Stability and Generalization of Stochastic Gradient Methods for Minimax Problems

Yunwen Lei [* 1] **Zhenhuan Yang** [* 2] **Tianbao Yang** [3] **Yiming Ying** [2]

Abstract

Many machine learning problems can be formulated as minimax problems such as Generative Adversarial Networks (GANs), AUC maximization and robust estimation, to mention but a few. A substantial amount of studies are devoted to studying the convergence behavior of their stochastic gradient-type algorithms. In contrast, there is relatively little work on their generalization, i.e., how the learning models built from training examples would behave on test examples. In this paper, we provide a comprehensive generalization analysis of stochastic gradient methods for minimax problems under both convex-concave and nonconvex-nonconcave cases through the lens of algorithmic stability. We establish a quantitative connection between stability and several generalization measures both in expectation and with high probability. For the convex-concave setting, our stability analysis shows that stochastic gradient descent ascent attains optimal generalization bounds for both smooth and nonsmooth minimax problems. We also establish generalization bounds for both weakly-convex-weakly-concave and gradient-dominated problems.

1. Introduction

In machine learning we often encounter minimax optimization problems, where the decision variables are partitioned into two groups: one for minimization and one for maximization. This framework covers many important problems as specific instantiations, including adversarial learning (Goodfellow et al., 2014), robust optimization (Chen et al., 2017; Namkoong & Duchi, 2017), reinforcement learning (Dai et al., 2018; Du et al., 2017) and AUC maximization (Gao et al., 2013; Lei & Ying, 2021b; Liu et al., 2018; Ying et al., 2016; Zhao et al., 2011). To solve these problems, researchers have proposed various efficient optimization algorithms, for which a representative algorithm is the stochastic gradient descent ascent (SGDA) due to its simplicity and widespread use in real-world applications.

There is a large amount of work on the convergence analysis of minimax optimization algorithms in different settings such as convex-concave (Nemirovski et al., 2009), strongly-convex-strongly-concave (SC-SC) (Balamurugan & Bach, 2016), nonconvex-concave (Rafique et al., 2018) and nonconvex-nonconcave (Liu et al., 2020; Yang et al., 2020) cases. However, there is relatively little work on studying the generalization, i.e., how the model trained based on the training examples would generalize to test examples. Indeed, a model with good performance in training examples may not generalize well if the models are too complex. It is imperative to study the generalization error of the trained models to foresee their prediction behavior which often entails the investigation of the tradeoff between optimization and estimation for an implicit regularization.

To our best knowledge, there is only two recent work on the generalization analysis for minimax optimization algorithms (Farnia & Ozdaglar, 2020; Zhang et al., 2020). The argument stability for the specific empirical saddle point (ESP) was studied (Zhang et al., 2020), which implies weak generalization and strong generalization bounds. However, the discussion there ignored optimization errors and nonconvex-nonconcave cases, which can be restrictive in practice. For SC-SC, convex-concave, nonconvex-nonconcave objective functions, the uniform stability of several gradient-based minimax learners was developed in a smooth setting (Farnia & Ozdaglar, 2020), including gradient descent ascent (GDA), proximal point method (PPM) and GDmax. While they developed optimal generalization bounds for PPM, their discussions did not yield vanishing risk bounds for GDA in the general convex-concave case since their generalization bounds grow exponentially in terms of the iteration number. Furthermore, the above mentioned papers only study generalization bounds in expectation, and there is a lack of high-probability analysis.

[*]Equal contribution [1]School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK [2]Department of Mathematics and Statistics, State University of New York at Albany, USA [3]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA. Correspondence to: Yiming Ying <yying@albany.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 108, 2021. Copyright 2021 by the author(s).

In this paper, we leverage the lens of algorithmic stability to study the generalization behavior of minimax learners for both convex-concave and nonconvex-nonconcave problems. Our discussion shows how the optimization and generalization should be balanced for good prediction performance. We list our main results in Table 1. In particular, our contributions can be summarized as follows.

1. We establish a quantitative connection between stability and generalization for minimax learners in different forms including weak/strong primal-dual generalization, primal generalization and generalization with high probability. For the technical contributions, we introduce novel decompositions to handle the correlation between the primal model and dual model for connecting stability and generalization.

2. We establish stability bounds of SGDA for convex-concave problems, from which we derive its optimal population risk bounds under an appropriate early-stopping strategy. We consider several measures of generalization and show that the optimal population risk bounds can be derived even in the nonsmooth case. To the best of our knowledge, our results are the first-ever known population risk bounds for minimax problems in the nonsmooth setting and the high-probability format.

3. We further extend our analysis to the nonconvex-nonconcave setting and give the first generalization bounds for nonsmooth objective functions. Our analysis relaxes the range of step size for a controllable stability and implies meaningful primal population risk bounds under some regularity assumptions of objective functions, e.g., a decay of weak-convexity-weak-concavity parameter along the optimization process or a two-sided PL condition.

The paper is organized as follows. The related work is discussed in Section 1.1 and the minimax problem formulation is given in Section 2. The connection between stability and generalization is studied in Section 3. We develop population risk bounds in the convex-concave case in Section 4 and extend our discussions to the nonconvex-nonconcave case in Section 5. We conclude the paper in Section 7.

1.1. Related Work

We first review related work of stochastic optimization for minimax problems. Convergence rates of order $O(1/\sqrt{T})$ were established for SGDA with T iterations in the convex-concave case (Nedić & Ozdaglar, 2009; Nemirovski et al., 2009), which can be further improved for SC-SC problems (Balamurugan & Bach, 2016; Hsieh et al., 2019). These discussions were extended to nonconvex-strongly-concave (Lin et al., 2020; Luo et al., 2020; Rafique et al., 2018; Yan et al., 2020), nonconvex-concave (Lin et al., 2020; Thekumparampil et al., 2019) and nonconvex-nonconcave (Liu et al., 2020; Loizou et al., 2020; Yang et al., 2020) minimax optimization problems. All the above mentioned work consider the convergence rate of optimization errors, while the generalization analysis was much less studied (Farnia & Ozdaglar, 2020; Zhang et al., 2020).

We now survey related work on stability and generalization. The framework of stability analysis was established in a seminal paper (Bousquet & Elisseeff, 2002), where the celebrated concept of uniform stability was introduced. This stability was extended to study randomized algorithms (Elisseeff et al., 2005). It was shown that stability is closely related to the fundamental problem of learnability (Rakhlin et al., 2005; Shalev-Shwartz et al., 2010). Hardt et al. (2016) pioneered the generalization analysis of SGD via stability, which inspired several upcoming work to understand stochastic optimization algorithms based on different algorithmic stability measures, e.g., uniform stability (Chen et al., 2018; Lin et al., 2016; Madden et al., 2020; Mou et al., 2018; Richards et al., 2020), argument stability (Bassily et al., 2020; Lei & Ying, 2020; Liu et al., 2017), on-average stability (Kuzborskij & Lampert, 2018; Lei & Ying, 2021a), hypothesis stability (Charles & Papailiopoulos, 2018; Foster et al., 2019; London, 2017), Bayes stability (Li et al., 2020) and locally elastic stability (Deng et al., 2020).

2. Problem Formulation

Let $\mathcal{W}$ and $\mathcal{V}$ be two parameter spaces in $\mathbb{R}^d$. Let $\mathbb{P}$ be a probability measure defined on a sample space $\mathcal{Z}$ and $f : \mathcal{W} \times \mathcal{V} \times \mathcal{Z} \mapsto \mathbb{R}$. We consider the following minimax optimization problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w}, \mathbf{v}) := \mathbb{E}_{z\sim\mathbb{P}}[f(\mathbf{w}, \mathbf{v}; z)]. \tag{2.1}$$

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 (2.1). Since the model $A(S)$ is trained based on the training dataset S, its empirical behavior as measured by F_S may not generalize well to a test example (Bousquet & Elisseeff, 2002). We are interested in studying the test error (population risk) of $A(S)$. Unlike the standard statistical learning theory (SLT) setting where there is only a minimization of $\mathbf{w}$, we have different measures of population risk due to the minimax structure (Zhang et al., 2020). We collect the notations of these performance measures in Table A.1 in Appendix A. Let $\mathbb{E}[\cdot]$ denote the expectation w.r.t. the randomness of both the algorithm A and the dataset S.

Algorithm	Reference	Assumption	Measure	Rate
ESP	Zhang et al. (2020)	ρ-SC-SC, Lip	Weak PD Risk	$O(1/(n\rho))$
		ρ-SC-SC, Lip, S	Strong PD Risk	$O(1/(n\rho^2))$
R-ESP		C-C, Lip	Weak PD Risk	$O(1/\sqrt{n})$
SGDA, SGDmax	Farnia & Ozdaglar (2020)	ρ-SC-SC, Lip, S	Weak PD Generalization[1]	$O(\log(n)/(n\rho))$
PPM		C-C, Lip, S	Weak PD Risk	$O(1/\sqrt{n})$
SGDA	This work	C-C, Lip (S)	Weak PD Risk	$O(1/\sqrt{n})$
		C-ρ-SC, Lip, S	(H.P.) Primal Risk	$O(1/(\sqrt{n}\rho))$
		C-C, Lip	H.P. Plain Risk	$O(\log(n)/\sqrt{n})$
		ρ-SC-SC, Lip	Weak PD Risk	$O(\sqrt{\log n}/(n\rho))$
SGDA	Farnia & Ozdaglar (2020)	Lip, S	Weak PD Generalization	$O\big(T^{\frac{Lc}{Lc+1}}/n\big)$
SGDA	This work	ρ-WC-WC, Lip	Weak PD Generalization	$O\big(T^{\frac{2c\rho}{2c\rho+3}}/n^{\frac{2c\rho+1}{2c\rho+3}}\big)$
		D, Lip, S	Weak PD Generalization	$O\big(1/\sqrt{n}+\sqrt{T}/n\big)$
AGDA		ρ-SC, PL, Lip, S	Primal Risk	$O\big(n^{-\frac{cL+1}{2cL+1}}\big)$

Table 1. Summary of Results. Bounds are stated in expectation or with high probability (H.P.). For risk bounds, the optimal T (number of iterations) is chosen to trade-off generalization and optimization. Here, C-C means convex-concave, C-ρ-SC means convex-ρ-strongly-concave, ρ-SC means nonconvex-ρ-strongly-concave, Lip means Lipschitz continuity, S means the smoothness, D means a decay of weak-convexity-weak-concavity parameter along the optimization process as Eq. (5.1) and PL means the two-sided condition as Assumption 3 . AGDA means Alternating Gradient Descent Ascent and (R)-ESP means the (regularized)-empirical risk saddle point. c is a parameter in the step size and L is given in Assumption 2.

Definition 1 (Weak Primal-Dual Risk)**.** The weak Primal-Dual (PD) population risk of a (randomized) model $(\mathbf{w},\mathbf{v})$ is defined as (Zhang et al., 2020)

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

The weak PD empirical risk of $(\mathbf{w},\mathbf{v})$ is defined as

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

We refer to $\triangle^w(\mathbf{w},\mathbf{v})-\triangle^w_S(\mathbf{w},\mathbf{v})$ as the weak PD generalization error of the model $(\mathbf{w},\mathbf{v})$.

Definition 2 (Strong Primal-Dual Risk)**.** The strong PD population risk of a model $(\mathbf{w},\mathbf{v})$ is defined as

$$\triangle^s(\mathbf{w},\mathbf{v})=\sup_{\mathbf{v}'\in\mathcal{V}}F(\mathbf{w},\mathbf{v}')-\inf_{\mathbf{w}'\in\mathcal{W}}F(\mathbf{w}',\mathbf{v}).$$

The strong PD empirical risk of $(\mathbf{w},\mathbf{v})$ is defined as

$$\triangle^s_S(\mathbf{w},\mathbf{v})=\sup_{\mathbf{v}'\in\mathcal{V}}F_S(\mathbf{w},\mathbf{v}')-\inf_{\mathbf{w}'\in\mathcal{W}}F_S(\mathbf{w}',\mathbf{v}).$$

We refer to $\triangle^s(\mathbf{w},\mathbf{v})-\triangle^s_S(\mathbf{w},\mathbf{v})$ as the strong PD generalization error of the model $(\mathbf{w},\mathbf{v})$.

Definition 3 (Primal Risk)**.** The primal population risk of a model $\mathbf{w}$ is defined as $R(\mathbf{w})=\sup_{\mathbf{v}\in\mathcal{V}}F(\mathbf{w},\mathbf{v})$. The primal empirical risk of $\mathbf{w}$ is defined as $R_S(\mathbf{w})=\sup_{\mathbf{v}\in\mathcal{V}}F_S(\mathbf{w},\mathbf{v})$. We refer to $R(\mathbf{w})-R_S(\mathbf{w})$ as the primal generalization error of the model $\mathbf{w}$, and $R(\mathbf{w})-\inf_{\mathbf{w}'}R(\mathbf{w}')$ as the excess primal population risk.

According to the above definitions, we know $\triangle^w(\mathbf{w},\mathbf{v})\leq\mathbb{E}\big[\triangle^s(\mathbf{w},\mathbf{v})\big]$ and $R(\mathbf{w})-R_S(\mathbf{w})$ is closely related to $\triangle^s(\mathbf{w},\mathbf{v})-\triangle^s_S(\mathbf{w},\mathbf{v})$. The key difference between the weak PD risk and the strong PD risk is that the expectation is inside of the supremum/infimum for weak PD risk, while outside of the supremum/infimum for strong PD risk. In this way, one does not need to consider the coupling between primal and dual models for studying the weak PD risk, and has to consider this coupling for studying the strong PD risk. Furthermore, we refer to $F(\mathbf{w},\mathbf{v})-F_S(\mathbf{w},\mathbf{v})$ as the plain generalization error as it is standard in SLT. A standard approach to handle a population risk is to decompose it into a generalization error (estimation error) and an empirical risk (optimization error) (Bousquet & Bottou, 2008). For example, the weak PD population risk can be decomposed as

$$\triangle^w(\mathbf{w},\mathbf{v})=\big(\triangle^w(\mathbf{w},\mathbf{v})-\triangle^w_S(\mathbf{w},\mathbf{v})\big)+\triangle^w_S(\mathbf{w},\mathbf{v}).\quad(2.2)$$

The generalization error comes from the approximation of $\mathbb{P}$ with S, while the empirical risk comes since the algorithm may not find the saddle point of F_S. Our basic idea is to use algorithmic stability to study the generalization error and use optimization theory to study the empirical risk.

We now introduce necessary definitions and assumptions. Denote $\|\cdot\|_2$ as the Euclidean norm and $\langle\cdot,\cdot\rangle$ as the inner product. A function $g:\mathcal{W}\mapsto\mathbb{R}$ is said to be ρ-strongly-convex ($\rho\geq0$) iff for all $\mathbf{w},\mathbf{w}'\in\mathcal{W}$ there holds

$$g(\mathbf{w})\geq g(\mathbf{w}')+\langle\mathbf{w}-\mathbf{w}',\nabla g(\mathbf{w}')\rangle+\frac{\rho}{2}\|\mathbf{w}-\mathbf{w}'\|_2^2,$$

where ∇ is the gradient operator. We say g is convex if g is 0-strongly-convex. We say g is ρ-strongly concave if $-g$ is ρ-strongly convex and concave if $-g$ is convex.

Definition 4. Let $\rho\geq0$ and $g:\mathcal{W}\times\mathcal{V}\mapsto\mathbb{R}$. We say

(a) g is ρ-strongly-convex-strongly-concave (ρ-SC-SC) if for any $\mathbf{v} \in \mathcal{V}$, the function $\mathbf{w} \mapsto g(\mathbf{w}, \mathbf{v})$ is ρ-strongly-convex and for any $\mathbf{w} \in \mathcal{W}$, the function $\mathbf{v} \mapsto g(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave.

