{pmatrix} \right\|_2 d\mathbf{u} \\
&\leq L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2,
\end{aligned}$$

where we use Assumption 1 for the last step. In this way, we show that $(\mathbf{w}, \mathbf{v}) \mapsto \tilde{f}_1(\mathbf{w}, \mathbf{v}; z)$ is L-smooth, and similarly, $(\mathbf{w}, \mathbf{v}) \mapsto \tilde{f}_2(\mathbf{w}, \mathbf{v}; z)$ is L-smooth.
We know

$$\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; S_r) = \nabla_{\mathbf{w}} \mathbb{E}_{\mathbf{u}}[f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; S_r)] = \mathbb{E}_{\mathbf{u}}[\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; S_r)],$$

where $\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Then, applying Jensen's inequality, we have

$$\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; S_r)\|_2^2 = \|\mathbb{E}_{\mathbf{u}}[\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; S_r)]\|_2^2 \leq \mathbb{E}_{\mathbf{u}}[\|\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; S_r)\|_2^2].$$

Taking expectation w.r.t r to both sides above,

$$\begin{aligned}
\mathbb{E}_r[\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}, \mathbf{v}; S_r)\|_2^2] &\leq \mathbb{E}_{\mathbf{u}, r}[\|\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; S_r)\|_2^2] \\
&= \mathbb{E}_{\mathbf{u}}\left[\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; S_r)\|_2^2]\right] \leq G_S^2,
\end{aligned}$$

where we use Assumption 3 for the last step. The results hold similarly for $\mathbb{E}_r[\|\nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}, \mathbf{v}; S_r)\|_2^2]$.
2. Follows directly from Lemma 4.4 in Shalev-Shwartz [46]. □

We now present the proof of Lemma 18 and Lemma 21, which show that Eq. (4.6) and Eq. (4.8) hold for Zeroth-order SGDA. To prove Lemma 18, we require the following lemma to show that the Hessian matrices can be bounded.

Lemma 17. *Let Assumption 1 hold. For any matrix A, the spectral norm is*

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2.$$

Then for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$,

$$\|\nabla_{\mathbf{w}}^2 f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq L, \quad \|\nabla_{\mathbf{v}}^2 f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq L.$$

Proof. From Assumption 1, we know $\|\nabla^2 f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq L$, where

$$\nabla^2 f(\mathbf{w}, \mathbf{v}; z) = \begin{pmatrix} \nabla_{\mathbf{w}}^2 f(\mathbf{w}, \mathbf{v}; z) & \nabla_{\mathbf{w}} \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z) \\ \nabla_{\mathbf{v}} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z) & \nabla_{\mathbf{v}}^2 f(\mathbf{w}, \mathbf{v}; z) \end{pmatrix}. \tag{B.11}$$

Then, for any $\mathbf{h}_{\mathbf{w}} \in \mathbb{R}^d$ such that $\|\mathbf{h}_{\mathbf{w}}\|_2 = 1$, we construct $\mathbf{h} = (\mathbf{h}_{\mathbf{w}}^\top, 0, \ldots, 0)^\top \in \mathbb{R}^{2d}$. According to the definition of the matrix norm and Eq. (B.11), we know

$$\begin{aligned}
L \geq \|\nabla^2 f(\mathbf{w}, \mathbf{v}; z) \cdot \mathbf{h}\|_2 &= \sqrt{\|\nabla_{\mathbf{w}}^2 f(\mathbf{w}, \mathbf{v}; z) \cdot \mathbf{h}_{\mathbf{w}}\|_2^2 + \|\nabla_{\mathbf{v}, \mathbf{w}}^2 f(\mathbf{w}, \mathbf{v}; z) \cdot \mathbf{h}_{\mathbf{w}}\|_2^2} \\
&\geq \|\nabla_{\mathbf{w}}^2 f(\mathbf{w}, \mathbf{v}; z) \cdot \mathbf{h}_{\mathbf{w}}\|_2
\end{aligned}$$

Therefore, $\|\nabla_{\mathbf{w}}^2 f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq L$. Similarly, $\|\nabla_{\mathbf{v}}^2 f(\mathbf{w}, \mathbf{v}; z)\|_2 \leq L$. □

Lemma 18. *Let Assumption 3 hold, $\tilde{f}_1, \tilde{f}_2$ be given in Eq.* (5.3)*. Then, Eq.* (4.8) *holds with $\widetilde{G}^2 = 8\mathbb{E}[G_S^2]$.*

Proof of Lemma 18. By the inequality that $(a+b)^2 \le 2a^2 + 2b^2$, we know

$$\begin{aligned}
&\left\| \begin{pmatrix} \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \\ \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') \end{pmatrix} \right\|_2^2 \\
&= \left\| \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \right\|_2^2 + \left\| \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') \right\|_2^2 \\
&\le 2\Big(\left\| \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') \right\|_2^2 + \left\| \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \right\|_2^2 + \left\| \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \right\|_2^2 + \left\| \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') \right\|_2^2 \Big).
\end{aligned}$$

Taking the expectation of the above inequality, then taking an average of $i \in [n]$, we get (r follows the uniform distribution over $[n]$)

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left\| \begin{pmatrix} \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') - \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \\ \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) - \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i') \end{pmatrix} \right\|_2^2 \right] \\
&\le 4\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left\| \nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \right\|_2^2 \right] + 4\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left\| \nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i) \right\|_2^2 \right] \\
&= 4\mathbb{E}\big[\|\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_r)\|_2^2 \big] + 4\mathbb{E}\big[\|\nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_r)\|_2^2 \big] \le 8\mathbb{E}[G_S^2],
\end{aligned}$$

where in the first step we have used the fact $\mathbb{E}[\|\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\|_2^2] = \mathbb{E}[\|\nabla_{\mathbf{w}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i)\|_2^2], \mathbb{E}[\|\nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_i')\|_2^2] = \mathbb{E}[\|\nabla_{\mathbf{v}} \tilde{f}(\mathbf{w}_t, \mathbf{v}_t; z_i)\|_2^2]$, and in the last step we have used Lemma 4. □

To prove Lemma 21, we need the following two lemmas.

Lemma 19 (Spherical Poincaré inequality; see, e.g., [15, 28])**.** *Let $d \ge 2$ and $\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. For any locally Lipschitz function $h : \mathbb{S}^{d-1} \to \mathbb{R}$, we have*

$$\mathbb{E}_{\mathbf{u}}[(h(\mathbf{u}) - \mathbb{E}_{\mathbf{u}}[h(\mathbf{u})])^2] \le \frac{1}{d-1} \mathbb{E}_{\mathbf{u}}[\|\nabla_{\mathbb{S}} h(\mathbf{u})\|_2^2],$$

where $\nabla_{\mathbb{S}} h(\mathbf{u})$ denotes the tangential gradient of h at $\mathbf{u}$. More precisely, if $\widetilde{h}$ is a differentiable extension of h to a neighborhood of $\mathbb{S}^{d-1}$, denoting $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ as the identity matrix, then $\nabla_{\mathbb{S}} h(\mathbf{u}) = (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top) \nabla_{\mathbf{u}} \widetilde{h}(\mathbf{u})$ (see, e.g., Section 4.6.1 in [2]).

Lemma 20. *Let Assumption 1 hold. For any $\mathbf{w}, \mathbf{v}, \mathbf{w}', \mathbf{v}'$, the function*

$$h(\mathbf{u}) := [f(\mathbf{w}' + \mu\mathbf{u}, \mathbf{v}'; z) - f(\mathbf{w}' - \mu\mathbf{u}, \mathbf{v}'; z)] - [f(\mathbf{w} + \mu\mathbf{u}, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}, \mathbf{v}; z)]$$

with $\mathbf{u} \in \mathbb{S}^{d-1}$ is $2\mu L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2$-Lipschitz continuous with respect to $\mathbf{u}$.

Proof. Take arbitrary $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{S}^{d-1}$. Decompose the difference $h(\mathbf{u}_1) - h(\mathbf{u}_2)$ as follows:

$$\begin{aligned}
&h(\mathbf{u}_1) - h(\mathbf{u}_2) \\
&= [f(\mathbf{w}' + \mu\mathbf{u}_1, \mathbf{v}'; z) - f(\mathbf{w}' - \mu\mathbf{u}_1, \mathbf{v}'; z) - f(\mathbf{w} + \mu\mathbf{u}_1, \mathbf{v}; z) + f(\mathbf{w} - \mu\mathbf{u}_1, \mathbf{v}; z)] \\
&\quad - [f(\mathbf{w}' + \mu\mathbf{u}_2, \mathbf{v}'; z) - f(\mathbf{w}' - \mu\mathbf{u}_2, \mathbf{v}'; z) - f(\mathbf{w} + \mu\mathbf{u}_2, \mathbf{v}; z) + f(\mathbf{w} - \mu\mathbf{u}_2, \mathbf{v}; z)] \\
&= [f(\mathbf{w}' + \mu\mathbf{u}_1, \mathbf{v}'; z) - f(\mathbf{w}' + \mu\mathbf{u}_2, \mathbf{v}'; z)] - [f(\mathbf{w}' - \mu\mathbf{u}_1, \mathbf{v}'; z) - f(\mathbf{w}' - \mu\mathbf{u}_2, \mathbf{v}'; z)] \\
&\quad - [f(\mathbf{w} + \mu\mathbf{u}_1, \mathbf{v}; z) - f(\mathbf{w} + \mu\mathbf{u}_2, \mathbf{v}; z)] + [f(\mathbf{w} - \mu\mathbf{u}_1, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}_2, \mathbf{v}; z)] \,. \quad \text{(B.12)}
\end{aligned}$$

Apply the Newton-Leibniz formula

$$f(\mathbf{a} + \mathbf{b}, \mathbf{c}; z) - f(\mathbf{a}, \mathbf{c}; z) = \int_0^1 \nabla_{\mathbf{w}} f(\mathbf{a} + t\mathbf{b}, \mathbf{c}; z) \cdot \mathbf{b} dt$$

to each of the four difference terms:

For the first term, let $\mathbf{a} = \mathbf{w}' + \mu\mathbf{u}_2$, $\mathbf{b} = \mu(\mathbf{u}_1 - \mathbf{u}_2)$, and $\mathbf{c} = \mathbf{v}'$:

$$f(\mathbf{w}' + \mu\mathbf{u}_1, \mathbf{v}'; z) - f(\mathbf{w}' + \mu\mathbf{u}_2, \mathbf{v}'; z) = \mu \int_0^1 \nabla_{\mathbf{w}} f\left(\mathbf{w}' + \mu(\mathbf{u}_2 + t(\mathbf{u}_1 - \mathbf{u}_2)), \mathbf{v}'; z\right) \cdot (\mathbf{u}_1 - \mathbf{u}_2) dt.$$

For the second term, let $\mathbf{a} = \mathbf{w}' - \mu\mathbf{u}_2$, $\mathbf{b} = -\mu(\mathbf{u}_1 - \mathbf{u}_2)$, and $\mathbf{c} = \mathbf{v}'$:

$$f(\mathbf{w}'-\mu\mathbf{u}_1, \mathbf{v}'; z)-f(\mathbf{w}'-\mu\mathbf{u}_2, \mathbf{v}'; z) = -\mu \int_0^1 \nabla_{\mathbf{w}} f\left(\mathbf{w}' - \mu(\mathbf{u}_2 + t(\mathbf{u}_1 - \mathbf{u}_2)), \mathbf{v}'; z\right) \cdot (\mathbf{u}_1 - \mathbf{u}_2) dt.$$

For the third term, let $\mathbf{a} = \mathbf{w} + \mu\mathbf{u}_2$, $\mathbf{b} = \mu(\mathbf{u}_1 - \mathbf{u}_2)$, and $\mathbf{c} = \mathbf{v}$:

$$f(\mathbf{w}+\mu\mathbf{u}_1, \mathbf{v}; z)-f(\mathbf{w}+\mu\mathbf{u}_2, \mathbf{v}; z) = \mu \int_0^1 \nabla_{\mathbf{w}} f\left(\mathbf{w} + \mu(\mathbf{u}_2 + t(\mathbf{u}_1 - \mathbf{u}_2)), \mathbf{v}; z\right) \cdot (\mathbf{u}_1 - \mathbf{u}_2) dt.$$

For the fourth term, let $\mathbf{a} = \mathbf{w} - \mu\mathbf{u}_2$, $\mathbf{b} = -\mu(\mathbf{u}_1 - \mathbf{u}_2)$, and $\mathbf{c} = \mathbf{v}$:

$$f(\mathbf{w}-\mu\mathbf{u}_1, \mathbf{v}; z)-f(\mathbf{w}-\mu\mathbf{u}_2, \mathbf{v}; z) = -\mu \int_0^1 \nabla_{\mathbf{w}} f\left(\mathbf{w} - \mu(\mathbf{u}_2 + t(\mathbf{u}_1 - \mathbf{u}_2)), \mathbf{v}; z\right) \cdot (\mathbf{u}_1 - \mathbf{u}_2) dt.$$

Let $\boldsymbol{\xi}_t = \mathbf{u}_2 + t(\mathbf{u}_1 - \mathbf{u}_2)$. Note that $\|\boldsymbol{\xi}_t\|_2 \leq 1$ for all $t \in [0, 1]$ since $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{S}^{d-1}$. Substitute the four integrals back into Eq. (B.12), we get

$$\begin{aligned} h(\mathbf{u}_1) - h(\mathbf{u}_2) = \mu \int_0^1 &[(\nabla_{\mathbf{w}} f(\mathbf{w}' + \mu\boldsymbol{\xi}_t, \mathbf{v}'; z) - \nabla_{\mathbf{w}} f(\mathbf{w} + \mu\boldsymbol{\xi}_t, \mathbf{v}; z)) \\ &+(\nabla_{\mathbf{w}} f(\mathbf{w}' - \mu\boldsymbol{\xi}_t, \mathbf{v}'; z) - \nabla_{\mathbf{w}} f(\mathbf{w} - \mu\boldsymbol{\xi}_t, \mathbf{v}; z))] \cdot (\mathbf{u}_1 - \mathbf{u}_2) dt. \end{aligned} \tag{B.13}$$

Apply the Cauchy-Schwarz inequality $|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$ followed by the triangle inequality $\|\mathbf{a} + \mathbf{b}\|_2 \leq \|\mathbf{a}\|_2 + \|\mathbf{b}\|_2$ to the integrand:

$$\begin{aligned} &\left|\left[(\nabla_{\mathbf{w}} f'_+ - \nabla_{\mathbf{w}} f_+) + (\nabla_{\mathbf{w}} f'_- - \nabla_{\mathbf{w}} f_-)\right] \cdot (\mathbf{u}_1 - \mathbf{u}_2)\right| \\ &\leq \left(\left\|\nabla_{\mathbf{w}} f'_+ - \nabla_{\mathbf{w}} f_+\right\|_2 + \left\|\nabla_{\mathbf{w}} f'_- - \nabla_{\mathbf{w}} f_-\right\|_2\right) \cdot \|\mathbf{u}_1 - \mathbf{u}_2\|_2, \end{aligned}$$

where we use the shorthand $f'_+ = f(\mathbf{w}' + \mu\boldsymbol{\xi}_t, \mathbf{v}'; z)$, $f'_- = f(\mathbf{w}' - \mu\boldsymbol{\xi}_t, \mathbf{v}'; z)$, $f_+ = f(\mathbf{w} + \mu\boldsymbol{\xi}_t, \mathbf{v}; z)$, and $f_- = f(\mathbf{w} - \mu\boldsymbol{\xi}_t, \mathbf{v}; z)$.

Now apply the joint L-smoothness in Assumption 1 to each gradient difference. For the first difference:

$$\begin{aligned} \left\|\nabla_{\mathbf{w}} f'_+ - \nabla_{\mathbf{w}} f_+\right\|_2 &= \|\nabla_{\mathbf{w}} f(\mathbf{w}' + \mu\boldsymbol{\xi}_t, \mathbf{v}'; z) - \nabla_{\mathbf{w}} f(\mathbf{w} + \mu\boldsymbol{\xi}_t, \mathbf{v}; z)\|_2 \\ &\leq L \left\| \begin{pmatrix} (\mathbf{w}' + \mu\boldsymbol{\xi}_t) - (\mathbf{w} + \mu\boldsymbol{\xi}_t) \\ \mathbf{v}' - \mathbf{v} \end{pmatrix} \right\|_2 = L \left\| \begin{pmatrix} \mathbf{w}' - \mathbf{w} \\ \mathbf{v}' - \mathbf{v} \end{pmatrix} \right\|_2. \end{aligned}$$

For the second difference, the $-\mu\boldsymbol{\xi}_t$ terms cancel identically:

$$\left\|\nabla_{\mathbf{w}} f'_- - \nabla_{\mathbf{w}} f_-\right\|_2 \leq L \left\| \begin{pmatrix} (\mathbf{w}' - \mu\boldsymbol{\xi}_t) - (\mathbf{w} - \mu\boldsymbol{\xi}_t) \\ \mathbf{v}' - \mathbf{v} \end{pmatrix} \right\|_2 = L \left\| \begin{pmatrix} \mathbf{w}' - \mathbf{w} \\ \mathbf{v}' - \mathbf{v} \end{pmatrix} \right\|_2.$$

Combining these estimates:

$$\left\|\nabla_{\mathbf{w}} f'_+ - \nabla_{\mathbf{w}} f_+\right\|_2 + \left\|\nabla_{\mathbf{w}} f'_- - \nabla_{\mathbf{w}} f_-\right\|_2 \leq 2L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2.$$

Take the absolute value of Eq. (B.13) and substitute the above back into the expression for $|h(\mathbf{u}_1) - h(\mathbf{u}_2)|$ and integrate over $t \in [0, 1]$:

$$|h(\mathbf{u}_1) - h(\mathbf{u}_2)| \leq \mu \int_0^1 2L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2 \cdot \|\mathbf{u}_1 - \mathbf{u}_2\|_2 dt = 2\mu L \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2 \cdot \|\mathbf{u}_1 - \mathbf{u}_2\|_2.$$

This completes the proof. □

Lemma 21. *Let $g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ be given in Eq.* (5.2)*. If Assumptions 1, 3 hold, then* (4.6) *holds with $A = L^2(1 + \frac{4d}{K})$, $B = \frac{8\mathbb{E}[G_S^2]}{n}(1 + \frac{2d}{K})$.*

Proof of Lemma 21. For $\mathbf{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, define the one-direction gradient estimators

$$\widehat{g}_{\mathbf{w}}(\mathbf{w}, \mathbf{v}; z, \mathbf{u}) = \frac{f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z) - f(\mathbf{w} - \mu_1 \mathbf{u}, \mathbf{v}; z)}{2\mu_1/d} \mathbf{u},$$
$$\widehat{g}_{\mathbf{v}}(\mathbf{w}, \mathbf{v}; z, \mathbf{u}) = \frac{f(\mathbf{w}, \mathbf{v} + \mu_2 \mathbf{u}; z) - f(\mathbf{w}, \mathbf{v} - \mu_2 \mathbf{u}; z)}{2\mu_2/d} \mathbf{u}.$$

Since $\mathbb{E}[X^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] + (\mathbb{E}[X])^2$ for a random variable X, we know

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}\left[\left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} \right\|_2^2\right] \\
&= \left\| \mathbb{E}_{\mathbf{u}}\left[\begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} \right] \right\|_2^2 \\
&\quad + \mathbb{E}_{\mathbf{u}}\left[\left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} - \mathbb{E}_{\mathbf{u}}\left[\begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} \right] \right\|_2^2\right].
\end{aligned} \tag{B.14}$$

Since an average of K independent random vectors divides the variance by a factor of K, we know

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}\left[\left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} - \mathbb{E}_{\mathbf{u}}\left[\begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} \right] \right\|_2^2\right] \\
&\le \frac{1}{K}\mathbb{E}_{\mathbf{u}}\left[\left\| \begin{pmatrix} \widehat{g}_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}, \mathbf{u}_{1,t,1}) - \widehat{g}_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, \mathbf{u}_{1,t,1}) \\ \widehat{g}_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}, \mathbf{u}_{2,t,1}) - \widehat{g}_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}, \mathbf{u}_{2,t,1}) \end{pmatrix} \right\|_2^2\right],
\end{aligned} \tag{B.15}$$

where we have also used the inequality $\mathbb{E}[(X - \mathbb{E}[X])^2] \le \mathbb{E}[X^2]$.