(b) g is convex-concave if g is 0-SC-SC.

(c) g is ρ-weakly-convex-weakly-concave (ρ-WC-WC) if $g + \frac{\rho}{2}\big(\|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2\big)$ is convex-concave.

The following two assumptions are standard (Farnia & Ozdaglar, 2020; Zhang et al., 2020). Assumption 1 amounts to saying f is Lipschitz continuous with respect to (w.r.t.) both $\mathbf{w}$ and $\mathbf{v}$. Let $\nabla_{\mathbf{w}} f$ denote the gradient w.r.t. $\mathbf{w}$.

Assumption 1. Let $G > 0$. Assume for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$, there holds $\big\|\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; z)\big\|_2 \leq G$ and $\big\|\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; z)\big\|_2 \leq G$.

Assumption 2. Let $L > 0$. For any z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is said to be L-smooth, if the following inequality holds for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$ and $z \in \mathcal{Z}$

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

2.1. Motivating Examples

The minimax formulation (2.1) has broad applications in machine learning. Here we give some examples.

AUC Maximization. Area Under ROC Curve (AUC) is a popular measure for binary classification. Let $h(\mathbf{w}; x)$ denote a scoring function parameterized by $\mathbf{w}$ at x. It was shown that AUC maximization for learning h under the square loss reduces to the problem (Ying et al., 2016)

$$\min_{(\mathbf{w},a,b) \in \mathbb{R}^{d+2}} \max_{\alpha \in \mathbb{R}} \mathbb{E}[f(\mathbf{w}, a, b, \alpha; z)], \tag{2.3}$$

where $p = \mathbb{P}[y = 1]$ and $f(\mathbf{w}, a, b, \alpha; z) = (h(\mathbf{w}; x) - a)^2 \mathbb{I}_{[y=1]}/p + (h(\mathbf{w}; x) - b)^2 \mathbb{I}_{[y=-1]}/(1 - p) + 2(1 + \alpha) h(\mathbf{w}; x)(\mathbb{I}_{[y=-1]}/(1 - p) - \mathbb{I}_{[y=1]}/p) - \alpha^2$ ($\mathbb{I}_{[\cdot]}$ is the indicator function). It is clear that $\alpha \mapsto f(\mathbf{w}, a, b, \alpha; z)$ is a (strongly) concave function. Depending on h, the function f can be convex, nonconvex, smooth or nonsmooth.

Generative Adversarial Networks. GAN (Goodfellow et al., 2014) refers to a popular class of generative models that consider generative modeling as a game between a generator network $G_{\mathbf{v}}$ and a discriminator network $D_{\mathbf{w}}$. The generator network produces synthetic data from random noise $\xi \sim \mathbb{P}_\xi$, while the discriminator network discriminates between the true data and the synthetic data. In particular, a popular variant of GAN named as WGAN (Arjovsky et al., 2017) can be written as a minimax problem

$$\min_{\mathbf{w}} \max_{\mathbf{v}} \mathbb{E}[f(\mathbf{w}, \mathbf{v}; z, \xi)] = \mathbb{E}_z[D_{\mathbf{w}}(z)] - \mathbb{E}_\xi[D_{\mathbf{w}}(G_{\mathbf{v}}(\xi))].$$

While this problem is generally nonconvex-nonconcave, it is weakly-convex-weakly-concave under smoothness assumptions on D and G (Liu et al., 2020).

Robust Estimation with minimax estimator. Audibert & Catoni (2011) formulated robust estimation as a minimax problem as follows

$$\min_{\mathbf{w}} \max_{\mathbf{v}} \; F_S(\mathbf{w}, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^n \psi(\ell_1(\mathbf{w}; z_i) - \ell_2(\mathbf{v}; z_i)),$$

where $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a truncated loss, and ℓ_1, ℓ_2 are Lipschitz continuous and convex loss functions. A typical truncated loss is $\psi(x) = \log(1 + |x| + x^2/2)\text{sign}(x)$ to compute a mean estimator under heavy-tailed distribution of data (Brownlees et al., 2015; Xu et al., 2020), where $\text{sign}(x)$ is the sign of x. The composition function F_S can be nonconvex and nonsmooth since ψ is nonconvex and ℓ_1, ℓ_2 can be nonsmooth. Following Xu et al. (2020), it can be shown that $F_S(\mathbf{w}, \mathbf{v})$ is weakly-convex-weakly-concave.

3. Connecting Stability and Generalization

A fundamental concept in our analysis is the algorithmic stability, which measures the sensitivity of an algorithm w.r.t. the perturbation of training datasets (Bousquet & Elisseeff, 2002). We say $S, S' \subset \mathcal{Z}$ are neighboring datasets if they differ by at most a single example. We introduce three stability measures to the minimax learning setting. The weak-stability and uniform-stability quantify the sensitivity measured by function values, while the argument-stability quantifies the sensitivity measured by arguments. We collect these notations of stabilities in Table A.2 in Appendix A.

Definition 5 (Algorithmic Stability)**.** Let A be a randomized algorithm, $\epsilon > 0$ and $\delta \in (0, 1)$. Then we say

(a) A is ϵ-weakly-stable if for all neighboring S and S' there holds

$$\sup_z \Big(\sup_{\mathbf{v}' \in \mathcal{V}} \mathbb{E}_A\big[f(A_{\mathbf{w}}(S), \mathbf{v}'; z) - f(A_{\mathbf{w}}(S'), \mathbf{v}'; z)\big] + \sup_{\mathbf{w}' \in \mathcal{W}} \mathbb{E}_A\big[f(\mathbf{w}', A_{\mathbf{v}}(S); z) - f(\mathbf{w}', A_{\mathbf{v}}(S'); z)\big]\Big) \leq \epsilon.$$

(b) A is ϵ-argument-stable in expectation if for all neighboring S and S' there holds

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S')\end{pmatrix}\right\|_2\right] \leq \epsilon.$$

---
[1]Primal generalization bounds were presented in Farnia & Ozdaglar (2020). However, the stability analysis there actually only implies bounds on weak PD risk.

A is ϵ-argument-stable with probability at least $1-\delta$ if with probability at least $1-\delta$

$$\left\| \begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S') \end{pmatrix} \right\|_2 \leq \epsilon.$$

(c) A is ϵ-uniformly-stable with probability at least $1-\delta$ if for all neighboring S and S', the following inequality holds with probability at least $1-\delta$

$$\sup_z \big[f(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S); z) - f(A_{\mathbf{w}}(S'), A_{\mathbf{v}}(S'); z) \big] \leq \epsilon.$$

Under Assumption 1, the argument stability implies weak stability and uniform stability. As we will see, argument stability plays an important role in deriving primal population risk bounds.

As our first main result, we establish a quantitative connection between algorithmic stability and generalization in the following theorem to be proved in Appendix B. Part (a) establishes the connection between weak-stability and weak PD generalization error. Part (b) and Part (c) establish the connection between argument stability and strong/primal generalization error under a further assumption on the strong convexity/concavity. Part (d) and Part (e) establish high-probability bounds based on the uniform stability, which are much more challenging to derive than bounds in expectation and are important to understand the variation of an algorithm in several independent runs (Bousquet et al., 2020; Feldman & Vondrak, 2019). Regarding the technical contributions, we introduce novel decompositions in handling the correlation between $A_{\mathbf{w}}(S)$ and $\mathbf{v}^* = \arg\sup_{\mathbf{v}} F(A_{\mathbf{w}}(S), \mathbf{v})$, especially for high-probability analysis.

Theorem 1. *Let A be a randomized algorithm and $\epsilon > 0$.*

(a) If A is ϵ-weakly-stable, then the weak PD generalization error of $(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ satisfies

$$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle^w_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \epsilon.$$

(b) If A is ϵ-argument-stable in expectation, the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave and Assumptions 1, 2 hold, then the primal generalization error satisfies

$$\mathbb{E}_{S,A}\Big[R(A_{\mathbf{w}}(S)) - R_S(A_{\mathbf{w}}(S))\Big] \leq \big(1 + L/\rho\big) G\epsilon.$$

(c) If A is ϵ-argument-stable in expectation, $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is ρ-SC-SC and Assumptions 1, 2 hold, then the strong PD generalization error satisfies

$$\mathbb{E}_{S,A}\Big[\triangle^s(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle^s_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))\Big] \leq \big(1 + L/\rho\big) G\sqrt{2}\epsilon.$$

(d) Assume $|f(\mathbf{w}, \mathbf{v}; z)| \leq R$ for some $R > 0$ and $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}, z \in \mathcal{Z}$. Assume for all $\mathbf{w}$, the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave and Assumptions 1, 2 hold. Let $\delta \in (0, 1)$. If A is ϵ-uniformly stable almost surely (a.s.), then with probability at least $1-\delta$

$$R(A_{\mathbf{w}}(S)) - R_S(A_{\mathbf{w}}(S)) = O\Big(GL\rho^{-1}\epsilon \log n \log(1/\delta) + Rn^{-\frac{1}{2}}\sqrt{\log(1/\delta)}\Big).$$

(e) Assume $|f(\mathbf{w}, \mathbf{v}; z)| \leq R$ for some $R > 0$ and $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}, z \in \mathcal{Z}$. Let $\delta \in (0, 1)$. If A is ϵ-uniformly-stable a.s., then with probability $1-\delta$ there holds

$$\big|F(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - F_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))\big| = O\Big(\epsilon \log n \log(1/\delta) + Rn^{-\frac{1}{2}}\sqrt{\log(1/\delta)}\Big).$$

Remark 1. We compare Theorem 1 with related work. Weak and strong PD generalization error bounds were established for (R)-ESP (Zhang et al., 2020). However, the discussion there does not consider the connection between stability and generalization. Primal generalization bounds were studied for stable algorithms (Farnia & Ozdaglar, 2020). However, the discussion there is not rigorous since they used an identity $nR_S(A_{\mathbf{w}}(S)) = \sum_{i=1}^n \max_{\mathbf{v}} f(A_{\mathbf{w}}(S), \mathbf{v}; z_i)$, which does not hold. To our best knowledge, Theorem 1 gives the first systematic connection between stability and generalization for minimax problems.

Remark 2. We provide some intuitive understanding of Theorem 1 here. Part (a) shows that weak-stability is sufficient for weak PD generalization. This is as expected since both the supremum over $\mathbf{w}'$ and $\mathbf{v}'$ are outside of the expectation operator in the definition of weak stability/generalization. We do not need to consider the correlation between $A_{\mathbf{w}}(S)$ and $\mathbf{v}'$. As a comparison, the primal generalization needs the much stronger argument-stability. The reason is that the supremum over $\mathbf{w}'$ is inside the expectation and $\mathbf{v}^{(i)} := \arg\sup_{\mathbf{v}'} F(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}')$ is different for different i ($\mathbf{v}^{(i)}$ correlates to $A_{\mathbf{w}}(S^{(i)})$). We need to estimate how $\mathbf{v}^{(i)}$ differs from each other due to the difference among $A_{\mathbf{w}}(S^{(i)})$. This explains why we need argument-stability and a strong-concavity in Parts (b), (d) for primal generalization. Similarly, the strong PD generalization assumes SC-SC functions.

4. SGDA: Convex-Concave Case

In this section, we are interested in SGDA for solving minimax optimization problems in the convex-concave case. Let $\mathbf{w}_1 = 0 \in \mathcal{W}$ and $\mathbf{v}_1 = 0 \in \mathcal{V}$ be the initial point. Let $\mathrm{Proj}_{\mathcal{W}}(\cdot)$ and $\mathrm{Proj}_{\mathcal{V}}(\cdot)$ denote the projections onto $\mathcal{W}$ and $\mathcal{V}$, respectively. Let $\{\eta_t\}_t$ be a sequence of positive stepsizes. At each iteration, we randomly draw i_t from the uniform distribution over $[n] := \{1, 2, \ldots, n\}$ and do the update

$$\begin{cases} \mathbf{w}_{t+1} = \text{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big), \\ \mathbf{v}_{t+1} = \text{Proj}_{\mathcal{V}}\big(\mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big). \end{cases} \tag{4.1}$$

4.1. Stability Bounds

In this section, we present the stability bounds for SGDA in the convex-concave case. We consider both the nonsmooth setting and smooth setting. Part (a) and Part (b) establish stability bounds in expectation, while Part (c) and Part (d) give stability bounds with high probability. Part (e) consider the SC-SC case. The proofs are given in Appendix C.

Theorem 2. *Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Let the algorithm A be SGDA (4.1) with t iterations. Let $\delta \in (0, 1)$.*

(a) Assume $\eta_t = \eta$. If Assumption 1 holds, then A is $4\eta G\big(\sqrt{t} + t/n\big)$-argument-stable in expectation.

(b) If Assumptions 1, 2 hold, then A is ϵ-argument-stable in expectation, where

$$\epsilon \leq \frac{\sqrt{8e(1+t/n)}G}{\sqrt{n}} \exp\Big(2^{-1}L^2 \sum_{j=1}^{t} \eta_j^2\Big)\Big(\sum_{k=1}^{t} \eta_k^2\Big)^{\frac{1}{2}}.$$

(c) Let $\eta_t = \eta$. If Assumption 1 holds, then A is ϵ-argument-stable with probability at least $1 - \delta$, where

$$\epsilon \leq \sqrt{8e}G\eta\Big(\sqrt{t} + t/n + \log(1/\delta) + \sqrt{2tn^{-1}\log(1/\delta)}\Big).$$

(d) Let $\eta_t = \eta$. If Assumptions 1, 2 hold, then A is ϵ-argument-stable with probability at least $1 - \delta$, where

$$\epsilon \leq \sqrt{8e}G\eta \exp\big(2^{-1}L^2 t\eta^2\big) \times \Big(1 + t/n + \log(1/\delta) + \sqrt{2tn^{-1}\log(1/\delta)}\Big).$$

(e) If $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is ρ-SC-SC, Assumption 1 holds and $\eta_t = 1/(\rho t)$, then A is ϵ-argument-stable in expectation, where $\epsilon \leq \frac{2\sqrt{2}G}{\rho}\big(\frac{\log(et)}{t} + \frac{1}{n(n-2)}\big)^{\frac{1}{2}}$.

Remark 3. If $t = O(n^2)$, then the stability bounds in a nonsmooth case (Part a) become $O(\eta\sqrt{t})$ and we can still get non-vacuous bounds by taking small step size $\eta = o(t^{-1/2})$. If we choose $\eta_j = 1/\sqrt{t}$ for $j \in [t]$, then the stability bound in Part (b) under a further smoothness assumption becomes $O(\sqrt{t}/n + n^{-\frac{1}{2}})$, which matches the existing result for SGD in a convex setting (Hardt et al., 2016). The high-probability bounds in Part (c) and Part (d) enjoy the same behavior.

Remark 4. The stability bounds of SGDA and GDA were discussed in Farnia & Ozdaglar (2020) for SC-SC, Lipschitz continuous and smooth problems. We remove the smoothness assumption in Part (e) in the SC-SC case. The stability of GDA was also studied there for convex-concave f, which, however, implies non-vanishing generalization bounds growing exponentially with the iteration count (Farnia & Ozdaglar, 2020). We extend their discussions to SGDA in this convex-concave case, and, as we will show in Theorem 3, our stability bounds imply optimal bounds on PD population risks. Furthermore, the existing discussions (Farnia & Ozdaglar, 2020) require the function f to be smooth, while we show that meaningful stability bounds can be achieved in a nonsmooth setting (Parts (a), (c), (e)).