We next establish an empirical second-moment bound for these one-direction estimators. For fixed $\mathbf{w}, \mathbf{v}$ and z_i, let

$$h_i(\mathbf{u}) = f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z_i) - f(\mathbf{w} - \mu_1 \mathbf{u}, \mathbf{v}; z_i).$$

Since $h_i(-\mathbf{u}) = -h_i(\mathbf{u})$, we have $\mathbb{E}_{\mathbf{u}}[h_i(\mathbf{u})] = 0$. By Lemma 19,

$$\mathbb{E}_{\mathbf{u}}[h_i(\mathbf{u})^2] \le \frac{1}{d-1}\mathbb{E}_{\mathbf{u}}\big[\|\nabla_{\mathbb{S}} h_i(\mathbf{u})\|_2^2\big],$$

where $\nabla_{\mathbb{S}}$ denotes the spherical gradient. Furthermore,

$$\begin{aligned}
\|\nabla_{\mathbb{S}} h_i(\mathbf{u})\|_2^2 &\le \mu_1^2 \big\|\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z_i) + \nabla_{\mathbf{w}} f(\mathbf{w} - \mu_1 \mathbf{u}, \mathbf{v}; z_i)\big\|_2^2 \\
&\le 2\mu_1^2 \Big(\|\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z_i)\|_2^2 + \|\nabla_{\mathbf{w}} f(\mathbf{w} - \mu_1 \mathbf{u}, \mathbf{v}; z_i)\|_2^2 \Big).
\end{aligned}$$

Then, by Assumption 3,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}\left[\frac{1}{n}\sum_{i=1}^n \|\widehat{g}_{\mathbf{w}}(\mathbf{w}, \mathbf{v}; z_i, \mathbf{u})\|_2^2\right] = \frac{d^2}{4\mu_1^2}\mathbb{E}_{\mathbf{u}}\Big[\frac{1}{n}\sum_{i=1}^n h_i(\mathbf{u})^2\Big] \le \frac{d^2}{4\mu_1^2(d-1)}\mathbb{E}_{\mathbf{u}}\Big[\frac{1}{n}\sum_{i=1}^n \|\nabla_{\mathbb{S}} h_i(\mathbf{u})\|_2^2\Big] \\
&\le \frac{d^2}{2(d-1)}\mathbb{E}_{\mathbf{u}}\left[\frac{1}{n}\sum_{i=1}^n \Big(\|\nabla_{\mathbf{w}} f(\mathbf{w} + \mu_1 \mathbf{u}, \mathbf{v}; z_i)\|_2^2 + \|\nabla_{\mathbf{w}} f(\mathbf{w} - \mu_1 \mathbf{u}, \mathbf{v}; z_i)\|_2^2 \Big)\right] \\
&\le \frac{d^2}{d-1} G_S^2 \le 2dG_S^2,
\end{aligned} \tag{B.16}$$

where we have used $d/(d-1) \le 2$ for $d \ge 2$. Analogously,

$$\mathbb{E}_{\mathbf{u}}\left[\frac{1}{n}\sum_{i=1}^n \|\widehat{g}_{\mathbf{v}}(\mathbf{w}, \mathbf{v}; z_i, \mathbf{u})\|_2^2\right] \le \frac{d^2}{d-1} G_S^2 \le 2dG_S^2. \tag{B.17}$$

We now estimate the right-hand side of Eq. (B.15) by considering the following two cases.

Case 1: $i_t = i$, which holds with probability $1/n$. Using $(a+b)^2 \le 2a^2 + 2b^2$, Eqs. (B.16) and (B.17), and the symmetry between $(S^{(i)}, z_i')$ and (S, z_i), we obtain

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{u}}\left[\left\|\begin{pmatrix}\widehat{g}_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i',\mathbf{u}_{1,t,1}) - \widehat{g}_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_i,\mathbf{u}_{1,t,1})\\ \widehat{g}_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_i,\mathbf{u}_{2,t,1}) - \widehat{g}_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i',\mathbf{u}_{2,t,1})\end{pmatrix}\right\|_2^2\right]\right] \le 16d\mathbb{E}[G_S^2]. \quad \text{(B.18)}$$

Case 2: $i_t = j \neq i$, which holds with probability $1 - 1/n$. Define

$$h_{\mathbf{w},j}^{(i)}(\mathbf{u}) = \big[f(\mathbf{w}_t^{(i)} + \mu_1\mathbf{u}, \mathbf{v}_t^{(i)}; z_j) - f(\mathbf{w}_t^{(i)} - \mu_1\mathbf{u}, \mathbf{v}_t^{(i)}; z_j)\big] - \big[f(\mathbf{w}_t + \mu_1\mathbf{u}, \mathbf{v}_t; z_j) - f(\mathbf{w}_t - \mu_1\mathbf{u}, \mathbf{v}_t; z_j)\big].$$

Since $h_{\mathbf{w},j}^{(i)}(-\mathbf{u}) = -h_{\mathbf{w},j}^{(i)}(\mathbf{u})$, we have $\mathbb{E}_{\mathbf{u}}[h_{\mathbf{w},j}^{(i)}(\mathbf{u})] = 0$. Furthermore, Lemma 20 implies

$$\|\nabla_{\mathbb{S}} h_{\mathbf{w},j}^{(i)}(\mathbf{u})\|_2 \le 2\mu_1 L\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}_t^{(i)}\\ \mathbf{v}_t - \mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2.$$

It then follows from Lemma 19 that

$$\mathbb{E}_{\mathbf{u}}\Big[\big\|\widehat{g}_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j,\mathbf{u}) - \widehat{g}_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_j,\mathbf{u})\big\|_2^2\Big] = \frac{d^2}{4\mu_1^2}\mathbb{E}_{\mathbf{u}}\big[h_{\mathbf{w},j}^{(i)}(\mathbf{u})^2\big]$$
$$\le \frac{d^2}{4\mu_1^2(d-1)}\mathbb{E}_{\mathbf{u}}\big[\|\nabla_{\mathbb{S}} h_{\mathbf{w},j}^{(i)}(\mathbf{u})\|_2^2\big] \le \frac{d^2L^2}{d-1}\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}_t^{(i)}\\ \mathbf{v}_t - \mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2 \le 2dL^2\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}_t^{(i)}\\ \mathbf{v}_t - \mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2.$$

The same argument applies to the estimator with respect to $\mathbf{v}$. Then,

$$\mathbb{E}_{\mathbf{u}}\left[\left\|\begin{pmatrix}\widehat{g}_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j,\mathbf{u}_{1,t,1}) - \widehat{g}_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_j,\mathbf{u}_{1,t,1})\\ \widehat{g}_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_j,\mathbf{u}_{2,t,1}) - \widehat{g}_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j,\mathbf{u}_{2,t,1})\end{pmatrix}\right\|_2^2\right] \le 4L^2d\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}_t^{(i)}\\ \mathbf{v}_t - \mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2. \quad \text{(B.19)}$$

Combining the two cases with Eq. (B.15), followed by taking an average over $i \in [n]$ and expectation, we obtain

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix} - \mathbb{E}_{\mathbf{u}}\left[\begin{pmatrix}g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right]\right\|_2^2\right]$$
$$\le \frac{4L^2d}{K}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}_t^{(i)}\\ \mathbf{v}_t - \mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2\right] + \frac{16d\,\mathbb{E}[G_S^2]}{Kn}. \quad \text{(B.20)}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbb{E}_{\mathbf{u}}\left[\begin{pmatrix}g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right]\right\|_2^2\right]\\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\end{pmatrix}\right\|_2^2\right]\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left[\left\|\begin{pmatrix}\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j) - \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_j)\\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_j) - \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_j)\end{pmatrix}\right\|_2^2\right]\\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\begin{pmatrix}\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i') - \nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_i)\\ \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t,\mathbf{v}_t;z_i) - \nabla_{\mathbf{v}}\tilde{f}_2(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_i')\end{pmatrix}\right\|_2^2\right]\\
&\le L^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}_t^{(i)}\\ \mathbf{v}_t - \mathbf{v}_t^{(i)}\end{pmatrix}\right\|_2^2\right] + \frac{8\mathbb{E}[G_S^2]}{n}, \qquad \text{(B.21)}
\end{aligned}$$

where we use the smoothness of $\tilde{f}_1, \tilde{f}_2$ (Lemma 4) for the first term in the identity,

$$\left\| \begin{pmatrix} \nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_j) - \nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_j) \\ \nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}_t, \mathbf{v}_t; z_j) - \nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_j) \end{pmatrix} \right\|_2^2 \leq L^2 \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2^2,$$

and Lemma 18 for the second term in the last identity.

Finally, taking an average over $i \in [n]$ and expectation in Eq. (B.14), and then plugging Eqs. (B.20) and (B.21) into it, we obtain

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)}; z_{i_t}^{(i)}) \end{pmatrix} \right\|_2^2 \right] \\ &\leq L^2 \Big(1 + \frac{4d}{K}\Big) \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2^2 \right] + \frac{8\mathbb{E}[G_S^2]}{n} \Big(1 + \frac{2d}{K}\Big). \end{aligned}$$

The proof is completed. □

We then apply Lemma 21 and Lemma 18 to derive Theorem 5. We omit the details for brevity.

We are now ready to control the optimization error for Zeroth-order SGDA, which requires the following lemmas and we show the proofs here.

Lemma 22. *If f is G_z-Lipschitz continuous w.r.t $\mathbf{w}$, then the function*

$$h(\mathbf{u}) = f(\mathbf{w} + \mu\mathbf{u}, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}, \mathbf{v}; z)$$

with $\mathbf{u} \in \mathbb{S}^{d-1}$ is $2\mu G_z$-Lipschitz continuous w.r.t $\mathbf{u}$.

Proof. Take arbitrary $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{S}^{d-1}$. Decompose the difference:

$$\begin{aligned} &h(\mathbf{u}_1) - h(\mathbf{u}_2) \\ &= [f(\mathbf{w} + \mu\mathbf{u}_1, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}_1, \mathbf{v}; z)] - [f(\mathbf{w} + \mu\mathbf{u}_2, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}_2, \mathbf{v}; z)] \\ &= [f(\mathbf{w} + \mu\mathbf{u}_1, \mathbf{v}; z) - f(\mathbf{w} + \mu\mathbf{u}_2, \mathbf{v}; z)] + [f(\mathbf{w} - \mu\mathbf{u}_2, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}_1, \mathbf{v}; z)]. \end{aligned}$$

For the first term: by G_z-Lipschitz continuity of f,

$$\|f(\mathbf{w} + \mu\mathbf{u}_1, \mathbf{v}; z) - f(\mathbf{w} + \mu\mathbf{u}_2, \mathbf{v}; z)\| \leq G_z \|\mu\mathbf{u}_1 - \mu\mathbf{u}_2\| = \mu G_z \|\mathbf{u}_1 - \mathbf{u}_2\|.$$

For the second term: similarly,

$$\|f(\mathbf{w} - \mu\mathbf{u}_2, \mathbf{v}; z) - f(\mathbf{w} - \mu\mathbf{u}_1, \mathbf{v}; z)\| \leq G_z \| - \mu\mathbf{u}_2 + \mu\mathbf{u}_1\| = \mu G_z \|\mathbf{u}_1 - \mathbf{u}_2\|.$$

Combining via triangle inequality gives $|h(\mathbf{u}_1) - h(\mathbf{u}_2)| \leq 2\mu G_z \|\mathbf{u}_1 - \mathbf{u}_2\|$. □

Lemma 23. *Let Assumption 3 hold. Let $g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ and $g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ be defined in Eq.* (5.2)*. Then*

$$\begin{aligned} \mathbb{E}_{i_t, \mathbf{u}}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] &\leq \mathbb{E}_{i_t}[\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] + \frac{2dG_S^2}{K}, \\ \mathbb{E}_{i_t, \mathbf{u}}[\|g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] &\leq \mathbb{E}_{i_t}[\|\nabla_{\mathbf{v}} \tilde{f}_2(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] + \frac{2dG_S^2}{K}. \end{aligned}$$

Proof. Since $\mathbb{E}[X^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] + (\mathbb{E}[X])^2$ for a random variable X, we know

$$\mathbb{E}_{\mathbf{u}}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] = \|\mathbb{E}_{\mathbf{u}}[g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})]\|_2^2 + \mathbb{E}_{\mathbf{u}}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \mathbb{E}_{\mathbf{u}}[g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})]\|_2^2]. \tag{B.22}$$

Since an average of K independent random variables divides the variance by a factor of K, we know

$$\begin{aligned} &\mathbb{E}_{\mathbf{u}}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \mathbb{E}_{\mathbf{u}}[g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})]\|_2^2] \leq \\ &\qquad \frac{1}{K} \mathbb{E}_{\mathbf{u}}\Big[\Big\| \frac{f(\mathbf{w}_t + \mu_1 \mathbf{u}_{1,t,1}, \mathbf{v}_t; z_{i_t}) - f(\mathbf{w}_t - \mu_1 \mathbf{u}_{1,t,1}, \mathbf{v}_t; z_{i_t})}{2\mu_1/d} \mathbf{u}_{1,t,1} \Big\|_2^2 \Big]. \end{aligned} \tag{B.23}$$

For each $i \in [n]$, define

$$h_i(\mathbf{u}) = f(\mathbf{w}_t + \mu_1 \mathbf{u}, \mathbf{v}_t; z_i) - f(\mathbf{w}_t - \mu_1 \mathbf{u}, \mathbf{v}_t; z_i).$$

Since $h_i(-\mathbf{u}) = -h_i(\mathbf{u})$, we have $\mathbb{E}_{\mathbf{u}}[h_i(\mathbf{u})] = 0$. By Lemma 19,

$$\mathbb{E}_{\mathbf{u}}[h_i(\mathbf{u})^2] \leq \frac{1}{d-1}\mathbb{E}_{\mathbf{u}}\big[\|\nabla_{\mathbb{S}} h_i(\mathbf{u})\|_2^2\big].$$

By the definition in Lemma 19,

$$\begin{aligned}\nabla_{\mathbb{S}} h_i(\mathbf{u}) &= (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top)\nabla_{\mathbf{u}} h_i(\mathbf{u}) \\ &= (\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top)\big(\mu_1 \nabla_{\mathbf{w}} f(\mathbf{w}_t + \mu_1 \mathbf{u}, \mathbf{v}_t; z_i) + \mu_1 \nabla_{\mathbf{w}} f(\mathbf{w}_t - \mu_1 \mathbf{u}, \mathbf{v}_t; z_i)\big).\end{aligned}$$

Note $\|(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top)\mathbf{y}\|_2 \leq \|\mathbf{y}\|_2$ for any $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{S}^{d-1}$. Applying this with $\mathbf{y} = \nabla_{\mathbf{u}} h_i(\mathbf{u})$ and using $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\begin{aligned}\|\nabla_{\mathbb{S}} h_i(\mathbf{u})\|_2^2 &\leq \|\nabla_{\mathbf{u}} h_i(\mathbf{u})\|_2^2 \\ &= \mu_1^2 \big\|\nabla_{\mathbf{w}} f(\mathbf{w}_t + \mu_1 \mathbf{u}, \mathbf{v}_t; z_i) + \nabla_{\mathbf{w}} f(\mathbf{w}_t - \mu_1 \mathbf{u}, \mathbf{v}_t; z_i)\big\|_2^2 \\ &\leq 2\mu_1^2 \Big(\|\nabla_{\mathbf{w}} f(\mathbf{w}_t + \mu_1 \mathbf{u}, \mathbf{v}_t; z_i)\|_2^2 + \|\nabla_{\mathbf{w}} f(\mathbf{w}_t - \mu_1 \mathbf{u}, \mathbf{v}_t; z_i)\|_2^2\Big).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{i_t,\mathbf{u}}\Big[\Big\|\frac{f(\mathbf{w}_t + \mu_1 \mathbf{u}_{1,t,1}, \mathbf{v}_t; z_{i_t}) - f(\mathbf{w}_t - \mu_1 \mathbf{u}_{1,t,1}, \mathbf{v}_t; z_{i_t})}{2\mu_1/d}\mathbf{u}_{1,t,1}\Big\|_2^2\Big] \\ &= \frac{d^2}{4\mu_1^2}\mathbb{E}_{i_t,\mathbf{u}}[h_{i_t}(\mathbf{u})^2] = \frac{d^2}{4\mu_1^2}\cdot\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{u}}[h_i(\mathbf{u})^2] \leq \frac{d^2}{4\mu_1^2(d-1)}\cdot\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{u}}\big[\|\nabla_{\mathbb{S}} h_i(\mathbf{u})\|_2^2\big] \\ &\leq \frac{d^2}{2(d-1)}\cdot\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{u}}\Big[\|\nabla_{\mathbf{w}} f(\mathbf{w}_t + \mu_1 \mathbf{u}, \mathbf{v}_t; z_i)\|_2^2 + \|\nabla_{\mathbf{w}} f(\mathbf{w}_t - \mu_1 \mathbf{u}, \mathbf{v}_t; z_i)\|_2^2\Big] \\ &\leq \frac{d^2}{d-1}G_S^2 \leq 2dG_S^2,\end{aligned}$$

where we use Assumption 3 and $d/(d-1) \leq 2$ for $d \geq 2$. Plugging into Eq. (B.23), we get

$$\mathbb{E}_{i_t,\mathbf{u}}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \mathbb{E}_{\mathbf{u}}[g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})]\|_2^2] \leq \frac{2dG_S^2}{K}. \tag{B.24}$$

Furthermore, we know

$$\|\mathbb{E}_{\mathbf{u}}[g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})]\|_2^2 = \|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2. \tag{B.25}$$

Taking expectation over i_t in Eq. (B.22), and plugging Eq. (B.24) and Eq. (B.25) into it, we know

$$\mathbb{E}_{i_t,\mathbf{u}}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] \leq \mathbb{E}_{i_t}[\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] + \frac{2dG_S^2}{K}.$$

The proof for $\mathbb{E}_{i_t,\mathbf{u}}[\|g_{\mathbf{v}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2]$ holds similarly. □

The following lemma presents convergence rate analysis for Zeroth-order SGDA.

Proof of Lemma 6. Note that the projection step is nonexpansive. For any $\mathbf{w} \in B_R$, we know

$$\begin{aligned}&\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \|\mathbf{w}_t - \eta_t g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \mathbf{w}\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}\|_2^2 - 2\eta_t\langle \mathbf{w}_t - \mathbf{w}, g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + \eta_t^2\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\rangle + 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\rangle \\ &+ \eta_t^2\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 2\eta_t\big(\widetilde{F}_{1,S}(\mathbf{w}, \mathbf{v}_t) - \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\big) \\ &+ 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\rangle + \eta_t^2\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2,\end{aligned}$$

where we have used the convexity of $\widetilde{F}_{1,S}$ established in Lemma 4. Taking a summation of the above inequality from $t = 1$ to $t = T$ ($\mathbf{w}_1 = 0$), we derive

$$\sum_{t=1}^{T} 2\eta_t(\widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - \widetilde{F}_{1,S}(\mathbf{w}, \mathbf{v}_t)) \leq \|\mathbf{w}\|_2^2 + \sum_{t=1}^{T} \eta_t^2 \|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2 + 2\sum_{t=1}^{T} \eta_t \langle \mathbf{w} - \mathbf{w}_t, g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\rangle.$$

Since $\eta_t = \eta$, we further get

$$\sum_{t=1}^{T} (\widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - \widetilde{F}_{1,S}(\mathbf{w}, \mathbf{v}_t)) \leq \frac{1}{2}\eta^{-1}\|\mathbf{w}\|_2^2 + \frac{1}{2}\eta \sum_{t=1}^{T} \|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2 + \sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + \sum_{t=1}^{T} \langle \mathbf{w}, g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\rangle.$$

It then follows from the concavity of $\widetilde{F}_{1,S}(\mathbf{w}, \cdot)$ and Schwartz's inequality that

$$\sum_{t=1}^{T} (\widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - \widetilde{F}_{1,S}(\mathbf{w}, \bar{\mathbf{v}}_T)) \leq \frac{1}{2}\eta^{-1}\|\mathbf{w}\|_2^2 + \frac{1}{2}\eta \sum_{t=1}^{T} \|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2 + \sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + \|\mathbf{w}\|_2 \Big\| \sum_{t=1}^{T} \big(g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2.$$

Since the above inequality holds for all $\mathbf{w} \in B_R$, we further get

$$\sum_{t=1}^{T} (\widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - \inf_{\mathbf{w} \in B_R} \widetilde{F}_{1,S}(\mathbf{w}, \bar{\mathbf{v}}_T)) \leq \frac{1}{2}\eta^{-1}R^2 + \frac{1}{2}\eta \sum_{t=1}^{T} \|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2 + \sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + R \Big\| \sum_{t=1}^{T} \big(g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2. \tag{B.26}$$

According to Lemma 23 and Assumption 3,

$$\mathbb{E}[\|g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] \leq \mathbb{E}[\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] + \frac{cd\mathbb{E}[G_S^2]}{K}. \tag{B.27}$$

Then, taking an expectation on both sides of Eq. (B.26), noting Eq. (B.27) and

$$\mathbb{E}_{i_t}[\langle \mathbf{w}_t, \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle] = 0, \tag{B.28}$$

we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t) - \inf_{\mathbf{w} \in B_R} \widetilde{F}_{1,S}(\mathbf{w}, \bar{\mathbf{v}}_T)] \\
&\leq \frac{R^2}{2\eta} + R\mathbb{E}\Big[\Big\| \sum_{t=1}^{T} \big(g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2\Big] + \frac{\eta}{2} \sum_{t=1}^{T} \big(\mathbb{E}[\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2] + \frac{2d\mathbb{E}[G_S^2]}{K}\big) \\
&\leq \frac{1}{2}\eta^{-1}R^2 + \frac{d\mathbb{E}[G_S^2]}{K}\eta T + R\mathbb{E}\Big[\Big\| \sum_{t=1}^{T} \big(g_{\mathbf{w}}(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla \widetilde{F}_{1,S}(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2\Big] \\
&+ \frac{1}{2}\eta \sum_{t=1}^{T} \mathbb{E}[\|\nabla_{\mathbf{w}} \tilde{f}_1(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\|_2^2].
\end{aligned} \tag{B.29}$$