Remark 5. We consider stability bounds under various assumptions on loss functions. We now sketch the technical difference in our analysis. Let $\delta_t := \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + \|\mathbf{v}_t - \mathbf{v}'_t\|_2^2$, where $(\mathbf{w}_t, \mathbf{v}_t), (\mathbf{w}'_t, \mathbf{v}'_t)$ are SGDA iterates for S and S' differing only by the last element. For convex-concave and nonsmooth problems, we show $\delta_{t+1} = \delta_t + O(\eta_t^2)$ if $i_t \neq n$. For convex-concave and smooth problems, we show $\delta_{t+1} = (1 + O(\eta_t^2))\delta_t$ if $i_t \neq n$. For ρ-SC-SC and nonsmooth problems, we show $\delta_{t+1} = (1 - 2\rho\eta_t)\delta_t + O(\eta_t^2)$ if $i_t \neq n$. For the above cases, we first control δ_{t+1} and then take expectation w.r.t. i_t. A key point to tackle *nonsmooth* problems is to consider the evolution δ_t instead of $\|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \|\mathbf{v}_t - \mathbf{v}'_t\|_2$, which yields nontrivial bounds by making $\sum_t \eta_t^2 = o(1)$ with small η_t.

4.2. Population Risks

We now use stability bounds in Theorem 2 to develop error bounds of SGDA which outputs an average of iterates

$$\bar{\mathbf{w}}_T = \frac{\sum_{t=1}^T \eta_t \mathbf{w}_t}{\sum_{t=1}^T \eta_t} \text{ and } \bar{\mathbf{v}}_T = \frac{\sum_{t=1}^T \eta_t \mathbf{v}_t}{\sum_{t=1}^T \eta_t}. \tag{4.2}$$

The underlying reason to introduce the average operator is to simplify the optimization error analysis (Nemirovski et al., 2009). Indeed, our stability and generalization analysis applies to any individual iterates. As a comparison, the optimization error analysis for the last iterate is much more difficult than that for the averaged iterate. We use the notation $B \asymp \widetilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1\widetilde{B} < B \leq c_2\widetilde{B}$. The following theorem to be proved in Appendix E gives weak PD population risk bounds.

Theorem 3 (Weak PD risk)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be produced by (4.1). Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Let A be defined by $A_{\mathbf{w}}(S) = \bar{\mathbf{w}}_T$ and $A_{\mathbf{v}}(S) = \bar{\mathbf{v}}_T$ for $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ in (4.2). Assume $\sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_2 \leq B_W$ and $\sup_{\mathbf{v}\in\mathcal{V}} \|\mathbf{v}\|_2 \leq B_V$.*

(a) If $\eta_t = \eta$ and Assumption 1 holds, then

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \leq 4\sqrt{2}\eta G^2\Big(\sqrt{T} + \frac{T}{n}\Big) + \eta G^2 + \frac{B_W^2 + B_V^2}{2\eta T} + \frac{G(B_W + B_V)}{\sqrt{T}}. \tag{4.3}$$

If we choose $T \asymp n^2$ and $\eta \asymp T^{-3/4}$, then we get $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(n^{-1/2})$.

(b) *If $\eta_t = \eta$ and Assumptions 1, 2 hold, then*

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le \frac{4\sqrt{e(T+T^2/n)}G^2\eta \exp\big(LT\eta^2/2\big)}{\sqrt{n}} + \eta G^2 + \frac{B_W^2 + B_V^2}{2\eta T} + \frac{G(B_W + B_V)}{\sqrt{T}}. \quad (4.4)$$

We can choose $T \asymp n$ and $\eta \asymp T^{-1/2}$ to derive $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(n^{-1/2})$.

(c) *If $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is ρ-SC-SC ($\rho > 0$), Assumption 1 holds, $\eta_t = 1/(\rho t)$ and $T \asymp n^2$, then*

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(\sqrt{\log n}/(n\rho)).$$

(d) *If $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is ρ-SC-SC ($\rho > 0$), Assumptions 1, 2 hold, $\eta_t = 1/(\rho(t + t_0))$ with $t_0 \ge L^2/\rho^2$ and $T \asymp n$, then $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(\log(n)/(n\rho))$.*

Remark 6. We first compare our bounds with the related work in a convex-concave setting. Weak PD population risk bounds were established for PPM under Assumptions 1, 2 (Farnia & Ozdaglar, 2020), which updates $(\mathbf{w}_{t+1}^{\text{PPM}}, \mathbf{v}_{t+1}^{\text{PPM}})$ as the saddle point of the following minimax problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}, \mathbf{v}) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t^{\text{PPM}}\|_2^2 + \frac{1}{2\eta_t}\|\mathbf{v} - \mathbf{v}_t^{\text{PPM}}\|_2^2.$$

In particular, they developed population risk bounds $O(1/\sqrt{n})$ by taking $T \asymp \sqrt{n}$ for PPM. However, the implementation of PPM requires to find the exact saddle point at each iteration, which is often computationally expensive. As a comparison, Part (b) shows the minimax optimal population risk bounds $O(1/\sqrt{n})$ for SGDA with $O(n)$ iterations. Weak PD population risk bounds $O(1/\sqrt{n})$ were established for R-ESP (Zhang et al., 2020) without a smoothness assumption, which, however, ignore the interplay between generalization and optimization. In this setting, we show SGDA achieves the same population risk bounds $O(1/\sqrt{n})$ by taking $\eta \asymp T^{-3/4}$ and $T \asymp n^2$ in Part (a). We now consider the SC-SC setting. Weak PD risk bounds $O(1/(n\rho))$ were established for ESP (Zhang et al., 2020). Since Farnia & Ozdaglar (2020) did not present an explicit risk bound, we use their stability analysis to give an explicit risk bound $O(\log(n)/(n\rho))$ in the smooth case (Part (d)). As a comparison, we establish the same population risk bounds for SGDA within a logarithmic factor by taking $\eta_t = 1/(\rho t)$ and $T \asymp n^2$ without the smoothness assumption (Part (c)).

We further develop bounds on primal population risks under a strong concavity assumption on $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$. Primal risk bounds measure the performance of primal variables, which are of real interest in some learning problems, e.g., AUC maximization and robust optimization. We consider both bounds in expectation and bounds with high-probability. Let $(\mathbf{w}^*, \mathbf{v}^*)$ be a saddle point of F, i.e., for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{v} \in \mathcal{V}$, there holds $F(\mathbf{w}^*, \mathbf{v}) \le F(\mathbf{w}, \mathbf{v}) \le F(\mathbf{w}, \mathbf{v}^*)$.

Theorem 4 (Excess primal risk)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be produced by (4.1) with $\eta_t = \eta$. Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave and the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave. Assume $\sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_2 \le B_W$ and $\sup_{\mathbf{v}\in\mathcal{V}} \|\mathbf{v}\|_2 \le B_V$. Let the algorithm A be defined by $A_{\mathbf{w}}(S) = \bar{\mathbf{w}}_T$ and $A_{\mathbf{v}}(S) = \bar{\mathbf{v}}_T$ for $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ in (4.2). If Assumptions 1, 2 hold, then*

$$\mathbb{E}\big[R(\bar{\mathbf{w}}_T)\big] - \inf_{\mathbf{w}} R(\mathbf{w}) \le \eta G^2 + \frac{B_W^2 + B_V^2}{2\eta T} + \frac{G(B_W + B_V)}{\sqrt{T}} + \frac{(1 + L/\rho)\sqrt{32e(T + T^2/n)}G^2\eta \exp\big(L^2T\eta^2/2\big)}{\sqrt{n}}.$$

In particular, if we choose $T \asymp n$ and $\eta \asymp T^{-1/2}$ then

$$\mathbb{E}\big[R(\bar{\mathbf{w}}_T)\big] - \inf_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w}) = O((L/\rho)n^{-1/2}). \quad (4.5)$$

Furthermore, for any $\delta \in (0, 1)$ we can choose $T \asymp n$ and $\eta \asymp T^{-1/2}$ to show with probability at least $1 - \delta$

$$R(\bar{\mathbf{w}}_T) - R(\mathbf{w}^*) = O\Big((L/\rho)n^{-\frac{1}{2}} \log n \log^2(1/\delta)\Big). \quad (4.6)$$

Theorem 4 is proved in Appendix E. In Theorem E.1, we also develop high-probability bounds of order $O(n^{-\frac{1}{2}} \log n)$ on plain generalization errors $|F(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \mathbf{v}^*)|$.

5. Nonconvex-Nonconcave Objectives

In this section, we extend our analysis to nonconvex-nonconcave minimax learning problems.

5.1. Stability and Generalization of SGDA

We first study the generalization bounds of SGDA for WC-WC problems. The proof is given in Appendix F.1. In Appendix F.2, we further give high-probability bounds.

Theorem 5 (Weak generalization bound)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be produced by (4.1) with T iterations. Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is ρ-WC-WC and $|f(\cdot, \cdot; z)| \le 1$. If Assumption 1 holds and $\eta_t = c/t$, then the weak PD generalization error of SGDA is bounded by*

$$O\Big(\Big(1 + \frac{\sqrt{T}}{n}\Big)T^{c\rho}\Big)^{\frac{2}{2c\rho+3}} \Big(\frac{1}{n}\Big)^{\frac{2c\rho+1}{2c\rho+3}}.$$

Remark 7. If $T = O(n^2)$, our weak PD generalization error bound is of the order $O\big(n^{-\frac{2c\rho+1}{2c\rho+3}} T^{\frac{2c\rho}{2c\rho+3}}\big)$. This is the first generalization bound of SGDA for nonsmooth and nonconvex-nonconcave objectives. Farnia

& Ozdaglar (2020) also studied generalization under nonconvex-nonconcave setting but required the objectives to be smooth, which is relaxed to a milder WC-WC assumption here. Our analysis readily applies to stochastic gradient descent (SGD) with nonsmooth weakly-convex functions, which has not been studied in the literature.

We further consider a variant of weak-convexity-weak-concavity. The proof is given in Appendix F.3.

Theorem 6 (Weak generalization bound)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be produced by (4.1) with T iterations. Let Assumptions 1, 2 hold. Assume there are non-negative numbers $\{\rho_t\}_{t\in\mathbb{N}}$ such that the following inequality holds a.s.*

$$\Big\langle \begin{pmatrix} \mathbf{w}_t - \mathbf{w} \\ \mathbf{v}_t - \mathbf{v} \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v}) \\ \nabla_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v}) - \nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t) \end{pmatrix} \Big\rangle \geq -\rho_t \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w} \\ \mathbf{v}_t - \mathbf{v} \end{pmatrix} \right\|_2^2, \ \forall \mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}. \quad (5.1)$$

Then the weak PD generalization error can be bounded by

$$O\Big(n^{-1} \sum_{t=1}^{T} \Big(\eta_t^2 + \frac{1}{n}\Big) \exp\Big(\sum_{k=t+1}^{T} \big(2\rho_k\eta_k + (L^2+1)\eta_k^2\big)\Big)\Big)^{\frac{1}{2}}.$$

Eq. (5.1) allows the empirical objective F_S to have varying weak-convexity-weak-concavity at different iterates encountered by the algorithm. This is motivated by the observation that the nonconvex-nonconcave function can have approximate convexity-concavity around a saddle point. For these problems, we can expect the weak-convexity-weak-concavity parameter ρ_t to decrease along the optimization process (Sagun et al., 2017; Yuan et al., 2019).

Remark 8. If F_S is convex-concave, then $\rho_t = 0$ and we can take $\eta_t \asymp 1/\sqrt{T}$ to show that SGDA with T iterations enjoys the generalization bound $O\big(1/\sqrt{n} + \sqrt{T}/n\big)$. This extends Theorem 3 since we only require the convexity-concavity of F_S here instead of $f(\cdot, \cdot; z)$ for all z in Theorem 3. If $\rho_t = O(t^{-\alpha})$ ($\alpha \in (0,1)$), then we can take $\eta_t \asymp t^{\min\{\alpha-1, -\frac{1}{2}\}}/\log T$ (note $\sum_{t=1}^T \eta_t^2 = O(1), \sum_{t=1}^T \eta_t\rho_t = O(1)$) to show that SGDA with T iterations enjoys the weak PD generalization bound $O\big(1/\sqrt{n} + \sqrt{T}/n\big)$. As compared to Theorem 5, the assumption (5.1) allows us to use much larger step sizes ($O(t^{-\beta}), \beta \in (0,1)$ vs $O(t^{-1})$). This larger step size allows for a better trade-off between generalization and optimization. We note that a recent work (Richards & Rabbat, 2021) considered gradient descent under an assumption similar to (5.1), and developed interesting generalization bounds for $\eta_t = O(t^{-\beta})$ ($\beta \in (0,1)$). However, their discussions do not apply to the important SGD and require an additional assumption on the Lipschitz continuity of Hessian matrix which may be restrictive. It is direct to extend Theorem 6 to SGD for learning with weakly-convex functions for relaxing the step size under Eq. (5.1). Therefore, our stability analysis even gives novel results in the standard nonconvex learning setting. We introduce a novel technique in achieving this improvement. Specifically, let $\delta_t := \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + \|\mathbf{v}_t - \mathbf{v}'_t\|_2^2$, where $(\mathbf{w}_t, \mathbf{v}_t), (\mathbf{w}'_t, \mathbf{v}'_t)$ are SGDA iterates for neighboring datasets S and S'. For the stability bounds in Section 4.1, we first handle δ_{t+1} according to different realization of i_t and then consider the expectation w.r.t. i_t. While for ρ-WC-WC problems, we first take expectation w.r.t. i_t and then show $\mathbb{E}_{i_t}[\delta_{t+1}]$ would change along the iterations.

5.2. Stability and Generalization of AGDA and Beyond

We now study the Alternating Gradient Descent Ascent (AGDA) proposed recently to optimize nonconvex-nonconcave problems (Yang et al., 2020). Let $\{\eta_{\mathbf{w},t}, \eta_{\mathbf{v},t}\}_t$ be a sequence of positive stepsizes for updating $\{\mathbf{w}_t, \mathbf{v}_t\}_t$. At each iteration, we randomly draw i_t and j_t from the uniform distribution over $[n]$ and do the update

$$\begin{cases} \mathbf{w}_{t+1} = \mathrm{Proj}_{\mathcal{W}}\big(\mathbf{w}_t - \eta_{\mathbf{w},t} \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\big), \\ \mathbf{v}_{t+1} = \mathrm{Proj}_{\mathcal{V}}\big(\mathbf{v}_t + \eta_{\mathbf{v},t} \nabla_{\mathbf{v}} f(\mathbf{w}_{t+1}, \mathbf{v}_t; z_{j_t})\big). \end{cases} \quad (5.2)$$

This algorithm differs from SGDA in two aspects. First, it randomly selects two examples to update $\mathbf{w}$ and $\mathbf{v}$ per iteration. Second, it uses the updated $\mathbf{w}_{t+1}$ when updating $\mathbf{v}_{t+1}$. Theorem 7 to be proved in Appendix G provides generalization bounds for AGDA.

Theorem 7 (Weak generalization bounds)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be the sequence produced by (5.2). If Assumptions 1, 2 hold and $\eta_{\mathbf{w},t} + \eta_{\mathbf{v},t} \leq \frac{c}{t}$ for some $c > 0$, then the weak PD generalization error can be upper bounded by $O\big(n^{-1} T^{\frac{cL}{cL+1}}\big)$.*

Global convergence on AGDA was studied based on the two-sided PL condition defined below (Yang et al., 2020), which means the suboptimality of function values can be bounded by gradients and were shown for several rich classes of functions (Karimi et al., 2016). We also refer to the two-sided PL condition as the gradient dominance condition.