According to Jensen's inequality and $\mathbb{E}[(X-\mathbb{E}[X])^2] \leq \mathbb{E}[X^2]$ (note that $\mathbb{E}_{\mathbf{u},i_t}[g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})] = \nabla \widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)$ by Lemma 4), we know

$$\Big(\mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\big(g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)\big)\Big\|_2\Big]\Big)^2 \leq \mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\big(g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)\big)\Big\|_2^2\Big]$$
$$= \sum_{t=1}^{T}\mathbb{E}\Big[\Big\|g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)\Big\|_2^2\Big] \leq \sum_{t=1}^{T}\mathbb{E}[\|g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2],$$

from which and Eq. (B.27), we get

$$R\mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\big(g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)\big)\Big\|_2\Big] \leq \frac{R^2}{4\eta}+\Big(\mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\big(g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)\big)\Big\|_2\Big]\Big)^2\eta$$
$$\leq \frac{\eta^{-1}R^2}{4}+\eta\sum_{t=1}^{T}\mathbb{E}[\|g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2] \leq \frac{\eta^{-1}R^2}{4}+\eta\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2]+\frac{2d\mathbb{E}[G_S^2]}{K}\eta T,$$

where we use Cauchy-Schwarz's inequality for the first step. Plugging the above inequality into Eq. (B.29), we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in B_R}\widetilde{F}_{1,S}(\mathbf{w},\bar{\mathbf{v}}_T)] &\leq \frac{3R^2}{4\eta}+\frac{3d\mathbb{E}[G_S^2]\eta T}{K}+\frac{3}{2}\eta\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\mathbf{w}}\tilde{f}_1(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2]\\ &\leq \frac{3R^2}{4\eta}+\frac{3d\mathbb{E}[G_S^2]}{K}\eta T+\frac{3}{2}\mathbb{E}[G_S^2]\eta T,\end{aligned}$$

where we use Lemma 4 and Assumption 3 for the last step. This yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\widetilde{F}_{1,S}(\mathbf{w}_t,\mathbf{v}_t)]-\mathbb{E}[\inf_{\mathbf{w}\in B_R}\widetilde{F}_{1,S}(\mathbf{w},\bar{\mathbf{v}}_T)] \lesssim \frac{R^2}{\eta T}+\big(1+\frac{d}{K}\big)\mathbb{E}[G_S^2]\eta.$$

Moreover, by the mean-value theorem and Cauchy–Schwarz's inequality, we know

$$\begin{aligned}\big|\widetilde{F}_{1,S}(\mathbf{w},\mathbf{v})-F_S(\mathbf{w},\mathbf{v})\big| &= \Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{u}}\big[f(\mathbf{w}+\mu_1\mathbf{u},\mathbf{v};z_i)-f(\mathbf{w},\mathbf{v};z_i)\big]\Big|\\ &\leq \mu_1\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{u}}\big[\|\nabla_{\mathbf{w}}f(\mathbf{w}+\theta_i\mu_1\mathbf{u},\mathbf{v};z_i)\|_2^2\big]\Big)^{\frac{1}{2}} \leq \mu_1 G_S,\end{aligned}$$

for some $\theta_i \in [0,1]$. Taking expectation then gives

$$\big|\mathbb{E}[\widetilde{F}_{1,S}(\mathbf{w},\mathbf{v})-F_S(\mathbf{w},\mathbf{v})]\big| \leq \mu_1\sqrt{\mathbb{E}[G_S^2]}.$$

Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t,\mathbf{v}_t)]-\mathbb{E}[\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)] \lesssim \frac{R^2}{\eta T}+\big(1+\frac{d}{K}\big)\mathbb{E}[G_S^2]\eta+\mu_1\sqrt{\mathbb{E}[G_S^2]}.$$

In a similar way, we can show that

$$\mathbb{E}[\sup_{\mathbf{v}\in B_R}F_S(\bar{\mathbf{w}}_T,\mathbf{v})]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t,\mathbf{v}_t)] \lesssim \frac{R^2}{\eta T}+\big(1+\frac{d}{K}\big)\mathbb{E}[G_S^2]\eta+\mu_2\sqrt{\mathbb{E}[G_S^2]}.$$

Then,

$$\mathbb{E}[\sup_{\mathbf{v}\in B_R}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)] \lesssim \frac{R^2}{\eta T}+\big(1+\frac{d}{K}\big)\mathbb{E}[G_S^2]\eta+\mu\sqrt{\mathbb{E}[G_S^2]}.$$

The proof is completed. □

Combining Theorem 5 and Lemma 6, we derive the following theorem.

Proof of Theorem 7. Define

$$\triangle_S^{w,R}(\mathbf{w},\mathbf{v}) = \sup_{\mathbf{v}'\in B_R} \mathbb{E}\big[F_S(\mathbf{w},\mathbf{v}')\big] - \inf_{\mathbf{w}'\in B_R} \mathbb{E}\big[F_S(\mathbf{w}',\mathbf{v})\big].$$

We have the decomposition

$$\Delta^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) = \Delta^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) - \Delta_S^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) + \Delta_S^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T). \tag{B.30}$$

From Theorem 5, we know

$$\epsilon \lesssim \Big(\frac{\sqrt{\mathbb{E}[G_S^2]}}{\sqrt{n}}\Big(1+\frac{d^{\frac{1}{2}}}{K^{\frac{1}{2}}}\Big)\Big)\eta\sqrt{T} + \frac{\sqrt{\mathbb{E}[G_S^2]}}{n}\eta T \lesssim \frac{\sqrt{\mathbb{E}[G_S^2]}/L}{\sqrt{n}} + \frac{\sqrt{\mathbb{E}[G_S^2]}}{n}\eta T,$$

where we used $L^2\big(1+\frac{4d}{K}\big)\eta^2 \le \frac{1}{T}$ in Theorem 5.

Furthermore, Assumption 3 implies that the data- and algorithm-dependent constant $G_{S,A}$ in Assumption 2 satisfies $\mathbb{E}_{S,A}[G_{S,A}^2] \le \mathbb{E}[G_S^2]$. According to Theorem 1 and Eq. (5.4) $L\epsilon^2 \lesssim \sqrt{\mathbb{E}[G_S^2]}\epsilon$, we know

$$\triangle^{w,R}(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S)) - \triangle_S^{w,R}(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S)) \lesssim \frac{\mathbb{E}[G_S^2]/L}{\sqrt{n}} + \frac{\mathbb{E}[G_S^2]}{n}\eta T.$$

According to Lemma 6, we know

$$\Delta_S^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) \lesssim \frac{R^2}{\eta T} + \Big(1+\frac{d}{K}\Big)\mathbb{E}[G_S^2]\eta + \mu\sqrt{\mathbb{E}[G_S^2]} \lesssim \frac{R^2}{\eta T} + \frac{\mathbb{E}[G_S^2]}{L^2\eta T} + \mu\sqrt{\mathbb{E}[G_S^2]},$$

where we use the condition $L^2\big(1+\frac{4d}{K}\big)\eta^2 \le 1/T$ for the last step. Combining the above two inequalities with Eq. (B.30), we obtain

$$\Delta^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) \lesssim \frac{\mathbb{E}[G_S^2]/L}{\sqrt{n}} + \frac{\mathbb{E}[G_S^2]}{n}\eta T + \frac{R^2}{\eta T} + \frac{\mathbb{E}[G_S^2]}{L^2\eta T} + \mu\sqrt{\mathbb{E}[G_S^2]}.$$

The bound then follows directly. □

B.3.3 Clipped-SGDA

To prove verify Eq. (4.6) for Clipped-SGDA, we introduce the following lemma which shows that the clipping operator is non-expansive. Actually, one can show that the clipping operator is a projector onto a convex set. We omit the proof for simplicity.

Lemma 24. *For any $\mathbf{v}_1,\mathbf{v}_2 \in \mathcal{W}$, we have $\|\mathrm{clip}(\mathbf{v}_1,\tau) - \mathrm{clip}(\mathbf{v}_2,\tau)\|_2 \le \|\mathbf{v}_1-\mathbf{v}_2\|_2$.*

Then, we apply Lemma 24 to verify Eq. (4.6) for Clipped-SGDA in the following lemma.

Lemma 25. *Let $g(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})$ be given in Eq.* (5.6)*, and Assumptions 1, 3 hold with $G_S \le G$ for some $G>0$. Then, Eq.* (4.6) *holds with $A = L^2, B = \frac{8G^2}{n}$.*

Proof of Lemma 25. According to Eq. (5.6) and Lemma 24, we know

$$\begin{aligned}
&\left\| \begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \end{pmatrix} \right\|_2^2 \\
&\le \|\nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - \nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2 + \|\nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)})\|_2^2 \\
&= \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - \nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \end{pmatrix} \right\|_2^2.
\end{aligned} \tag{B.31}$$

It then follows that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} g_{\mathbf{w}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - g_{\mathbf{w}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \\ g_{\mathbf{v}}(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - g_{\mathbf{v}}(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \end{pmatrix}\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) - \nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t^{(i)},\mathbf{v}_t^{(i)};z_{i_t}^{(i)}) \end{pmatrix}\right\|_2^2\right] \\
&\leq L^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\|\begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix}\right\|_2^2\right] + \frac{8\mathbb{E}[G_S^2]}{n}, \qquad \text{(B.32)}
\end{aligned}$$

where we use Eq. (B.9) for the last step. The proof is completed. □

The following lemma shows the Lipschitz continuity of the bias of clipping operator.

Lemma 26 ([56])**.** *For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{W}$, we have*

$$\|(\mathrm{clip}(\mathbf{v}_1,\tau) - \mathrm{clip}(\mathbf{v}_2,\tau)) - (\mathbf{v}_1 - \mathbf{v}_2)\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_2 \mathbb{I}\big[\|\mathbf{v}_1\|_2 > \tau \text{ or } \|\mathbf{v}_2\|_2 > \tau\big], \quad \text{(B.33)}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function, i.e., returning 1 if the argument holds and 0 otherwise.