Assumption 3. Assume F_S satisfies the two-sided PL condition, i.e., there exist constants $\beta_1(S), \beta_2(S) > 0$ such that the following inequalities hold for all $\mathbf{w} \in \mathcal{W}, \mathbf{v} \in \mathcal{V}$

$$2\beta_1(S)\big(F_S(\mathbf{w}, \mathbf{v}) - \inf_{\mathbf{w}' \in \mathcal{W}} F_S(\mathbf{w}', \mathbf{v})\big) \leq \|\nabla_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v})\|_2^2,$$

$$2\beta_2(S)\big(\sup_{\mathbf{v}' \in \mathcal{V}} F_S(\mathbf{w}, \mathbf{v}') - F_S(\mathbf{w}, \mathbf{v})\big) \leq \|\nabla_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v})\|_2^2.$$

As a combination of our generalization bounds and optimization error bounds in Yang et al. (2020), we can derive the following informal corollary on primal population risks by early stopping the algorithm to balance the optimization and generalization. It gives the first primal risk bounds for learning with nonconvex-nonconcave functions. The precise statement can be found in Corollary G.3.

Corollary 8 (Informal)**.** *Let $\beta_1, \rho > 0$. Assume for all $\mathbf{w}$, the function $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ is ρ-strongly concave. Let Assumptions 1, 2, 3 with $\beta_1(S) \geq \beta_1, \beta_2(S) \geq \rho$ hold. Then AGDA with some appropriate step size and $T \asymp \left(n\beta_1^{-2}\rho^{-3}\right)^{\frac{cL+1}{2cL+1}}$ satisfy ($c \asymp 1/(\beta_1\rho^2)$)*

$$\mathbb{E}\big[R(\mathbf{w}_T)\big] - R(\mathbf{w}^*) = O\Big(n^{-\frac{cL+1}{2cL+1}}\beta_1^{-\frac{2cL}{2cL+1}}\rho^{-\frac{5cL+1}{2cL+1}}\Big).$$

For gradient dominated problems, we further have the following error bounds to be proved in Appendix H. Note we do not need the smoothness assumption here.

Theorem 9. *Let A be an algorithm. Let Assumptions 1, 3 hold. Let $\mathbf{u}_S = (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ and $\mathbf{u}_S^{(S)}$ be the projection of $\mathbf{u}_S$ onto the set of stationary points of F_S. Then,*

$$\big|\mathbb{E}\big[F(\mathbf{u}_S) - F_S(\mathbf{u}_S)\big]\big| \leq \frac{2G^2}{n}\max\Big\{\mathbb{E}\big[1/\beta_1(S)\big], \mathbb{E}\big[1/\beta_2(S)\big]\Big\} + 2G\mathbb{E}\big[\|\mathbf{u}_S - \mathbf{u}_S^{(S)}\|_2\big].$$

Remark 9. Note $\|\mathbf{u}_S - \mathbf{u}_S^{(S)}\|_2$ measures how far the point found by A is from the set of stationary points of F_S, and can be interpreted as an optimization error. Therefore, Theorem 9 gives a connection between generalization error and optimization error. For a variant of AGDA with noiseless stochastic gradients, it was shown that $\|(\mathbf{w}_T, \mathbf{v}_T) - (\mathbf{w}_T, \mathbf{v}_T)^{(S)}\|_2$ decays linearly w.r.t. T (Yang et al., 2020). We can plug this linear convergent optimization bound into Theorem 9 to directly get generalization bounds. If A returns a saddle point of F_S, then $\|\mathbf{u}_S - \mathbf{u}_S^{(S)}\|_2 = 0$ and therefore $\big|\mathbb{E}\big[F(\mathbf{u}_S) - F_S(\mathbf{u}_S)\big]\big| = O\big(n^{-1}\max\{\mathbb{E}[1/\beta_1(S)], \mathbb{E}[1/\beta_2(S)]\}\big)$. Generalization errors of this particular ESP were studied in Zhang et al. (2020) for SC-SC minimax problems, which were extended to more general gradient-dominated problems in Theorem 9. Furthermore, Theorem 9 applies to any optimization algorithm instead of the specific ESP. It should be mentioned that Zhang et al. (2020) addressed PD population risks, while we consider plain generalization errors.

6. Experiments

In this subsection, we report preliminary experimental results to verify our theoretical results. We consider two datasets available at the LIBSVM website: svmguide3 and w5a (Chang & Lin, 2011). For each dataset, we use 80 percents of the dataset for as the training dataset S. We follow the experimental setup in Hardt et al. (2016) to study how the stability of SGDA would behave along the learning process. To this aim, we build a neighboring dataset S' by changing the last example of the training set S. We apply the same randomized algorithm to S and S' and get two model sequences $\{(\mathbf{w}_t, \mathbf{v}_t)\}$ and $\{(\mathbf{w}'_t, \mathbf{v}'_t)\}$. We then evaluate the Euclidean distance $\triangle_t = \big(\|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + \|\mathbf{v}_t - \mathbf{v}'_t\|_2^2\big)^{\frac{1}{2}}$. Firstly, we consider the algorithm SOLAM (Ying et al., 2016), which is the SGDA for the solving the problem (2.3), which is a minimax reformulation of the AUC maximization problem. We consider step sizes $\eta_t = \eta/\sqrt{T}$ with $\eta \in \{0.1, 0.3, 1, 3\}$. We repeat the experiments 100 times and report the average of the experimental results as well as the standard deviation. In Figure 6, we report $\triangle_t$ as a function of the number of passes (iteration number divided by n). It is clear the Euclidean distances continue to increase in the learning process. Furthermore, the Euclidean distances increase if we consider larger step sizes. This phenomenon is consistent with our stability bounds in Theorem 2.

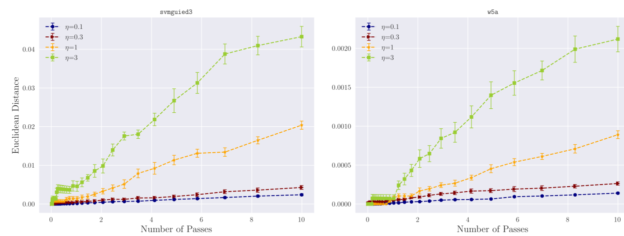

Figure 1. $\triangle_t$ versus the number of passes. Left: svmguide3, right: w8a.

7. Conclusion

We present a comprehensive stability and generalization analysis of stochastic algorithms for minimax objective functions. We introduce various generalization measures and stability measures, and present a systematic study on their quantitative relationship. We develop the first minimax optimal risk bounds for SGDA in a general convex-concave case, covering both smooth and nonsmooth setting. We also give the first non-trivial risk bounds for nonconvex-nonconcave problems. Our analysis shows how to early-stop the algorithm in practice to train a model with better generalization. Our theoretical results have potential applications in developing differentially private algorithms to handle sensitive data.

There remain some interesting problems for further investigation. Our primal generalization bounds require a strong concavity assumption. It would be interesting and challenging to remove this assumption. Also, it would be interesting to show how the strong concavity in dual variables would help generalization in a nonconvex setting.

Acknowledgments

We are grateful to the anonymous reviewers for their constructive comments and suggestions. The work of Yiming Ying is partially supported by NSF grants IIS-1816227 and IIS-2008532. Tianbao's work was partially supported by

NSF #1933212 and NSF CAREER Award #1844403.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Audibert, J.-Y. and Catoni, O. Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794, 2011.

Balamurugan, P. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *Advance In Neural Information Processing Systems*, pp. 1416–1424, 2016.

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

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

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

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

Brownlees, C., Joly, E., Lugosi, G., et al. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

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

Charles, Z. and Papailiopoulos, D. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pp. 744–753, 2018.

Chen, R. S., Lucier, B., Singer, Y., and Syrgkanis, V. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems*, pp. 4705–4714, 2017.

Chen, Y., Jin, C., and Yu, B. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pp. 1125–1134, 2018.

Deng, Z., He, H., and Su, W. J. Toward better generalization bounds with locally elastic stability. *arXiv preprint arXiv:2010.13988*, 2020.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058, 2017.

Elisseeff, A., Evgeniou, T., and Pontil, M. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

Farnia, F. and Ozdaglar, A. Train simultaneously, generalize better: Stability of gradient-based minimax learners. *arXiv preprint arXiv:2010.12561*, 2020.

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

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, pp. 6726–6736. 2019.

Gao, W., Jin, R., Zhu, S., and Zhou, Z.-H. One-pass auc optimization. In *International Conference on Machine Learning*, pp. 906–914, 2013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014.

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

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, pp. 6938–6948, 2019.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning*, pp. 795–811, 2016.

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

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Interna-*

tional Conference on Machine Learning, pp. 5809–5819, 2020.

Lei, Y. and Ying, Y. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021a.

Lei, Y. and Ying, Y. Stochastic proximal auc maximization. *Journal of Machine Learning Research*, 22(61): 1–45, 2021b.

Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

Lin, J., Camoriano, R., and Rosasco, L. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pp. 2340–2348, 2016.

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

Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic AUC maximization with O(1/n)-convergence rate. In *International Conference on Machine Learning*, pp. 3195–3203, 2018.

Liu, M., Rafique, H., Lin, Q., and Yang, T. First-order convergence theory for weakly-convex-weakly-concave min-max problems, 2020.

Liu, T., Lugosi, G., Neu, G., and Tao, D. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pp. 2159–2167, 2017.

Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pp. 6370–6381, 2020.

London, B. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2931–2940, 2017.

Luo, L., Ye, H., Huang, Z., and Zhang, T. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Madden, L., Dall'Anese, E., and Becker, S. High probability convergence and uniform stability bounds for nonconvex stochastic gradient descent. *arXiv preprint arXiv:2006.05610*, 2020.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pp. 605–638, 2018.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pp. 2971–2980, 2017.

Nedić, A. and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

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

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Raghavan, P. Probabilistic construction of deterministic algorithms: approximating packing integer programs. *Journal of Computer and System Sciences*, 37(2):130–143, 1988.

Rakhlin, A., Mukherjee, S., and Poggio, T. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

Richards, D. and Rabbat, M. Learning with gradient descent and weakly convex losses. *arXiv preprint arXiv:2101.04968*, 2021.

Richards, D. et al. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(2020), 2020.

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

Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

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

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

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32: 12680–12691, 2019.

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

Xu, Y., Zhu, S., Yang, S., Zhang, C., Jin, R., and Yang, T. Learning with non-convex truncated losses by sgd. In *Uncertainty in Artificial Intelligence*, pp. 701–711, 2020.

Yan, Y., Xu, Y., Lin, Q., Liu, W., and Yang, T. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Yang, J., Kiyavash, N., and He, N. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Ying, Y., Wen, L., and Lyu, S. Stochastic online AUC maximization. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2016.

Yuan, Z., Yan, Y., Jin, R., and Yang, T. Stagewise training accelerates convergence of testing error over sgd. In *Advances in Neural Information Processing Systems*, pp. 2604–2614, 2019.

Zhang, J., Hong, M., Wang, M., and Zhang, S. Generalization bounds for stochastic saddle point problems. *arXiv preprint arXiv:2006.02067*, 2020.

Zhao, P., Hoi, S. C., Jin, R., and Yang, T. Online AUC maximization. In *International Conference on Machine Learning*, pp. 233–240, 2011.

Appendix for "Stability and Generalization of Stochastic Gradient Methods for Minimax Problems"

A. Notations

We collect in Table A.1 the notations of performance measures used in this paper.

	Notation	Meaning	Definition
Weak Measure	$\triangle^w(\mathbf{w},\mathbf{v})$	weak PD population risk	$\sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}\big[F(\mathbf{w},\mathbf{v}')\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}\big[F(\mathbf{w}',\mathbf{v})\big]$
	$\triangle^w_S(\mathbf{w},\mathbf{v})$	weak PD empirical risk	$\sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}\big[F_S(\mathbf{w},\mathbf{v}')\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}\big[F_S(\mathbf{w}',\mathbf{v})\big]$
	$\triangle^w(\mathbf{w},\mathbf{v}) - \triangle^w_S(\mathbf{w},\mathbf{v})$	weak PD generalization error	$\big(\sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}\big[F(\mathbf{w},\mathbf{v}')\big] - \sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}\big[F_S(\mathbf{w},\mathbf{v}')\big]\big)$ $+\big(\inf_{\mathbf{w}'\in\mathcal{V}} \mathbb{E}\big[F_S(\mathbf{w}',\mathbf{v})\big] - \inf_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}\big[F(\mathbf{w}',\mathbf{v})\big]\big)$
Strong Measure	$\triangle^s(\mathbf{w},\mathbf{v})$	strong PD population Risk	$\sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w},\mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} F(\mathbf{w}',\mathbf{v})$
	$\triangle^s_S(\mathbf{w},\mathbf{v})$	strong PD empirical Risk	$\sup_{\mathbf{v}'\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}',\mathbf{v})$
	$\triangle^s(\mathbf{w},\mathbf{v}) - \triangle^s_S(\mathbf{w},\mathbf{v})$	strong PD generalization error	$\big(\sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w},\mathbf{v}') - \sup_{\mathbf{v}'\in\mathcal{W}} F_S(\mathbf{w},\mathbf{v}')\big)$ $+\big(\inf_{\mathbf{w}'\in\mathcal{V}} F_S(\mathbf{w}',\mathbf{v}) - \inf_{\mathbf{w}'\in\mathcal{W}} F(\mathbf{w}',\mathbf{v})\big)$
Primal Measure	$R(\mathbf{w}) - \inf_{\mathbf{w}'\in\mathcal{W}} R(\mathbf{w}')$	excess primal population risk	$\sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w},\mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} \sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w}',\mathbf{v}')$
	$R_S(\mathbf{w}) - \inf_{\mathbf{w}'\in\mathcal{W}} R_S(\mathbf{w}')$	excess primal empirical risk	$\sup_{\mathbf{v}'\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} \sup_{\mathbf{v}'\in\mathcal{V}} F_S(\mathbf{w}',\mathbf{v}')$
	$R(\mathbf{w}) - R_S(\mathbf{w})$	primal generalization error	$\sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w},\mathbf{v}') - \sup_{\mathbf{v}'\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v}')$
	$F(\mathbf{w},\mathbf{v}) - F_S(\mathbf{w},\mathbf{v})$	plain generalization error	

Table A.1. Notations on Measures of Performance.

We collect in Table A.2 the stability measures for a (randomized) algorithm A.

Stability Measure	Definition
Weak Stability	$\sup_z \Big(\sup_{\mathbf{v}'\in\mathcal{V}} \mathbb{E}_A\big[f(A_\mathbf{w}(S),\mathbf{v}';z) - f(A_\mathbf{w}(S'),\mathbf{v}';z)\big] + \sup_{\mathbf{w}'\in\mathcal{W}} \mathbb{E}_A\big[f(\mathbf{w}',A_\mathbf{v}(S);z) - f(\mathbf{w}',A_\mathbf{v}(S');z)\big]\Big)$
Argument Stability	$\mathbb{E}_A\left[\left\|\begin{pmatrix} A_\mathbf{w}(S) - A_\mathbf{w}(S') \\ A_\mathbf{v}(S) - A_\mathbf{v}(S')\end{pmatrix}\right\|_2\right]$ or $\left\|\begin{pmatrix} A_\mathbf{w}(S) - A_\mathbf{w}(S') \\ A_\mathbf{v}(S) - A_\mathbf{v}(S')\end{pmatrix}\right\|_2$
Uniform Stability	$\sup_z \big[f(A_\mathbf{w}(S),A_\mathbf{v}(S);z) - f(A_\mathbf{w}(S'),A_\mathbf{v}(S');z)\big]$

Table A.2. Stability Measures. Here S and S' are neighboring datasets.

B. Proof of Theorem 1

In this section, we prove Theorem 1 on the connection between stability measure and generalization.