We can easily generalize Lemma 26 with two variables $\mathbf{v}_1, \mathbf{v}_2$ in the following lemma, which shows the Lipschitz continuity of the bias of the clipping operator with four variables $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$. We omit the details for brevity.

Lemma 27. *For any $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$, we have*

$$\begin{aligned}
&\left\|\begin{pmatrix} \mathrm{clip}(\mathbf{v}_1,\tau_1) - \mathrm{clip}(\mathbf{v}_2,\tau_1)) - (\mathbf{v}_1 - \mathbf{v}_2) \\ \mathrm{clip}(\mathbf{v}_3,\tau_2) - \mathrm{clip}(\mathbf{v}_4,\tau_2)) - (\mathbf{v}_3 - \mathbf{v}_4) \end{pmatrix}\right\|_2 \\
&\qquad \leq \left\|\begin{pmatrix} \mathbf{v}_1 - \mathbf{v}_2 \\ \mathbf{v}_3 - \mathbf{v}_4 \end{pmatrix}\right\|_2 \mathbb{I}\big[\max\{\|\mathbf{v}_1\|_2/\tau_1, \|\mathbf{v}_2\|_2/\tau_1, \|\mathbf{v}_3\|_2/\tau_2, \|\mathbf{v}_4\|_2/\tau_2\} > 1\big].
\end{aligned}$$

We now prove that the bias of Clipped-SGDA satisfies a generalized Lipschitz-type condition in Lemma 29, which uses Lemma 27 and Lemma 28. Lemma 28 helps deal with the indicator function in the proof of Lemma 29. Let S_r denote the r-th elemtn of S.

Lemma 28. *Let $\mathbf{w}, \mathbf{w}' \in \mathcal{W}, \mathbf{v}, \mathbf{v}' \in \mathcal{V}$ and r follow the uniform distribution over $[n]$. Denote $\theta = (\mathbf{w}, \mathbf{v}), \theta' = (\mathbf{w}', \mathbf{v}')$. Then*

$$\begin{aligned}
&\frac{1}{n}\sum_{j:j\neq i}\mathbb{I}\big[\max\{\|\nabla_{\mathbf{w}} f(\theta;z_j)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta';z_j)\|_2/\tau_1, \|\nabla_{\mathbf{v}} f(\theta';z_j)\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta;z_j)\|_2/\tau_2\} > 1\big] \\
&\quad \leq \frac{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\theta;S_r)\|_2^2]}{\tau_1^2} + \frac{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\theta';S_r^{(i)})\|_2^2]}{\tau_1^2} + \frac{\mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\theta;S_r)\|_2^2]}{\tau_2^2} + \frac{\mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\theta';S_r^{(i)})\|_2^2]}{\tau_2^2}.
\end{aligned}$$

Proof. We know

$$\begin{aligned}
&\frac{1}{n}\sum_{j:j\neq i}\mathbb{I}\big[\max\{\|\nabla_{\mathbf{w}} f(\theta;z_j)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta';z_j)\|_2/\tau_1, \|\nabla_{\mathbf{v}} f(\theta';z_j)\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta;z_j)\|_2/\tau_2\} > 1\big] \\
&\leq \mathbb{E}_r\big[\mathbb{I}\big[\max\{\|\nabla_{\mathbf{w}} f(\theta;S_r)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta';S_r^{(i)})\|_2/\tau_1, \|\nabla_{\mathbf{v}} f(\theta';S_r^{(i)})\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta;S_r)\|_2/\tau_2\} > 1\big]\big] \\
&\leq \Pr\big\{\|\nabla_{\mathbf{w}} f(\theta;S_r)\|_2 > \tau_1\big\} + \Pr\big\{\|\nabla_{\mathbf{w}} f(\theta';S_r^{(i)})\|_2 > \tau_1\big\} + \Pr\big\{\|\nabla_{\mathbf{v}} f(\theta';S_r^{(i)})\|_2 > \tau_2\big\} \\
&+ \Pr\big\{\|\nabla_{\mathbf{v}} f(\theta;S_r)\|_2 > \tau_2\big\}.
\end{aligned}$$

By Markov's inequality, we know

$$\Pr\big\{\|\nabla_{\mathbf{w}} f(\theta;S_r)\|_2 > \tau_1\big\} = \Pr\big\{\|\nabla_{\mathbf{w}} f(\theta;S_r)\|_2^2 > \tau_1^2\big\} \leq \frac{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\theta;S_r)\|_2^2]}{\tau_1^2},$$

and similarly,

$$\Pr\big\{\|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2 > \tau_1\big\} \le \frac{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2^2]}{\tau_1^2},$$
$$\Pr\big\{\|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2 > \tau_2\big\} \le \frac{\mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2^2]}{\tau_2^2},$$
$$\Pr\big\{\|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2 > \tau_2\big\} \le \frac{\mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2^2]}{\tau_2^2}.$$

We complete the proof by combining all these inequalities together. □

Lemma 29. *Let $\mathbf{w}, \mathbf{w}' \in \mathcal{W}, \mathbf{v}, \mathbf{v}' \in \mathcal{V}$ and r follow the uniform distribution over $[n]$. Denote $\theta = (\mathbf{w}, \mathbf{v}), \theta' = (\mathbf{w}', \mathbf{v}')$. Then*

$$\begin{aligned}
&\left\| \begin{pmatrix} \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{w}} f(\theta; S_r), \tau_1)] - \nabla_{\mathbf{w}} F_S(\theta) - \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}), \tau_1)] + \nabla_{\mathbf{w}} F_{S^{(i)}}(\theta') \\ \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}), \tau_2)] - \nabla_{\mathbf{v}} F_{S^{(i)}}(\theta') - \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{v}} f(\theta; S_r), \tau_2)] + \nabla_{\mathbf{v}} F_S(\theta) \end{pmatrix} \right\|_2 \\
&\le L\|\theta - \theta'\|_2 \left(\frac{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\theta; S_r)\|_2^2 + \|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2^2]}{\tau_1^2} + \frac{\mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2^2 + \|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2^2]}{\tau_2^2} \right) \\
&\qquad + \frac{1}{n} \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; z_i) - \nabla_{\mathbf{w}} f(\theta'; z_i') \\ \nabla_{\mathbf{v}} f(\theta'; z_i') - \nabla_{\mathbf{v}} f(\theta; z_i) \end{pmatrix} \right\|_2.
\end{aligned}$$

Proof. Noting that

$$\nabla_{\mathbf{w}} F_S(\theta) = \mathbb{E}_r[\nabla_{\mathbf{w}} f(\theta; S_r)], \qquad \nabla_{\mathbf{w}} F_{S^{(i)}}(\theta') = \mathbb{E}_r[\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})],$$
$$\nabla_{\mathbf{v}} F_S(\theta) = \mathbb{E}_r[\nabla_{\mathbf{v}} f(\theta; S_r)], \qquad \nabla_{\mathbf{v}} F_{S^{(i)}}(\theta') = \mathbb{E}_r[\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})],$$

we know (by Jensen's inequality)

$$\begin{aligned}
&\left\| \begin{pmatrix} \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{w}} f(\theta; S_r), \tau_1)] - \nabla_{\mathbf{w}} F_S(\theta) - \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}), \tau_1)] + \nabla_{\mathbf{w}} F_{S^{(i)}}(\theta') \\ \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}), \tau_2)] - \nabla_{\mathbf{v}} F_{S^{(i)}}(\theta') - \mathbb{E}_r[\mathrm{clip}(\nabla_{\mathbf{v}} f(\theta; S_r), \tau_2)] + \nabla_{\mathbf{v}} F_S(\theta) \end{pmatrix} \right\|_2 \\
&\le \mathbb{E}_r \left[\left\| \begin{pmatrix} \mathrm{clip}(\nabla_{\mathbf{w}} f(\theta; S_r), \tau_1) - \nabla_{\mathbf{w}} f(\theta; S_r) - \mathrm{clip}(\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}), \tau_1) + \nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}) \\ \mathrm{clip}(\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}), \tau_2) - \nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}) - \mathrm{clip}(\nabla_{\mathbf{v}} f(\theta; S_r), \tau_2) + \nabla_{\mathbf{v}} f(\theta; S_r) \end{pmatrix} \right\|_2 \right] \\
&\le \mathbb{E}_r \left[\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; S_r) - \nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}) \\ \nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}) - \nabla_{\mathbf{v}} f(\theta; S_r) \end{pmatrix} \right\|_2 \mathbb{I}\left[\begin{matrix} \max\{\|\nabla_{\mathbf{w}} f(\theta; S_r)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2/\tau_1, \\ \|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2/\tau_2\} > 1 \end{matrix} \right] \right],
\end{aligned} \tag{B.34}$$

where we use Lemma 27 for the last step. Since r follows the uniform distribution over $[n]$, we know

$$\begin{aligned}
&\mathbb{E}_r \left[\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; S_r) - \nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}) \\ \nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}) - \nabla_{\mathbf{v}} f(\theta; S_r) \end{pmatrix} \right\|_2 \mathbb{I}\left[\begin{matrix} \max\{\|\nabla_{\mathbf{w}} f(\theta; S_r)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2/\tau_1, \\ \|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2/\tau_2\} > 1 \end{matrix} \right] \right] \\
&= \frac{1}{n} \sum_{j: j \ne i} \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; z_j) - \nabla_{\mathbf{w}} f(\theta'; z_j) \\ \nabla_{\mathbf{v}} f(\theta'; z_j) - \nabla_{\mathbf{v}} f(\theta; z_j) \end{pmatrix} \right\|_2 \mathbb{I}\left[\begin{matrix} \max\{\|\nabla_{\mathbf{w}} f(\theta; z_j)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta'; z_j)\|_2/\tau_1, \\ \|\nabla_{\mathbf{v}} f(\theta'; z_j)\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta; z_j)\|_2/\tau_2\} > 1 \end{matrix} \right] \\
&+ \frac{1}{n} \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; z_i) - \nabla_{\mathbf{w}} f(\theta'; z_i') \\ \nabla_{\mathbf{v}} f(\theta'; z_i') - \nabla_{\mathbf{v}} f(\theta; z_i) \end{pmatrix} \right\|_2 \mathbb{I}\left[\begin{matrix} \max\{\|\nabla_{\mathbf{w}} f(\theta; z_i)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta'; z_i')\|_2/\tau_1, \\ \|\nabla_{\mathbf{v}} f(\theta'; z_i')\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta; z_i)\|_2/\tau_2\} > 1 \end{matrix} \right].
\end{aligned}$$