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 $g,\tilde{g}$, we have the basic inequalities

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

B.1. Proof of Part (a)

We first prove Part (a). It follows from (B.1) that

$$\begin{aligned}\triangle^w(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))-\triangle^w_S(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))\leq&\sup_{\mathbf{v}'\in\mathcal{V}}\mathbb{E}[F(A_{\mathbf{w}}(S),\mathbf{v}')-F_S(A_{\mathbf{w}}(S),\mathbf{v}')]\\&+\sup_{\mathbf{w}'\in\mathcal{W}}\mathbb{E}[F_S(\mathbf{w}',A_{\mathbf{v}}(S))-F(\mathbf{w}',A_{\mathbf{v}}(S))].\end{aligned}$$

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

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

The stated bound in Part (a) then follows directly from the definition of stability.

B.2. Proof of Part (b)

The following lemma quantifies the sensitivity of the optimal $\mathbf{v}$ w.r.t. the perturbation of $\mathbf{w}$.

Lemma B.1 (Lin et al. 2020)**.** *Let $\phi:\mathcal{W}\times\mathcal{V}\mapsto\mathbb{R}$. Assume that for any $\mathbf{w}$, the function $\mathbf{v}\mapsto\phi(\mathbf{w},\mathbf{v})$ is ρ-strongly-concave. Suppose for any $(\mathbf{w},\mathbf{v}),(\mathbf{w}',\mathbf{v}')$ we have*

$$\big\|\nabla_{\mathbf{v}}\phi(\mathbf{w},\mathbf{v})-\nabla_{\mathbf{v}}\phi(\mathbf{w}',\mathbf{v})\big\|_2\leq L\|\mathbf{w}-\mathbf{w}'\|_2.$$

For any $\mathbf{w}$, denote $\mathbf{v}^*(\mathbf{w})=\arg\max_{\mathbf{v}\in\mathcal{V}}\phi(\mathbf{w},\mathbf{v})$. Then for any $\mathbf{w},\mathbf{w}'\in\mathcal{W}$, we have

$$\big\|\mathbf{v}^*(\mathbf{w})-\mathbf{v}^*(\mathbf{w}')\|_2\leq\frac{L}{\rho}\|\mathbf{w}-\mathbf{w}'\|_2.$$

We now prove Part (b). For any S, let $\mathbf{v}_S^* = \arg\max_{\mathbf{v}\in\mathcal{V}} F(A_{\mathbf{w}}(S), \mathbf{v})$. According to the symmetry between z_i and z_i' we know

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

where the last identity holds since z_i is independent of $A_{\mathbf{w}}(S^{(i)})$ and $\mathbf{v}^*_{S^{(i)}}$.

According to Assumption 1, we know

$$\begin{aligned}&f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i) - f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; z_i) \\ &= f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i) - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_S; z_i) + f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_S; z_i) - f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; z_i) \\ &\leq G\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2 + G\big\|\mathbf{v}^*_{S^{(i)}} - \mathbf{v}^*_S\big\|_2 \leq \big(1 + L/\rho\big)G\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2,\end{aligned} \tag{B.2}$$

where in the last inequality we have used Lemma B.1 due to the strong concavity of $\mathbf{v} \mapsto F(\mathbf{w}, \mathbf{v})$ for any $\mathbf{w}$. As a combination of the above two inequalities, we get

$$\begin{aligned}\mathbb{E}\big[\sup_{\mathbf{v}'\in\mathcal{V}} F(A_{\mathbf{w}}(S), \mathbf{v}')\big] &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; z_i)\big] + \frac{\big(1+L/\rho\big)G}{n}\sum_{i=1}^n \mathbb{E}\big[\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2\big] \\ &= \mathbb{E}\big[F_S(A_{\mathbf{w}}(S), \mathbf{v}^*_S)\big] + \frac{\big(1+L/\rho\big)G}{n}\sum_{i=1}^n \mathbb{E}\big[\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2\big] \\ &\leq \mathbb{E}\big[\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}')\big] + \frac{\big(1+L/\rho\big)G}{n}\sum_{i=1}^n \mathbb{E}\big[\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2\big].\end{aligned} \tag{B.3}$$

The stated bound in Part (b) then follows.

B.3. Proof of Part (c)

In a similar way, one can show that

$$\mathbb{E}\big[\inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}', A_{\mathbf{v}}(S))\big] - \mathbb{E}\big[\inf_{\mathbf{w}'\in\mathcal{W}} F(\mathbf{w}', A_{\mathbf{v}}(S))\big] \leq \frac{\big(1+L/\rho\big)G}{n}\sum_{i=1}^n \mathbb{E}\big[\|A_{\mathbf{v}}(S^{(i)}) - A_{\mathbf{v}}(S)\|_2\big].$$

The above inequality together with (B.3) then implies

$$\begin{aligned}&\mathbb{E}\big[\triangle^s(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))\big] - \mathbb{E}\big[\triangle^s_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))\big] \\ &= \mathbb{E}\big[\sup_{\mathbf{v}'\in\mathcal{V}} F(A_{\mathbf{w}}(S), \mathbf{v}')\big] - \mathbb{E}\big[\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}')\big] + \mathbb{E}\big[\inf_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}', A_{\mathbf{v}}(S))\big] - \mathbb{E}\big[\inf_{\mathbf{w}'\in\mathcal{W}} F(\mathbf{w}', A_{\mathbf{v}}(S))\big] \\ &\leq \big(1+L/\rho\big)G\mathbb{E}\big[\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\|_2\big] + \big(1+L/\rho\big)G\mathbb{E}\big[\|A_{\mathbf{v}}(S^{(i)}) - A_{\mathbf{v}}(S)\|_2\big] \\ &\leq \big(1+L/\rho\big)G\sqrt{2}\mathbb{E}\Big[\Big\|\begin{pmatrix}A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S) \\ A_{\mathbf{v}}(S^{(i)}) - A_{\mathbf{v}}(S)\end{pmatrix}\Big\|_2\Big],\end{aligned}$$

where we have used the elementary inequality $a + b \leq \sqrt{2(a^2+b^2)}$. This proves the stated bound in Part (c).

B.4. Proof of Part (d)

To prove Part (d) on high-probability bounds, we need to introduce some lemmas.

The following lemma establishes a concentration inequality for a summation of weakly-dependent random variables. We denote $S\backslash\{z_i\}$ the set $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$. The L_p-norm of a random variable Z is denoted by $\|Z\|_p := \big(\mathbb{E}[|Z|^p]\big)^{\frac{1}{p}}, p \geq 1$.

Lemma B.2 (Bousquet et al. 2020). *Let $S = \{z_1, \dots, z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$ and $M > 0$. Let $g_1, \dots, g_n$ be some functions $g_i : \mathcal{Z}^n \mapsto \mathbb{R}$ such that the following holds for any $i \in [n]$*

- $\big|\mathbb{E}_{S\setminus\{z_i\}}[g_i(S)]\big| \le M$ *almost surely (a.s.),*
- $\mathbb{E}_{z_i}\big[g_i(S)\big] = 0$ *a.s.,*
- *for any $j \in [n]$ with $j \neq i$, and $z''_j \in \mathcal{Z}$*

$$\big|g_i(S) - g_i(z_1, \dots, z_{j-1}, z''_j, z_{j+1}, \dots, z_n)\big| \le \beta. \tag{B.4}$$

Then, for any $p \ge 2$

$$\Big\|\sum_{i=1}^n g_i(S)\Big\|_p \le 12\sqrt{6}pn\beta\lceil\log_2 n\rceil + 3\sqrt{2}M\sqrt{pn}.$$

The following lemma shows how to relate moment bounds of random variables to tail behavior.

Lemma B.3 (Bousquet et al. 2020; Vershynin 2018). *Let $a, b \in \mathbb{R}_+$ and $\delta \in (0, 1/e)$. Let Z be a random variable with $\|Z\|_p \le \sqrt{p}a + pb$ for any $p \ge 2$. Then with probability at least $1-\delta$*

$$|Z| \le e\Big(a\sqrt{\log(e/\delta)} + b\log(e/\delta)\Big).$$

With the above lemmas we are now ready to prove Part (d). For any S, denote

$$\mathbf{v}^*_S = \arg\max_{\mathbf{v}\in\mathcal{V}} F(A_{\mathbf{w}}(S), \mathbf{v}). \tag{B.5}$$

We have the following error decomposition

$$nF(A_{\mathbf{w}}(S), \mathbf{v}^*_S) - n\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}) = \sum_{i=1}^n \mathbb{E}_Z\big[f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; Z) - \mathbb{E}_{z'_i}[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)]\big] +$$
$$\sum_{i=1}^n \mathbb{E}_{z'_i}\Big[\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big] + \sum_{i=1}^n \mathbb{E}_{z'_i}\Big[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big] - n\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}).$$

By the definition of $\mathbf{v}^*_{S^{(i)}}$ we know $\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] \ge \mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_S; Z)]$. It then follows that

$$nF(A_{\mathbf{w}}(S), \mathbf{v}^*_S) - n\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}') \le \sum_{i=1}^n \mathbb{E}_Z\big[f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; Z) - \mathbb{E}_{z'_i}[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_S; Z)]\big] +$$
$$\sum_{i=1}^n \mathbb{E}_{z'_i}\Big[\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big] + \sum_{i=1}^n \mathbb{E}_{z'_i}\Big[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big] - n\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}').$$

According to (B.2), we know

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}_{z'_i}\Big[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big] &\le \big(1 + L/\rho\big)G\sum_{i=1}^n \mathbb{E}_{z'_i}\big[\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2\big] + \sum_{i=1}^n f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; z_i)\\ &= \big(1 + L/\rho\big)G\sum_{i=1}^n \mathbb{E}_{z'_i}\big[\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2\big] + nF_S(A_{\mathbf{w}}(S), \mathbf{v}^*_S)\\ &\le \big(1 + L/\rho\big)G\sum_{i=1}^n \mathbb{E}_{z'_i}\big[\big\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S)\big\|_2\big] + n\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}').\end{aligned}$$

As a combination of the above two inequalities, we derive

$$nF(A_{\mathbf{w}}(S), \mathbf{v}^*_S) - n\sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}') \le \big(2 + L/\rho\big)nG\epsilon + \sum_{i=1}^n g_i(S), \tag{B.6}$$

where we introduce

$$g_i(S) = \mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big]$$

and use the inequality

$$f(A_{\mathbf{w}}(S), \mathbf{v}^*_S; Z) - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_S; Z) \le G\|A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S^{(i)})\|_2 \le G\epsilon.$$

Due to the symmetry between z_i and Z, we know $\mathbb{E}_{z_i}[g_i(S)] = 0$. The inequality $|\mathbb{E}_{S\backslash\{z_i\}}[g_i(S)]| \le 2R$ is also clear.

For any $j \ne i$ and any $z_j'' \in \mathcal{Z}$, we know

$$\begin{aligned}
&\Big|\mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i)\Big] - \mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}_j), \mathbf{v}^*_{S^{(i)}_j}; Z)] - f(A_{\mathbf{w}}(S^{(i)}_j), \mathbf{v}^*_{S^{(i)}_j}; z_i)\Big]\Big| \\
&\le \Big|\mathbb{E}_{z_i'}\Big[\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] - \mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}_j), \mathbf{v}^*_{S^{(i)}_j}; Z)]\Big]\Big| + \Big|\mathbb{E}_{z_i'}\Big[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z_i) - f(A_{\mathbf{w}}(S^{(i)}_j), \mathbf{v}^*_{S^{(i)}_j}; z_i)\Big]\Big|,
\end{aligned}$$

where $S^{(i)}_j$ is the set derived by replacing the j-th element of $S^{(i)}$ with z_j''. For any z, there holds

$$\begin{aligned}
&\big|f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z) - f(A_{\mathbf{w}}(S^{(i)}_j), \mathbf{v}^*_{S^{(i)}_j}; z)\big| \\
&\le \big|f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; z) - f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}_j}; z)\big| + \big|f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}_j}; z) - f(A_{\mathbf{w}}(S^{(i)}_j), \mathbf{v}^*_{S^{(i)}_j}; z)\big| \\
&\le G\|\mathbf{v}^*_{S^{(i)}} - \mathbf{v}^*_{S^{(i)}_j}\|_2 + G\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S^{(i)}_j)\|_2 \le \big(L/\rho + 1\big)G\|A_{\mathbf{w}}(S^{(i)}) - A_{\mathbf{w}}(S^{(i)}_j)\|_2,
\end{aligned}$$

where in the last inequality we have used the definition of $\mathbf{v}^*_{S^{(i)}}$ and Lemma B.1 with $\phi = F$. Therefore $g_i(S)$ satisfies the condition (B.4) with $\beta = \big(L/\rho + 1\big)G\epsilon$. Therefore, all the conditions of Lemma B.2 hold and we can apply Lemma B.2 to derive the following inequality for any $p \ge 2$

$$\Big\|\sum_{i=1}^n g_i(S)\Big\|_p \le 12\sqrt{6}pn\big(L/\rho + 1\big)G\epsilon\lceil\log_2 n\rceil + 6\sqrt{2}R\sqrt{pn}.$$

This together with Lemma B.3 implies the following inequality with probability $1-\delta$

$$\Big|\sum_{i=1}^n g_i(S)\Big| \le e\Big(6R\sqrt{2n\log(e/\delta)} + 12\sqrt{6}n\big(L/\rho+1\big)G\epsilon\log(e/\delta)\lceil\log_2 n\rceil\Big).$$

We can plug the above inequality back into (B.6) and derive the following inequality with probability at least $1-\delta$

$$F(A_{\mathbf{w}}(S), \mathbf{v}^*_S) - \sup_{\mathbf{v}'\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v}') \le \big(2 + L/\rho\big)G\epsilon + e\Big(6R\sqrt{2n^{-1}\log(e/\delta)} + 12\sqrt{6}\big(L/\rho+1\big)G\epsilon\log(e/\delta)\lceil\log_2 n\rceil\Big).$$

This prove the stated bound in Part (d).

B.5. Proof of Part (e)

Part (e) is standard in the literature (Bousquet et al., 2020).

C. Proof of Theorem 2

In this section, we present the proof of Theorem 2 on the argument stability of SGDA.

C.1. Approximate Nonexpansiveness of Gradient Map

To prove stability bounds, we need to study the expansiveness of the gradient map

$$G_{f,\eta} : \begin{pmatrix}\mathbf{w}\\ \mathbf{v}\end{pmatrix} \mapsto \begin{pmatrix}\mathbf{w} - \eta\nabla_{\mathbf{w}}f(\mathbf{w},\mathbf{v})\\ \mathbf{v} + \eta\nabla_{\mathbf{v}}f(\mathbf{w},\mathbf{v})\end{pmatrix}$$

associated with a (strongly) convex-concave f. The following lemma shows that $G_{f,\eta}$ is approximately nonexpansive in both the Lipschitz continuous case and the smooth case. It also shows that $G_{f,\eta}$ is nonexpansive if f is SC-SC and the step size is small. Part (b) can be found in Farnia & Ozdaglar (2020).

Lemma C.1. *Let f be ρ-SC-SC with $\rho \geq 0$.*

(a) If Assumption 1 holds, then

$$\left\| \begin{pmatrix} \mathbf{w} - \eta \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}' - \eta \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') \\ \mathbf{v}' + \eta \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 \leq (1 - 2\rho\eta) \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2 + 8G^2\eta^2.$$

(b) If Assumption 2 holds, then

$$\left\| \begin{pmatrix} \mathbf{w} - \eta \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}' - \eta \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') \\ \mathbf{v}' + \eta \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 \leq (1 - 2\rho\eta + L^2\eta^2) \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2.$$

To prove Lemma C.1 we require the following standard lemma (Rockafellar, 1976).

Lemma C.2. *Let f be a ρ-SC-SC function, $\rho \geq 0$. Then*

$$\left\langle \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) - \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') \\ \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') - \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) \end{pmatrix} \right\rangle \geq \rho \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2. \tag{C.1}$$

Proof of Lemma C.1. It is clear that

$$\begin{aligned} A := \left\| \begin{pmatrix} \mathbf{w} - \eta \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}' - \eta \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') \\ \mathbf{v}' + \eta \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2 \\ + \eta^2 \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') - \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) \\ \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) - \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 - 2\eta \left\langle \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) - \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') \\ \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') - \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) \end{pmatrix} \right\rangle. \end{aligned}$$

Plugging (C.1) to the above inequality, we derive

$$A \leq (1 - 2\rho\eta) \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2 + \eta^2 \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}', \mathbf{v}') - \nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}) \\ \nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}) - \nabla_{\mathbf{v}} f(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2.$$

We can combine the above inequality with the Lipschitz continuity to derive Part (a). We refer to the interested readers to Farnia & Ozdaglar (2020) for the proof of Part (b). ☐

We now prove Theorem 2. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z_1, \ldots, z_{n-1}, z_n'\}$. Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ and $\{\mathbf{w}_t', \mathbf{v}_t'\}$ be the sequence produced by (4.1) w.r.t. S and S', respectively.

C.2. Proof of Part (a)

We first prove Part (a). Note that the projection step is nonexpansive. We consider two cases at the t-th iteration. If $i_t \neq n$, then it follows from Part (a) of Lemma C.1 that

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_{t+1}' \\ \mathbf{v}_{t+1} - \mathbf{v}_{t+1}' \end{pmatrix} \right\|_2^2 &\leq \left\| \begin{pmatrix} \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \mathbf{w}_t' + \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_{i_t}) \\ \mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \mathbf{v}_t' - \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_{i_t}) \end{pmatrix} \right\|_2^2 \\ &\leq \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \right\|_2^2 + 8G^2\eta_t^2. \end{aligned} \tag{C.2}$$

If $i_t = n$, then it follows from the elementary inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$ $(p > 0)$ that

$$\begin{aligned} \left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_{t+1}' \\ \mathbf{v}_{t+1} - \mathbf{v}_{t+1}' \end{pmatrix} \right\|_2^2 &\leq \left\| \begin{pmatrix} \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \mathbf{w}_t' + \eta_t \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') \\ \mathbf{v}_t + \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \mathbf{v}_t' - \eta_t \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') \end{pmatrix} \right\|_2^2 \\ &\leq (1+p) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \right\|_2^2 + (1 + 1/p)\eta_t^2 \left\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') \end{pmatrix} \right\|_2^2. \end{aligned} \tag{C.3}$$

Note that the event $i_t \neq n$ happens with probability $1 - 1/n$ and the event $i_t = n$ happens with probability $1/n$. Therefore, we know

$$\mathbb{E}_{i_t}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq \frac{n-1}{n}\left(\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + 8G^2\eta_t^2\right) + \frac{1+p}{n}\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{8(1+1/p)}{n}\eta_t^2 G^2$$
$$= (1+p/n)\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + 8\eta_t^2 G^2(1+1/(np)).$$

Applying this inequality recursively implies that

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq 8\eta^2 G^2\big(1+1/(np)\big)\sum_{k=1}^{t}\Big(1+\frac{p}{n}\Big)^{t-k} = 8\eta^2 G^2\Big(1+\frac{1}{np}\Big)\frac{n}{p}\Big(\Big(1+\frac{p}{n}\Big)^t - 1\Big)$$
$$= 8\eta^2 G^2\Big(\frac{n}{p}+\frac{1}{p^2}\Big)\Big(\Big(1+\frac{p}{n}\Big)^t - 1\Big).$$

By taking $p = n/t$ in the above inequality and using $(1+1/t)^t \leq e$, we get

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq 16\eta^2 G^2\Big(t+\frac{t^2}{n^2}\Big).$$

The stated bound then follows by Jensen's inequality.

C.3. Proof of Part (b)

We now prove Part (b). Analogous to (C.2), we can use Part (b) of Lemma C.1 to derive

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq (1+L^2\eta_t^2)\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2$$

in the case $i_t \neq n$. We can combine the above inequality and (C.3) to derive

$$\mathbb{E}_{i_t}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq \frac{(n-1)(1+L^2\eta_t^2)}{n}\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{1+p}{n}\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{8(1+1/p)}{n}\eta_t^2 G^2$$
$$\leq \Big(1+L^2\eta_t^2+p/n\Big)\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{8(1+1/p)}{n}\eta_t^2 G^2.$$

Applying this inequality recursively, we derive

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq \frac{8G^2(1+1/p)}{n}\sum_{k=1}^{t}\eta_k^2\prod_{j=k+1}^{t}\Big(1+L^2\eta_j^2+p/n\Big).$$

By the elementary inequality $1 + a \leq \exp(a)$, we further derive

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq \frac{8G^2(1+1/p)}{n}\sum_{k=1}^{t}\eta_k^2\prod_{j=k+1}^{t}\exp\Big(L^2\eta_j^2+p/n\Big)$$
$$= \frac{8G^2(1+1/p)}{n}\sum_{k=1}^{t}\eta_k^2\exp\Big(L^2\sum_{j=k+1}^{t}\eta_j^2+p(t-k)/n\Big)$$
$$\leq \frac{8G^2(1+1/p)}{n}\exp\Big(L^2\sum_{j=1}^{t}\eta_j^2+pt/n\Big)\sum_{k=1}^{t}\eta_k^2.$$

By taking $p = n/t$ we get

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq \frac{8eG^2(1+t/n)}{n}\exp\Big(L^2\sum_{j=1}^{t}\eta_j^2\Big)\sum_{k=1}^{t}\eta_k^2.$$

The stated result then follows from the Jensen's inequality.

C.4. Proof of Part (c)

To prove stability bounds with high probability, we first introduce a concentration inequality (Chernoff, 1952).

Lemma C.3 (Chernoff's Bound)**.** *Let $X_1,\ldots,X_t$ be independent random variables taking values in $\{0,1\}$. Let $X=\sum_{j=1}^t X_j$ and $\mu=\mathbb{E}[X]$. Then for any $\tilde{\delta}>0$ with probability at least $1-\exp\big(-\mu\tilde{\delta}^2/(2+\tilde{\delta})\big)$ we have $X\leq(1+\tilde{\delta})\mu$. Furthermore, for any $\delta\in(0,1)$ with probability at least $1-\delta$ we have*

$$X\leq\mu+\log(1/\delta)+\sqrt{2\mu\log(1/\delta)}.$$

We now prove Part (c). According to the analysis in Part (a), we know

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\leq\left(\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8G^2\eta_t^2\right)\mathbb{I}_{[i_t\neq n]}+\left((1+p)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8(1+1/p)\eta_t^2G^2\right)\mathbb{I}_{[i_t=n]}.$$

It then follows that

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\leq\big(1+p\mathbb{I}_{[i_t=n]}\big)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8G^2\eta_t^2\big(1+\mathbb{I}_{[i_t=n]}/p\big). \tag{C.4}$$

Applying this inequality recursively gives

$$\begin{aligned}\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2&\leq 8G^2\eta^2\sum_{k=1}^t\big(1+\mathbb{I}_{[i_k=n]}/p\big)\prod_{j=k+1}^t\big(1+p\mathbb{I}_{[i_j=n]}\big)\\&=8G^2\eta^2\sum_{k=1}^t\big(1+\mathbb{I}_{[i_k=n]}/p\big)\prod_{j=k+1}^t(1+p)^{\mathbb{I}_{[i_j=n]}}\\&\leq 8G^2\eta^2(1+p)^{\sum_{j=1}^t\mathbb{I}_{[i_j=n]}}\Big(t+\sum_{k=1}^t\mathbb{I}_{[i_k=n]}/p\Big).\end{aligned}$$

Applying Lemma C.3 with $X_j=\mathbb{I}_{[i_j=n]}$ and $\mu=t/n$ (note $\mathbb{E}_A[X_j]=1/n$), with probability $1-\delta$ there holds

$$\sum_{j=1}^t\mathbb{I}_{[i_j=n]}\leq t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}. \tag{C.5}$$

The following inequality then holds with probability at least $1-\delta$

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\leq 8G^2\eta^2(1+p)^{t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}}\Big(t+t/(pn)+p^{-1}\log(1/\delta)+p^{-1}\sqrt{2tn^{-1}\log(1/\delta)}\Big).$$

We can choose $p=\frac{1}{t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}}$ (note $(1+x)^{1/x}\leq e$) and derive the following inequality with probability at least $1-\delta$

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\leq 8eG^2\eta^2\Big(t+\big(t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}\big)^2\Big).$$

This finishes the proof of Part (c).

C.5. Proof of Part (d)

We now turn to Part (d). Under the smoothness assumption, the analysis in Part (b) implies

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\leq(1+L^2\eta_t^2)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2\mathbb{I}_{[i_t\neq n]}+\left((1+p)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8(1+1/p)\eta_t^2G^2\right)\mathbb{I}_{[i_t=n]}.$$

It then follows that

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq \big(1+L^2\eta_t^2+p\mathbb{I}_{[i_t=n]}\big)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2 + 8(1+1/p)\eta_t^2G^2\mathbb{I}_{[i_t=n]}.$$

We can apply the above inequality recursively and derive

$$\begin{aligned}\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 &\leq 8(1+1/p)G^2\sum_{k=1}^{t}\eta_k^2\mathbb{I}_{[i_k=n]}\prod_{j=k+1}^{t}\big(1+L^2\eta_j^2+p\mathbb{I}_{[i_j=n]}\big)\\
&\leq 8(1+1/p)G^2\eta^2\sum_{k=1}^{t}\mathbb{I}_{[i_k=n]}\prod_{j=k+1}^{t}\big(1+L^2\eta_j^2\big)\prod_{j=k+1}^{t}\big(1+p\mathbb{I}_{[i_j=n]}\big)\\
&= 8(1+1/p)G^2\eta^2\sum_{k=1}^{t}\mathbb{I}_{[i_k=n]}\prod_{j=k+1}^{t}\big(1+L^2\eta_j^2\big)\prod_{j=k+1}^{t}\big(1+p\big)^{\mathbb{I}_{[i_j=n]}}\\
&\leq 8(1+1/p)G^2\eta^2\prod_{j=1}^{t}\big(1+L^2\eta_j^2\big)\prod_{j=1}^{t}\big(1+p\big)^{\mathbb{I}_{[i_j=n]}}\sum_{k=1}^{t}\mathbb{I}_{[i_k=n]}.\end{aligned}$$

It then follows from the elementary inequality $1+x\leq e^x$ that

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq 8(1+1/p)G^2\eta^2\exp\Big(L^2\sum_{j=1}^{t}\eta_j^2\Big)\big(1+p\big)^{\sum_{j=1}^{t}\mathbb{I}_{[i_j=n]}}\sum_{k=1}^{t}\mathbb{I}_{[i_k=n]}$$

According to (C.5), we get the following inequality with probability at least $1-\delta$

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq 8(1+1/p)G^2\eta^2\exp\big(L^2t\eta^2\big)\big(1+p\big)^{t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}}\big(t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}\big).$$

We can choose $p=\frac{1}{t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}}$ and derive the following inequality with probability at least $1-\delta$

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq 8eG^2\eta^2\exp\big(L^2t\eta^2\big)\Big(1+t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}\Big)^2.$$

The stated bound then follows.

C.6. Proof of Part (e)

If $i_t\neq n$, we can analyze analogously to (C.2) excepting using the strong convexity, and show

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq (1-2\rho\eta_t)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2 + 8G^2\eta_t^2.$$

If $i_t=n$, then (C.3) holds. We can combine the above two cases and derive

$$\begin{aligned}&\mathbb{E}_{i_t}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right]\\
&\leq \frac{n-1}{n}\left((1-2\rho\eta_t)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8G^2\eta_t^2\right)+\frac{1+p}{n}\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+\frac{8(1+1/p)}{n}\eta_t^2G^2\\
&= (1-2\rho\eta_t+(2\rho\eta_t+p)/n)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8\eta_t^2G^2(1+1/(np)).\end{aligned}$$

We can choose $p=\rho\eta_t(n-2)$ to derive

$$\mathbb{E}_{i_t}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \leq (1-\rho\eta_t)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2+8\eta_t^2G^2\Big(1+\frac{1}{n(n-2)\rho\eta_t}\Big).$$

It then follows that

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \le 8G^2\sum_{j=1}^t \eta_j\Big(\eta_j+\frac{1}{n(n-2)\rho}\Big)\prod_{k=j+1}^t(1-\rho\eta_k).$$

For $\eta_t = 1/(\rho t)$, it follows from the identity $\prod_{k=j+1}^t(1-1/k)=j/t$ that

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \le \frac{8G^2}{t\rho}\sum_{j=1}^t\Big((\rho j)^{-1}+\frac{1}{n(n-2)\rho}\Big) \le \frac{8G^2}{\rho^2}\Big(\frac{\log(et)}{t}+\frac{1}{n(n-2)}\Big).$$

The stated result then follows from the Jensen's inequality.

D. Optimization Error Bounds: Convex-Concave Case

In this section, we present optimization error bounds for SGDA, which are standard in the literature (Nedić & Ozdaglar, 2009; Nemirovski et al., 2009). We give both bounds in expectation and bounds with high probability. The high-probability analysis requires to use concentration inequalities for martingales. Lemma D.1 is a Azuma-Hoeffding inequality for real-valued martingale difference sequence (Hoeffding, 1963), while Lemma D.2 is a Bernstein-type inequality for martingale difference sequences in a Hilbert space (Tarres & Yao, 2014).