By the smoothness of $\theta \mapsto f(\theta; z)$ (Assumption 1) and Lemma 28, we further get

$$\begin{aligned}
&\mathbb{E}_r \left[\left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; S_r) - \nabla_{\mathbf{w}} f(\theta'; S_r^{(i)}) \\ \nabla_{\mathbf{v}} f(\theta'; S_r^{(i)}) - \nabla_{\mathbf{v}} f(\theta; S_r) \end{pmatrix} \right\|_2 \mathbb{I}\left[\begin{matrix} \max\{\|\nabla_{\mathbf{w}} f(\theta; S_r)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2/\tau_1, \\ \|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2/\tau_2\} > 1 \end{matrix} \right] \right] \\
&\le L\|\theta - \theta'\|_2 \left(\frac{\mathbb{E}_r[\|\nabla_{\mathbf{w}} f(\theta; S_r)\|_2^2 + \|\nabla_{\mathbf{w}} f(\theta'; S_r^{(i)})\|_2^2]}{\tau_1^2} + \frac{\mathbb{E}_r[\|\nabla_{\mathbf{v}} f(\theta'; S_r^{(i)})\|_2^2 + \|\nabla_{\mathbf{v}} f(\theta; S_r)\|_2^2]}{\tau_2^2} \right) \\
&+ \frac{1}{n} \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta; z_i) - \nabla_{\mathbf{w}} f(\theta'; z_i') \\ \nabla_{\mathbf{v}} f(\theta'; z_i') - \nabla_{\mathbf{v}} f(\theta; z_i) \end{pmatrix} \right\|_2,
\end{aligned} \tag{B.35}$$

where we simply use the inequality

$$\mathbb{I}\big[\max\{\|\nabla_{\mathbf{w}} f(\theta;z_i)\|_2/\tau_1, \|\nabla_{\mathbf{w}} f(\theta';z_i')\|_2/\tau_1, \|\nabla_{\mathbf{v}} f(\theta';z_i')\|_2/\tau_2, \|\nabla_{\mathbf{v}} f(\theta;z_i)\|_2/\tau_2\} > 1\big] \leq 1.$$

Plugging Eq. (B.35) into Eq. (B.34) completes the proof. □

We now apply the above lemma to show that Eq. (4.7) holds for Clipped-SGDA.

Lemma 30. *Let $g(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})$ be defined in Eq.* (5.6)*. Let Assumptions 1, 3 hold with $G_S \leq G$ for some $G > 0$. Then, Eq.* (4.7) *holds with $\bar{A} = 8L^2G^4(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2})^2, \bar{B} = \frac{16G^2}{n^2}$.*

Proof of Lemma 30. Denote $\theta_t = (\mathbf{w}_t, \mathbf{v}_t)$ and $\theta_t^{(i)} = (\mathbf{w}_t^{(i)}, \mathbf{v}_t^{(i)})$. From Lemma 29, we know

$$\begin{aligned}
&\left\| \begin{pmatrix} b_{t,\mathbf{w}} - b_{t,\mathbf{w}}^{(i)} \\ b_{t,\mathbf{v}}^{(i)} - b_{t,\mathbf{v}} \end{pmatrix} \right\|_2 \\
&= \left\| \begin{pmatrix} \mathbb{E}_{i_t}[\mathrm{clip}(\nabla_{\mathbf{w}} f(\theta_t; z_{i_t}), \tau_1)] - \nabla_{\mathbf{w}} F_S(\theta_t) - \mathbb{E}_{i_t}[\mathrm{clip}(\nabla_{\mathbf{w}} f(\theta_t^{(i)}; z_{i_t}^{(i)}), \tau_1)] + \nabla_{\mathbf{w}} F_{S^{(i)}}(\theta_t^{(i)}) \\ \mathbb{E}_{i_t}[\mathrm{clip}(\nabla_{\mathbf{v}} f(\theta_t^{(i)}; z_{i_t}^{(i)}), \tau_2)] - \nabla_{\mathbf{v}} F_{S^{(i)}}(\theta_t^{(i)}) - \mathbb{E}_{i_t}[\mathrm{clip}(\nabla_{\mathbf{v}} f(\theta_t; z_{i_t}), \tau_2)] + \nabla_{\mathbf{v}} F_S(\theta_t) \end{pmatrix} \right\|_2 \\
&\leq \frac{1}{n} \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta_t; z_i) - \nabla_{\mathbf{w}} f(\theta_t^{(i)}; z_i') \\ \nabla_{\mathbf{v}} f(\theta_t^{(i)}; z_i') - \nabla_{\mathbf{v}} f(\theta_t; z_i) \end{pmatrix} \right\|_2 + 2LG_S^2 \Big(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2}\Big) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2, \qquad \text{(B.36)}
\end{aligned}$$

where we use Assumption 3 in the last step. It then follows from $(a+b)^2 \leq 2(a^2+b^2)$ that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\| \begin{pmatrix} b_{t,\mathbf{w}} - b_{t,\mathbf{w}}^{(i)} \\ b_{t,\mathbf{v}}^{(i)} - b_{t,\mathbf{v}} \end{pmatrix} \right\|_2^2\right] \\
&\leq \frac{2}{n^2}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\theta_t; z_i) - \nabla_{\mathbf{w}} f(\theta_t^{(i)}; z_i') \\ \nabla_{\mathbf{v}} f(\theta_t^{(i)}; z_i') - \nabla_{\mathbf{v}} f(\theta_t; z_i) \end{pmatrix} \right\|_2^2\right] + 8L^2\Big(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2}\Big)^2 \mathbb{E}\left[\frac{G_S^4}{n}\sum_{i=1}^n \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2^2\right] \\
&\leq 8L^2\Big(\frac{1}{\tau_1^2} + \frac{1}{\tau_2^2}\Big)^2 \mathbb{E}\left[\frac{G_S^4}{n}\sum_{i=1}^n \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t^{(i)} \\ \mathbf{v}_t - \mathbf{v}_t^{(i)} \end{pmatrix} \right\|_2^2\right] + \frac{16\mathbb{E}[G_S^2]}{n^2}, \qquad \text{(B.37)}
\end{aligned}$$

where we use Lemma 12 and Assumption 3 for the last step. The proof is completed. □

We then apply Lemma 25 and Lemma 30 to directly derive Theorem 8. We omit the details for brevity.

The following lemma presents the optimization error bounds for Clipped-SGDA.

Proof of Lemma 9. Let $\mathbf{w} \in B_R$. We denote $\hat{g}_{\mathbf{w},t} = \mathrm{clip}(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}), \tau_1)$. By the update of Clipped-SGD and note that the projection step is nonexpansive, we know that

$$\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \big\|\mathbf{w}_t - \eta_t \hat{g}_{\mathbf{w},t} - \mathbf{w}\big\|_2^2 = \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\hat{g}_{\mathbf{w},t}\|_2^2 + 2\eta_t \langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t}\rangle.$$

By $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$, we know

$$\|\hat{g}_{\mathbf{w},t}\|_2^2 \leq 3\|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3\|\mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + 3\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2$$

and

$$\begin{aligned}
\langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t}\rangle = \langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\rangle + \langle \mathbf{w} - \mathbf{w}_t, \mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle \\
+ \langle \mathbf{w} - \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle.
\end{aligned}$$

We combine the above three inequalities together to derive

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 3\eta_t^2\|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3\eta_t^2\|\mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
+ 3\eta_t^2\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\rangle + 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle \\
+ 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle.
\end{aligned}$$

It was shown that [57]

$$\big\|\mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big\|_2 \leq G^2\tau_1^{-1}.$$

Then we derive that

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 3\eta_t^2\|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3\eta_t^2 G^4\tau_1^{-2} + 3\eta_t^2\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
&+ 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\rangle + 2G\tau_1^{-1}\eta_t\|\mathbf{w} - \mathbf{w}_t\|_2 + 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle \\
&\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 3\eta_t^2\|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3\eta_t^2 G^4\tau_1^{-2} + 3\eta_t^2\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
&+ 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\rangle + 2G^2\tau_1^{-1}\eta_t\|\mathbf{w} - \mathbf{w}_t\|_2 + 2\eta_t\big(F_S(\mathbf{w}, \mathbf{v}_t) - F_S(\mathbf{w}_t, \mathbf{v}_t)\big),
\end{aligned}$$

where we use the convexity of $\mathbf{w} \mapsto F_S(\mathbf{w}, \mathbf{v})$ for the last step above. Taking a summation of the above inequality from $t = 1$ to $t = T$ ($\mathbf{w}_1 = 0$), we derive

$$\begin{aligned}
&\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + 2\sum_{t=1}^T \eta_t\big(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)\big) \\
&\leq \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + 3\sum_{t=1}^T \eta_t^2\|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3G^4\tau_1^{-2}\sum_{t=1}^T \eta_t^2 + 3\sum_{t=1}^T \eta_t^2\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
&+ 2\sum_{t=1}^T \eta_t\langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\rangle + 2G^2\tau_1^{-1}\sum_{t=1}^T \eta_t\|\mathbf{w}_t - \mathbf{w}\|_2.
\end{aligned}$$

Setting $\eta_t = \eta$, we further get

$$\begin{aligned}
&\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + 2\eta\sum_{t=1}^T \big(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)\big) \\
&\leq \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + 3\eta^2\sum_{t=1}^T \|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3G^4\tau_1^{-2}\eta^2 T + 3\eta^2\sum_{t=1}^T \|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
&+ 2\eta\sum_{t=1}^T \langle \mathbf{w} - \mathbf{w}_t, \hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\rangle + 2G^2\tau_1^{-1}\eta\sum_{t=1}^T \|\mathbf{w}_t - \mathbf{w}\|_2.
\end{aligned} \tag{B.38}$$

It then follows from the concavity of $F_S(\mathbf{w}, \cdot)$ and Schwartz's inequality that

$$\begin{aligned}
&\|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2 + 2\eta\sum_{t=1}^T \big(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big) \\
&\leq \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + 3\eta^2\sum_{t=1}^T \|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3G^4\tau_1^{-2}\eta^2 T + 3\eta^2\sum_{t=1}^T \|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
&+ 2\eta\sum_{t=1}^T \langle \mathbf{w}_t, \mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \hat{g}_{\mathbf{w},t}\rangle + 2\eta R\Big\|\sum_{t=1}^T \big(\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big)\Big\|_2 + 2G^2\tau_1^{-1}\eta\sum_{t=1}^T \|\mathbf{w}_t - \mathbf{w}\|_2.
\end{aligned}$$