Lemma D.1. *Let $\{\xi_k : k\in\mathbb{N}\}$ be a martingale difference sequence taking values in $\mathbb{R}$, i.e., $\mathbb{E}[\xi_k|\xi_1,\ldots,\xi_{k-1}]=0$. Assume that $|\xi_k-\mathbb{E}_{z_k}[\xi_k]|\le b_k$ for each k. For $\delta\in(0,1)$, with probability at least $1-\delta$ we have*

$$\sum_{k=1}^n \xi_k \le \Big(2\sum_{k=1}^n b_k^2\log\frac{1}{\delta}\Big)^{\frac{1}{2}}. \tag{D.1}$$

Lemma D.2. *Let $\{\xi_k : k\in\mathbb{N}\}$ be a martingale difference sequence in a Hilbert space with the norm $\|\cdot\|_2$. Suppose that almost surely $\|\xi_k\|\le B$ and $\sum_{k=1}^t\mathbb{E}[\|\xi_k\|^2|\xi_1,\ldots,\xi_{k-1}]\le\sigma_t^2$ for $\sigma_t\ge 0$. Then, for any $0<\delta<1$, the following inequality holds with probability at least $1-\delta$*

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

Lemma D.3. *Let $\{\mathbf{w}_t,\mathbf{v}_t\}$ be the sequence produced by (4.1) with $\eta_t=\eta$. Let Assumption 1 hold and F_S be convex-concave. Assume $\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{w}\|_2\le B_W$ and $\sup_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|_2\le B_V$. Then the following inequality holds*

$$\mathbb{E}_A\big[\sup_{\mathbf{v}\in\mathcal{V}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-\inf_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\le\eta G^2+\frac{B_W^2+B_V^2}{2\eta T}+\frac{G(B_W+B_V)}{\sqrt{T}}, \tag{D.2}$$

where $(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)$ is defined in (4.2). Let $\delta\in(0,1)$. Then with probability at least $1-\delta$ we have

$$\sup_{\mathbf{v}\in\mathcal{V}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-\inf_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\le\eta G^2+\frac{B_W^2+B_V^2}{2T\eta}+\frac{G\big(B_W+B_V\big)\big(9\log(6/\delta)+2\big)}{\sqrt{T}}. \tag{D.3}$$

Proof. According to the non-expansiveness of projection and (4.1), we know

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2\le\|\mathbf{w}_t-\eta_t\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\mathbf{w}\|_2^2\\&=\|\mathbf{w}_t-\mathbf{w}\|_2^2+\eta_t^2\|\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2+2\eta_t\langle\mathbf{w}-\mathbf{w}_t,\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\rangle\\&\le\|\mathbf{w}_t-\mathbf{w}\|_2^2+\eta_t^2G^2+2\eta_t\langle\mathbf{w}-\mathbf{w}_t,\nabla_{\mathbf{w}}F_S(\mathbf{w}_t;\mathbf{v}_t)\rangle+2\eta_t\langle\mathbf{w}-\mathbf{w}_t,\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\rangle,\end{aligned}$$

where we have used Assumption 1. According to the convexity of $F_S(\cdot,\mathbf{v}_t)$, we know

$$\begin{aligned}2\eta_t\big(F_S(\mathbf{w}_t,\mathbf{v}_t)-F_S(\mathbf{w},\mathbf{v}_t)\big)\le\|\mathbf{w}_t-\mathbf{w}\|_2^2-\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2+&\\ \eta_t^2G^2+2\eta_t\langle\mathbf{w}-\mathbf{w}_t,\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\rangle.&\end{aligned} \tag{D.4}$$

Taking a summation of the above inequality from $t = 1$ to $t = T$ ($\mathbf{w}_1 = 0$), we derive

$$2\eta \sum_{t=1}^{T} \big(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)\big) \le \|\mathbf{w}\|_2^2 + T\eta^2 G^2 + 2\eta \sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + 2\eta \sum_{t=1}^{T} \langle \mathbf{w}, \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle.$$

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

$$2\eta \sum_{t=1}^{T} \big(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big) \le B_W^2 + T\eta^2 G^2 + 2\eta \sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + 2\eta B_W \Big\| \sum_{t=1}^{T} \big(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2.$$

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

$$2\eta \sum_{t=1}^{T} \big(F_S(\mathbf{w}_t, \mathbf{v}_t) - \inf_{\mathbf{w}} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big) \le B_W^2 + T\eta^2 G^2 + 2\eta \sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle + 2\eta B_W \Big\| \sum_{t=1}^{T} \big(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2. \quad \text{(D.5)}$$

Note

$$\mathbb{E}_{i_t}\big[\langle \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle\big] = 0. \quad \text{(D.6)}$$

We can take an expectation over both sides of (D.5) and get

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_A\big[F_S(\mathbf{w}_t, \mathbf{v}_t)\big] - \mathbb{E}_A\big[\inf_{\mathbf{w}} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big] \le \frac{B_W^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{B_W}{T}\mathbb{E}_A\Big[\Big\| \sum_{t=1}^{T} \big(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2\Big].$$

According to Jensen's inequality and (D.6), we know

$$\Big(\mathbb{E}_A\Big[\Big\| \sum_{t=1}^{T} \big(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2\Big]\Big)^2 \le \mathbb{E}_A\Big[\Big\| \sum_{t=1}^{T} \big(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2^2\Big]$$
$$= \sum_{t=1}^{T} \mathbb{E}_A\Big[\Big\| \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) \Big\|_2^2\Big] \le TG^2.$$

It then follows that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_A\big[F_S(\mathbf{w}_t, \mathbf{v}_t)\big] - \mathbb{E}_A\big[\inf_{\mathbf{w}} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big] \le \frac{B_W^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{B_W G}{\sqrt{T}}. \quad \text{(D.7)}$$

In a similar way, we can show that

$$\mathbb{E}_A\big[\sup_{\mathbf{v}} F_S(\bar{\mathbf{w}}_T, \mathbf{v})\big] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_A\big[F_S(\mathbf{w}_t, \mathbf{v}_t)\big] \le \frac{B_V^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{B_V G}{\sqrt{T}}. \quad \text{(D.8)}$$

The stated bound (D.2) then follows from (D.7) and (D.8).

We now turn to (D.3). It is clear that $\big|\langle \mathbf{w}_t, \nabla F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle\big| \le 2GB_W$, and therefore we can apply Lemma D.1 to derive the following inequality with probability at least $1 - \delta/6$ that

$$\sum_{t=1}^{T} \langle \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})\rangle \le 2GB_W\Big(2T\log(6/\delta)\Big)^{\frac{1}{2}}. \quad \text{(D.9)}$$

For any $t \in \mathbb{N}$, define $\xi_t = \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)$. Then it is clear that $\|\xi_t\|_2 \le 2G$ and

$$\sum_{t=1}^{T} \mathbb{E}[\|\xi_t\|_2^2 | \xi_1, \ldots, \xi_{t-1}] \le 4TG^2.$$

Therefore, we can apply Lemma D.2 to derive the following inequality with probability at least $1 - \delta/3$

$$\Big\| \sum_{t=1}^{T} \xi_t \Big\|_2 \le 2\Big(\frac{2G}{3} + 2G\sqrt{T}\Big) \log(6/\delta).$$

Then, the following inequality holds with probability at least $1 - \delta/3$

$$\Big\| \sum_{t=1}^{T} \big(\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\big) \Big\|_2 \le 4G\Big(1 + \sqrt{T}\Big) \log(6/\delta).$$

We can plug the above inequality and (D.9) back into (D.5), and derive the following inequality with probability at least $1 - \delta/2$

$$\frac{1}{T}\sum_{t=1}^{T} F_S(\mathbf{w}_t, \mathbf{v}_t) - \inf_{\mathbf{w}} F_S(\mathbf{w}, \bar{\mathbf{v}}_T) \le \frac{B_W^2}{2T\eta} + \frac{\eta G^2}{2} + \frac{2GB_W\sqrt{2\log(6/\delta)}}{\sqrt{T}} + \frac{8B_W G \log(6/\delta)}{\sqrt{T}}.$$

In a similar way, we can get the following inequality with probability at least $1 - \delta/2$

$$\sup_{\mathbf{v}\in\mathcal{V}} F_S(\bar{\mathbf{w}}_T, \mathbf{v}) - \frac{1}{T}\sum_{t=1}^{T} F_S(\mathbf{w}_t, \mathbf{v}_t) \le \frac{B_V^2}{2T\eta} + \frac{\eta G^2}{2} + \frac{9B_V G \log(6/\delta) + 2B_V G}{\sqrt{T}}.$$

Combining the above two inequalities together we get the stated inequality with probability at least $1 - \delta$. The proof is complete. □

The following lemma gives optimization error bounds for SC-SC problems.

Lemma D.4. *Let Assumption 1 hold, $t_0 \ge 0$ and $F_S(\cdot,\cdot)$ be ρ-SC-SC with $\rho > 0$. Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be the sequence produced by (4.1) with $\eta_t = 1/(\rho(t+t_0))$. If $t_0 = 0$, then for $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ defined in (4.2) we have*

$$\mathbb{E}_A\big[\sup_{\mathbf{v}\in\mathcal{V}} F_S(\bar{\mathbf{w}}_T, \mathbf{v}) - \inf_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big] \le \frac{G^2\log(eT)}{\rho T} + \frac{(B_W + B_V)G}{\sqrt{T}}. \tag{D.10}$$

If $\sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_2 \le B_W$ and $\sup_{\mathbf{v}\in\mathcal{V}} \|\mathbf{v}\|_2 \le B_V$, then

$$\triangle_S^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le \frac{2\rho t_0(B_W^2 + B_V^2)}{T} + \frac{G^2\log(eT)}{\rho T}. \tag{D.11}$$

Proof. Analyzing analogous to (D.4) but using the strong convexity of $\mathbf{w} \mapsto F_S(\mathbf{w}, \mathbf{v})$, we derive

$$2\eta_t\big[F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)\big] \le (1 - \eta_t\rho)\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + \eta_t^2 G^2 + \xi_t(\mathbf{w}),$$

where $\xi_t(\mathbf{w}) = 2\eta_t\langle \mathbf{w} - \mathbf{w}_t, \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle$. Since $\eta_t = 1/(\rho(t+t_0))$, we further get

$$\frac{2}{\rho(t+t_0)}\big[F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)\big] \le (1 - 1/(t+t_0))\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + \frac{G^2}{\rho^2(t+t_0)^2} + \xi_t(\mathbf{w}).$$

Multiplying both sides by $t + t_0$ gives

$$\frac{2}{\rho}\big[F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)\big] \le (t + t_0 - 1)\|\mathbf{w}_t - \mathbf{w}\|_2^2 - (t+t_0)\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + (t+t_0)\xi_t(\mathbf{w}) + \frac{G^2}{\rho^2(t+t_0)}.$$

Taking a summation of the above inequality further gives

$$\sum_{t=1}^{T}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-F_S(\mathbf{w},\mathbf{v}_t)\big]\leq 2\rho t_0B_W^2+\frac{G^2\log(eT)}{2\rho}+\frac{\rho}{2}\sum_{t=1}^{T}(t+t_0)\xi_t(\mathbf{w}),$$

where we have used $\sum_{t=1}^{T}t^{-1}\leq\log(eT)$. This together with the concavity of $\mathbf{v}\mapsto F_S(\mathbf{w},\mathbf{v})$ gives

$$\sum_{t=1}^{T}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq 2\rho t_0B_W^2+\frac{G^2\log(eT)}{2\rho}+\frac{\rho}{2}\sum_{t=1}^{T}(t+t_0)\xi_t(\mathbf{w}). \tag{D.12}$$

Since the above inequality holds for any $\mathbf{w}$, we know

$$\sum_{t=1}^{T}\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq 2\rho t_0B_W^2+\frac{G^2\log(eT)}{2\rho}+\frac{\rho}{2}\sup_{\mathbf{w}\in\mathcal{W}}\sum_{t=1}^{T}(t+t_0)\xi_t(\mathbf{w}).$$

Since $\mathbb{E}_A[\langle\mathbf{w}_t,\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\rangle]=0$ we know

$$\begin{aligned}\mathbb{E}_A\Big[\sup_{\mathbf{w}\in\mathcal{W}}\sum_{t=1}^{T}(t+t_0)\xi_t(\mathbf{w})\Big]&=2\mathbb{E}_A\Big[\sup_{\mathbf{w}\in\mathcal{W}}\sum_{t=1}^{T}(t+t_0)\eta_t\langle\mathbf{w},\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\rangle\Big]\\&\leq 2\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{w}\|_2\mathbb{E}_A\Big\|\sum_{t=1}^{T}(t+t_0)\eta_t\big(\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\big)\Big\|_2\\&\leq 2B_W\left(\mathbb{E}_A\Big\|\sum_{t=1}^{T}(t+t_0)\eta_t\big(\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\big)\Big\|_2^2\right)^{1/2}\\&\leq 2B_W\left(\sum_{t=1}^{T}(t+t_0)^2\eta_t^2\mathbb{E}_A\|\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\|_2^2\right)^{1/2}\leq 2B_WG\rho^{-1}\sqrt{T}.\end{aligned}$$

We can combine the above two inequalities together and derive

$$\sum_{t=1}^{T}\mathbb{E}_A\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-\inf_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq 2\rho t_0B_W^2+\frac{G^2\log(eT)}{2\rho}+B_WG\sqrt{T}.$$

In a similar way one can show

$$\sum_{t=1}^{T}\mathbb{E}_A\big[\sup_{\mathbf{v}\in\mathcal{V}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-F_S(\mathbf{w}_t,\mathbf{v}_t)\big]\leq 2\rho t_0B_V^2+\frac{G^2\log(eT)}{2\rho}+B_VG\sqrt{T}.$$

We can combine the above two inequalities together, and get the following optimization error bounds

$$T\mathbb{E}_A\big[\sup_{\mathbf{v}\in\mathcal{V}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-\inf_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq 2\rho t_0(B_W^2+B_V^2)+\frac{G^2\log(eT)}{\rho}+(B_W+B_V)G\sqrt{T}.$$

This proves (D.10) with $t_0=0$.

We now turn to (D.11). Since $\mathbb{E}_A[\xi_t(\mathbf{w})]=0$, it follows from (D.12) that

$$\sum_{t=1}^{T}\mathbb{E}_A\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq 2\rho t_0B_W^2+\frac{G^2\log(eT)}{2\rho}.$$

In a similar way, one can show

$$\sum_{t=1}^{T}\mathbb{E}_A\big[F_S(\bar{\mathbf{w}}_T,\mathbf{v})-F_S(\mathbf{w}_t,\mathbf{v}_t)\big]\leq 2\rho t_0B_V^2+\frac{G^2\log(eT)}{2\rho}.$$

We can combine the above two inequalities together and derive

$$\mathbb{E}\big[F_S(\bar{\mathbf{w}}_T,\mathbf{v})-F_S(\mathbf{w},\bar{\mathbf{v}}_T\big]\leq\frac{2\rho t_0(B_W^2+B_V^2)}{T}+\frac{G^2\log(eT)}{\rho T}.$$

The stated bound (D.11) then follows by taking the supremum over $\mathbf{w}$ and $\mathbf{v}$. The proof is complete. □

E. Proofs on Generalization Bounds: Convex-Concave Case

In this section, we prove the generalization bounds of SGDA in a convex-concave case. We first prove Theorem 3 on bounds of weak PD population risks in expectation.

Proof of Theorem 3. We first prove Part (a). We have the decomposition

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = \triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) + \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T). \tag{E.1}$$

According to Part (a) of Theorem 2 we know the following inequality for all t

$$\mathbb{E}_A\Big[\Big\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\Big\|_2\Big] \le 4\eta G\Big(\sqrt{t} + \frac{t}{n}\Big).$$

It then follows from the convexity of a norm that

$$\mathbb{E}_A\Big[\Big\|\begin{pmatrix}\bar{\mathbf{w}}_{T} - \bar{\mathbf{w}}'_{T} \\ \bar{\mathbf{v}}_{T} - \bar{\mathbf{v}}'_{T}\end{pmatrix}\Big\|_2\Big] \le 4\eta G\Big(\sqrt{T} + \frac{T}{n}\Big)$$

and therefore

$$\begin{aligned}\sup_z\Big(\sup_{\mathbf{v}'\in\mathcal{V}}\mathbb{E}_A\big[f(\bar{\mathbf{w}}_T, \mathbf{v}'; z) - f(\bar{\mathbf{w}}'_T, \mathbf{v}'; z)\big] + \sup_{\mathbf{w}'\in\mathcal{W}}\mathbb{E}_A\big[f(\mathbf{w}', \bar{\mathbf{v}}_T; z) - f(\mathbf{w}', \bar{\mathbf{v}}'_T; z)\big]\Big)\\ \le G\Big(\mathbb{E}_A\big[\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\|_2\big] + \mathbb{E}_A\big[\|\bar{\mathbf{v}}_T - \bar{\mathbf{v}}'_T\|_2\big]\Big) \le 4\sqrt{2}\eta G^2\Big(\sqrt{T} + \frac{T}{n}\Big).\end{aligned}$$

According to Part (a) of Theorem 1, we know

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le 4\sqrt{2}\eta G^2\Big(\sqrt{T} + \frac{T}{n}\Big).$$

According to Eq. (D.2), we know

$$\triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le \eta G^2 + \frac{B_W^2 + B_V^2}{2\eta T} + \frac{G(B_W + B_V)}{\sqrt{T}}.$$

The bound (4.3) then follows directly from (E.1).

Eq. (4.4) in Part (b) can be proved in a similar way (e.g., by combining the stability bounds in Part (b) of Theorem 2 and optimization error bounds in Eq. (D.2) together). We omit the proof for brevity.