Define $\mathbf{w}_{S,1} = \arg\min_{\mathbf{w}\in B_R} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)$, then

$$\begin{aligned}
&\|\mathbf{w}_{T+1} - \mathbf{w}_{S,1}\|_2^2 + 2\eta\sum_{t=1}^T \big(F_S(\mathbf{w}_t, \mathbf{v}_t) - \inf_{\mathbf{w}\in B_R} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big) \\
&\leq \|\mathbf{w}_1 - \mathbf{w}_{S,1}\|_2^2 + 3\eta^2\sum_{t=1}^T \|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2 + 3G^4\tau_1^{-2}\eta^2 T + 3\eta^2\sum_{t=1}^T \|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\
&+ 2\eta\sum_{t=1}^T \langle \mathbf{w}_t, \mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \hat{g}_{\mathbf{w},t}\rangle + 2\eta R\Big\|\sum_{t=1}^T \big(\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big)\Big\|_2 + 2G^2\tau_1^{-1}\eta\sum_{t=1}^T \|\mathbf{w}_t - \mathbf{w}_{S,1}\|_2.
\end{aligned} \tag{B.39}$$

Note $\mathbb{E}\big[\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2\big] \leq G^2$, $\mathbb{E}[\|\hat{g}_{\mathbf{w},t} - \mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2] \leq G^2$ and

$$\mathbb{E}_t[\langle \mathbf{w}_t, \mathbb{E}_t[\hat{g}_{\mathbf{w},t}] - \hat{g}_{\mathbf{w},t}\rangle] = 0. \tag{B.40}$$

Then taking an expectation to both sides of Eq. (B.39), we have

$$\mathbb{E}[\|\mathbf{w}_{T+1}-\mathbf{w}_{S,1}\|_2^2]+2\eta\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]$$

$$\leq \mathbb{E}[\|\mathbf{w}_1-\mathbf{w}_{S,1}\|_2^2]+3G^2\eta^2T+3G^4\tau_1^{-2}\eta^2T+3G^2\eta^2T+2\eta R\mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\big(\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big)\Big\|_2\Big]$$

$$+2G^2\tau_1^{-1}\eta\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_{S,1}\|_2]. \tag{B.41}$$

According to Jensen's inequality and Eq. (B.40), we know

$$\Big(\mathbb{E}_A\Big[\Big\|\sum_{t=1}^{T}\big(\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big)\Big\|_2\Big]\Big)^2\leq\mathbb{E}_A\Big[\Big\|\sum_{t=1}^{T}\big(\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big)\Big\|_2^2\Big]=\sum_{t=1}^{T}\mathbb{E}_A\big[\big\|\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big\|_2^2\big]$$

and

$$\mathbb{E}\Big[\Big\|\sum_{t=1}^{T}\big(\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big)\Big\|_2\Big]\leq\mathbb{E}\big[\big(\sum_{t=1}^{T}\mathbb{E}_A\big[\big\|\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big\|_2^2\big]\big)^{\frac{1}{2}}\big]$$
$$\leq\big(\sum_{t=1}^{T}\mathbb{E}[\big\|\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\big\|_2^2]\big)^{\frac{1}{2}}\leq G\sqrt{T}.$$

Plugging the above inequality into Eq. (B.41) and using $\mathbb{E}[\|\hat{g}_{\mathbf{w},t}-\mathbb{E}_t[\hat{g}_{\mathbf{w},t}]\|_2^2]\leq G^2$, we have

$$\mathbb{E}[\|\mathbf{w}_{T+1}-\mathbf{w}_{S,1}\|_2^2]+2\eta\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq\mathbb{E}[\|\mathbf{w}_1-\mathbf{w}_{S,1}\|_2^2]$$
$$+3\big(2+G^2\tau_1^{-2}\big)G^2\eta^2T+2RG\eta\sqrt{T}+2G^2\tau_1^{-1}\eta\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_{S,1}\|_2]. \tag{B.42}$$

Note if $\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\geq 0$, Eq. (B.41) implies the following inequality for any $t\in[T]$:

$$\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_{S,1}\|_2^2]\leq\mathbb{E}[\|\mathbf{w}_1-\mathbf{w}_{S,1}\|_2^2]+3\big(2+\tau_1^{-2}\big)G^2\eta^2T+2RG\eta\sqrt{T}$$
$$+2G^2\tau_1^{-1}\eta\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_{S,1}\|_2].$$

If we denote $\bar{\triangle}=\big(\max_{t\in[T]}\mathbb{E}\big[\|\mathbf{w}_t-\mathbf{w}_{S,1}\|_2^2\big]\big)^{\frac{1}{2}}$, then it follows that

$$\bar{\triangle}^2\leq\mathbb{E}\big[\|\mathbf{w}_1-\mathbf{w}_{S,1}\|_2^2\big]+3\big(2+G^2\tau_1^{-2}\big)G^2\eta^2T+2RG\eta\sqrt{T}+2G^2\tau_1^{-1}\eta T\bar{\triangle}.$$

Solving this quadratic inequality of $\bar{\triangle}$ implies that

$$\bar{\triangle}\leq\big(\mathbb{E}\big[\|\mathbf{w}_1-\mathbf{w}_{S,1}\|_2^2\big]\big)^{\frac{1}{2}}+\sqrt{3}\big(2+G^2\tau_1^{-2}\big)^{\frac{1}{2}}G\eta\sqrt{T}+\sqrt{2}R^{\frac{1}{2}}G^{\frac{1}{2}}\eta^{\frac{1}{2}}T^{\frac{1}{4}}+2G^2\tau_1^{-1}\eta T.$$

We plug this inequality back into Eq. (B.42) and derive (noting $\tau_1\gtrsim G$)

$$\eta\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\lesssim\mathbb{E}[\|\mathbf{w}_{S,1}\|_2^2]+G^2\eta^2T+RG\eta\sqrt{T}$$
$$+G^2\tau_1^{-1}\eta T\Big(\big(\mathbb{E}\big[\|\mathbf{w}_{S,1}\|_2^2\big]\big)^{\frac{1}{2}}+G\eta\sqrt{T}+R^{\frac{1}{2}}G^{\frac{1}{2}}\eta^{\frac{1}{2}}T^{\frac{1}{4}}+G^2\tau_1^{-1}\eta T\Big).$$

It then follows that

$$\eta\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\lesssim\mathbb{E}[\|\mathbf{w}_{S,1}\|_2^2]+G^2\eta^2T+RG\eta\sqrt{T}$$
$$+G^2\tau^{-1}\eta T\big(\mathbb{E}\big[\|\mathbf{w}_{S,1}\|_2^2\big]\big)^{\frac{1}{2}}+G^3\tau_1^{-1}\eta^2T^{\frac{3}{2}}+G^{\frac{5}{2}}\tau_1^{-1}R^{\frac{1}{2}}\eta^{\frac{3}{2}}T^{\frac{5}{4}}+G^4\tau_1^{-2}\eta^2T^2.$$

It is clear that $2G^2\tau_1^{-1}(\mathbb{E}\big[\|\mathbf{w}_{S,1}\|_2^2\big])^{\frac{1}{2}}\eta T \leq \mathbb{E}\big[\|\mathbf{w}_{S,1}\|_2^2\big] + G^4\tau_1^{-2}\eta^2T^2$

$$2G^3\tau_1^{-1}\eta^2T^{\frac{3}{2}} \leq G^2\eta^2T + G^4\tau_1^{-2}\eta^2T^2$$

and

$$2G^{\frac{5}{2}}\tau_1^{-1}R^{\frac{1}{2}}\eta^{\frac{3}{2}}T^{\frac{5}{4}} \leq RG\eta\sqrt{T} + G^4\tau_1^{-2}\eta^2T^2.$$

Then,

$$\eta\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t) - \inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big] \lesssim \mathbb{E}[\|\mathbf{w}_{S,1}\|_2^2] + G^2\eta^2T + RG\eta\sqrt{T} + G^4\tau_1^{-2}\eta^2T^2,$$

which yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)\big] - \mathbb{E}\big[\inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big] \lesssim \frac{R^2}{\eta T} + G^2\eta + \frac{GR}{\sqrt{T}} + G^4\tau_1^{-2}\eta T.$$

If $\sum_{t=1}^{T}\mathbb{E}\big[F_S(\mathbf{w}_t,\mathbf{v}_t) - \inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big] < 0$, the above inequality also holds. In a similar way, we can show that

$$\mathbb{E}[\sup_{\mathbf{v}\in B_R}F_S(\bar{\mathbf{w}}_T,\mathbf{v})] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F_S(\mathbf{w}_t,\mathbf{v}_t)] \lesssim \frac{R^2}{\eta T} + G^2\eta + \frac{GR}{\sqrt{T}} + G^4\tau_2^{-2}\eta T.$$

Combining these two inequalities, we have

$$\mathbb{E}\big[\sup_{\mathbf{v}\in B_R}F_S(\bar{\mathbf{w}}_T,\mathbf{v}) - \inf_{\mathbf{w}\in B_R}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big] \lesssim \frac{R^2}{\eta T} + G^2\eta + \frac{GR}{\sqrt{T}} + G^4(\tau_1^{-2}+\tau_2^{-2})\eta T.$$

The proof is completed by noting $\frac{GR}{\sqrt{T}} \leq \frac{R^2}{\eta T} + G^2\eta$. □

Combining Theorem 8 and Lemma 9, we derive the following theorem.

Proof of Theorem 10. According to Part (b) of Theorem 1 and Theorem 8, we know

$$\triangle^{w,R}(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S)) - \triangle_S^{w,R}(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S)) \lesssim \frac{LG^2\eta^2T^2}{n^2} + \frac{LG^2\eta^2T}{n} + \frac{G^2\eta T}{n} + \frac{G^2\eta\sqrt{T}}{\sqrt{n}}.$$

According to Lemma 9, we know

$$\Delta_S^{w,R}(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) \lesssim \frac{R^2}{\eta T} + G^2\eta + G^4(\tau_1^{-2}+\tau_2^{-2})\eta T.$$

The bound then follows directly from Eq. (B.30). □

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clarify our contributions and the scope of the paper in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: In Section 6, we list several limitations of the paper, e.g., our analysis only considers the convex-concave case.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We present all the necessary assumptions and include the proof of all theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: As a theoretical paper, our main contributions are theoretical analyses and proofs. The numerical experiments are designed solely to provide basic validation of our theoretical findings. We have fully disclosed all experimental details in Section A of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The details and code of experiments can be found in Section A.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details of experiments are provided in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The details of experiments are provided in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Given the simplicity of our numerical experiments (which are minimal and for verification only), we conducted our experiments with only one NVIDIA GeForce RTX 3090 24GB GPU.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper confirm with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theoretical paper, and there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This is a theoretical paper, which poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: This is a theoretical paper, and we do not use any existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not release any new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This is a theoretical paper, which does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLM in preparing the submission.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.