We now turn to Part (c). According to Part (e) of Theorem 2 and the convexity of norm, we know

$$\mathbb{E}_A\Big[\Big\|\begin{pmatrix}\bar{\mathbf{w}}_{T} - \bar{\mathbf{w}}'_{T} \\ \bar{\mathbf{v}}_{T} - \bar{\mathbf{v}}'_{T}\end{pmatrix}\Big\|_2\Big] \le \frac{2\sqrt{2}G}{\rho}\Big(\frac{\log^{\frac{1}{2}}(eT)}{\sqrt{T}} + \frac{1}{\sqrt{n(n-2)}}\Big).$$

Analyzing analogous to Part (a), we further know

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le \frac{4G^2}{\rho}\Big(\frac{\log^{\frac{1}{2}}(eT)}{\sqrt{T}} + \frac{1}{\sqrt{n(n-2)}}\Big).$$

This together with the optimization error bounds in Lemma D.4 and (E.1) gives

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le \frac{4G^2}{\rho}\Big(\frac{\log^{\frac{1}{2}}(eT)}{\sqrt{T}} + \frac{1}{\sqrt{n(n-2)}}\Big) + \frac{G^2\log(eT)}{\rho T} + \frac{(B_W + B_V)G}{\sqrt{T}}.$$

The stated bound then follows the choice of T. The proof is complete.

Finally, we consider Part (d). Since $t_0 \geq L^2/\rho^2$ we know $\eta_t = 1/(\rho(t+t_0)) \leq \rho/L^2$. The stability analysis in Farnia & Ozdaglar (2020)[2] then shows that A is ϵ-argument stable with $\epsilon = O(1/(\rho n))$. This together with Part (a) of Theorem 1 then shows that

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(1/(\rho n)).$$

We can combine the above generalization bound and the optimization error bound in (D.11) together, and get

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O(1/(\rho n)) + O\Big(\frac{\rho}{T} + \frac{\log(eT)}{\rho T}\Big).$$

The stated bound then follows from $T \asymp n$. The proof is complete. □

We now present proofs of Theorem 4 on primal population risks.

Proof of Theorem 4. We have the decomposition

$$R(\bar{\mathbf{w}}_T) - R(\mathbf{w}^*) = \big(R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T)\big) + \big(R_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big) \\ + \big(F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big) + \big(F(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \mathbf{v}^*)\big).$$

Since $F(\mathbf{w}^*, \bar{\mathbf{v}}_T) \leq F(\mathbf{w}^*, \mathbf{v}^*)$, it then follows that

$$R(\bar{\mathbf{w}}_T) - R(\mathbf{w}^*) \leq \big(R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T)\big) + \big(R_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big) + \big(F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big). \tag{E.2}$$

Taking an expectation on both sides gives

$$\mathbb{E}\big[R(\bar{\mathbf{w}}_T) - R(\mathbf{w}^*)\big] \leq \mathbb{E}\big[R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T)\big] + \mathbb{E}\big[R_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big] + \mathbb{E}\big[F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big]. \tag{E.3}$$

Note that the first and the third term on the right-hand side is related to generalization, while the second term $R_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T)$ is related to optimization. According to Part (b) of Theorem 2 we know the following inequality for all t

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2\right] \leq \frac{G\sqrt{8e(t+t^2/n)}}{\sqrt{n}} \exp\Big(L^2 t\eta^2/2\Big)\eta.$$

It then follows from the convexity of a norm that

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T \\ \bar{\mathbf{v}}_T - \bar{\mathbf{v}}'_T\end{pmatrix}\right\|_2\right] \leq \frac{G\sqrt{8e(T+T^2/n)}}{\sqrt{n}} \exp\Big(L^2 T\eta^2/2\Big)\eta. \tag{E.4}$$

This together with Part (b) of Theorem 1 implies that

$$\mathbb{E}_{S,A}\Big[R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T)\Big] \leq \frac{(1+L/\rho)G^2\eta\sqrt{8e(T+T^2/n)}\exp\big(L^2T\eta^2/2\big)}{\sqrt{n}}.$$

Similarly, the stability bound (E.4) also implies the following bound on the gap between population and empirical risk

$$\mathbb{E}_{S,A}\Big[F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \bar{\mathbf{v}}_T)\Big] \leq \frac{(1+L/\rho)G^2\eta\sqrt{8e(T+T^2/n)}\exp\big(L^2T\eta^2/2\big)}{\sqrt{n}}.$$

According to Lemma D.3, we know

$$\mathbb{E}_A\big[R_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T)\big] \leq \mathbb{E}_A\big[\sup_{\mathbf{v}\in\mathcal{V}} F_S(\bar{\mathbf{w}}_T, \mathbf{v}) - \inf_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w}, \bar{\mathbf{v}}_T)\big] \leq \eta G^2 + \frac{B_W^2 + B_V^2}{2\eta T} + \frac{G(B_W + B_V)}{\sqrt{T}}.$$

We can plug the above three inequalities back into (E.3), and derive the stated bound on the excess primal population risk in expectation.

[2] Farnia & Ozdaglar (2020) considered the constant step size $\eta_t = \eta \leq \rho/L^2$. It is direct to extend the analysis there to any step size $\eta_t \leq \rho/L^2$ since an algorithm would be more stable if the step size decreases.

We now turn to the high-probability bounds. According to Assumption 1 and Part (d) of Theorem 2, we know that with probability at least $1-\delta/4$ that SGDA is ϵ-uniformly stable, where ϵ satisfies

$$\epsilon = O\Big(\eta \exp(L^2T\eta^2/2)\big(Tn^{-1}+\log(1/\delta)\big)\Big). \tag{E.5}$$

This together with Part (d) of Theorem 1 implies the following inequality with probability at least $1-\delta/2$

$$R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T) = O\Big(L\rho^{-1}\epsilon \log n \log(1/\delta) + n^{-\frac{1}{2}}\sqrt{\log(1/\delta)}\Big),$$

where ϵ satisfies (E.5). In a similar way, one can use Part (d) of Theorem 1 and stability bounds in Part (d) of Theorem 2 to show the following inequality with probability at least $1-\delta/4$

$$F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \bar{\mathbf{v}}_T) = O\Big(\log n \log(1/\delta)\epsilon\Big) + O(n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)). \tag{E.6}$$

According to (D.3), we derive the following inequality with probability at least $1-\delta/4$

$$R_S(\bar{\mathbf{w}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) = \sup_{\mathbf{v}\in\mathcal{V}} F_S(\bar{\mathbf{w}}_T, \mathbf{v}) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) = O\Big(\eta + (T\eta)^{-1} + T^{-\frac{1}{2}}\log(1/\delta)\Big).$$

We can plug the above three inequalities back into (E.2) and derive the following inequality with probability at least $1-\delta$

$$R(\bar{\mathbf{w}}_T) - R(\mathbf{w}^*) = O\Big(L\rho^{-1}\eta \exp(L^2T\eta^2/2)\log n \log(1/\delta)\big(Tn^{-1}+\log(1/\delta)\big)\Big) + O(n^{-\frac{1}{2}}\sqrt{\log(1/\delta)}) + O\Big(\eta + (T\eta)^{-1} + T^{-\frac{1}{2}}\log(1/\delta)\Big). \tag{E.7}$$

The high-probability bound (4.6) then follows from the choice of T and η. The proof is complete. □

Finally, we present high-probability bounds of plain generalization errors for SGDA.

Theorem E.1 (High-probability bounds)**.** *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be the sequence produced by (4.1) with $\eta_t = \eta$. Assume for all z, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is convex-concave. Let A be defined by $A_{\mathbf{w}}(S) = \bar{\mathbf{w}}_T$ and $A_{\mathbf{v}}(S) = \bar{\mathbf{v}}_T$ for $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ in (4.2). Let $\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{w}\|_2 \le B_W, \sup_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|_2 \le B_V$ and $\delta\in(0,1)$. Let $\widetilde{\triangle}_T = \big|F(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \mathbf{v}^*)\big|$.*

(a) If Assumption 1 holds, then with probability at least $1-\delta$

$$\widetilde{\triangle}_T = O\Big(\eta \log n \log(1/\delta)\big(\sqrt{T} + Tn^{-1} + \log(1/\delta)\big)\Big) + O(n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)) + O\Big((T\eta)^{-1} + T^{-\frac{1}{2}}\log(1/\delta)\Big).$$

If we choose $T \asymp n^2$ and $\eta \asymp T^{-3/4}$ then we get the following inequality with probability at least $1-\delta$

$$\widetilde{\triangle}_T = O(n^{-1/2}\log n \log^2(1/\delta)). \tag{E.8}$$

(b) If Assumptions 1, 2 hold, then the following inequality holds with probability at least $1-\delta$

$$\widetilde{\triangle}_T = O\Big(\eta \log n \log(1/\delta)\exp\big(L^2T\eta^2/2\big)\big(Tn^{-1}+\log(1/\delta)\big) + n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta) + (T\eta)^{-1} + T^{-\frac{1}{2}}\log(1/\delta)\Big).$$

In particular, we can choose $T \asymp n$ and $\eta \asymp T^{-1/2}$ to derive (E.8) with probability at least $1-\delta$.

Proof. We use the error decomposition

$$F(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \mathbf{v}^*) = F(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - F_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) + F_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) + F_S(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \bar{\mathbf{v}}_T) + F(\mathbf{w}^*, \bar{\mathbf{v}}_T) - F(\mathbf{w}^*, \mathbf{v}^*). \tag{E.9}$$

We first prove Part (a). According to Assumption 1 and Part (c) of Theorem 2, we know that SGDA is ϵ-uniformly stable with probability at least $1-\delta/4$, where

$$\epsilon = O\Big(\eta\big(\sqrt{T} + Tn^{-1} + \log(1/\delta)\big)\Big).$$

This together with Part (e) of Theorem 1 implies the following inequality with probability at least $1-\delta/2$

$$F(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)-F_S(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)=O\Big(\eta\log n\log(1/\delta)\big(\sqrt{T}+Tn^{-1}+\log(1/\delta)\big)\Big)+O(n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)). \tag{E.10}$$

Similarly, the following inequality holds with probability at least $1-\delta/4$

$$F_S(\mathbf{w}^*,\bar{\mathbf{v}}_T)-F(\mathbf{w}^*,\bar{\mathbf{v}}_T)=O\Big(\eta\log n\log(1/\delta)\big(\sqrt{T}+Tn^{-1}+\log(1/\delta)\big)\Big)+O(n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta)). \tag{E.11}$$

According to Lemma D.3, the following inequality holds with probability at least $1-\delta/4$

$$F_S(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)-F_S(\mathbf{w}^*,\bar{\mathbf{v}}_T)\le\sup_{\mathbf{v}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-\inf_{\mathbf{w}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)=O\Big(\eta+(T\eta)^{-1}+T^{-\frac{1}{2}}\log(1/\delta)\Big). \tag{E.12}$$

According to the definition of $(\mathbf{w}^*,\mathbf{v}^*)$, we know $F(\mathbf{w}^*,\bar{\mathbf{v}}_T)\le F(\mathbf{w}^*,\mathbf{v}^*)$. We can plug this inequality and (E.10), (E.11), (E.12) back into (E.9), and derive the following inequality with probability at least $1-\delta/2$

$$\begin{aligned}F(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)-F(\mathbf{w}^*,\mathbf{v}^*)&=O\Big(\eta\log n\log(1/\delta)\big(\sqrt{T}+Tn^{-1}+\log(1/\delta)\big)\Big)\\&\quad+O(n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta))+O\Big((T\eta)^{-1}+T^{-\frac{1}{2}}\log(1/\delta)\Big).\end{aligned}$$

Analyzing in a similar way but using the error decomposition

$$\begin{aligned}F(\mathbf{w}^*,\mathbf{v}^*)-F(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)&=F(\mathbf{w}^*,\mathbf{v}^*)-F(\bar{\mathbf{w}}_T,\mathbf{v}^*)+F(\bar{\mathbf{w}}_T,\mathbf{v}^*)-F_S(\bar{\mathbf{w}}_T,\mathbf{v}^*)\\&\quad+F_S(\bar{\mathbf{w}}_T,\mathbf{v}^*)-F_S(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)+F_S(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)-F(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T),\end{aligned}$$

one can derive the following inequality with probability at least $1-\delta/2$

$$\begin{aligned}F(\mathbf{w}^*,\mathbf{v}^*)-F(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)&=O\Big(\eta\log n\log(1/\delta)\big(\sqrt{T}+Tn^{-1}+\log(1/\delta)\big)\Big)\\&\quad+O(n^{-\frac{1}{2}}\log^{\frac{1}{2}}(1/\delta))+O\Big((T\eta)^{-1}+T^{-\frac{1}{2}}\log(1/\delta)\Big).\end{aligned}$$

The stated bound then follows as a combination of the above two inequalities.

Part (b) can be derived similarly excepting using the stability bounds in Part (d) of Theorem 2. We omit the proof for brevity. The proof is complete. ☐

F. Stability and Generalization Bounds of SGDA on Non-Convex Objectives

F.1. Proof of Theorem 5

In this section, we show the stability and generalization bounds of SGDA for weakly-convex-weakly-concave objectives. We first introduce some lemmas. As an extension of a lemma in Hardt et al. (2016), the next lemma is motivated by the fact that SGDA typically runs several iterations before encountering the different example between S and S'.

Lemma F.1. *Assume $|f(\cdot,\cdot,z)|\le 1$ for any z and let Assumption 1 hold. Let $S=\{z_1,\ldots,z_n\}$ and $S'=\{z_1,\ldots,z_{n-1},z'_n\}$. Let $\{\mathbf{w}_t,\mathbf{v}_t\}$ and $\{\mathbf{w}'_t,\mathbf{v}'_t\}$ be the sequence produced by (4.1) w.r.t. S and S', respectively. Denote*

$$\Delta_t=\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2. \tag{F.1}$$

Then for any $t_0\in\mathbb{N}$ and any $\mathbf{w}',\mathbf{v}'$ we have

$$\mathbb{E}[f(\mathbf{w}_T,\mathbf{v}';z)-f(\mathbf{w}'_T,\mathbf{v}';z)+f(\mathbf{w}',\mathbf{v}_T;z)-f(\mathbf{w}',\mathbf{v}'_T;z)]\le\frac{4t_0}{n}+\sqrt{2}G\mathbb{E}[\Delta_T|\Delta_{t_0}=0].$$

Proof. According to Assumption 1, we know

$$f(\mathbf{w}_T,\mathbf{v}';z)-f(\mathbf{w}'_T,\mathbf{v}';z)+f(\mathbf{w}',\mathbf{v}_T;z)-f(\mathbf{w}',\mathbf{v}'_T;z)\leq G\sqrt{2}\Delta_T. \tag{F.2}$$

Let $\mathcal{E}$ denote the event that $\Delta_{t_0}=0$. Then we have

$$\begin{aligned}&\mathbb{E}[f(\mathbf{w}_T,\mathbf{v}';z)-f(\mathbf{w}'_T,\mathbf{v}';z)+f(\mathbf{w}',\mathbf{v}_T;z)-f(\mathbf{w}',\mathbf{v}'_T;z)]\\=&\mathbb{P}[\mathcal{E}]\mathbb{E}[f(\mathbf{w}_T,\mathbf{v}';z)-f(\mathbf{w}'_T,\mathbf{v}';z)+f(\mathbf{w}',\mathbf{v}_T;z)-f(\mathbf{w}',\mathbf{v}'_T;z)|\mathcal{E}]\\&+\mathbb{P}[\mathcal{E}^c]\mathbb{E}[f(\mathbf{w}_T,\mathbf{v}';z)-f(\mathbf{w}'_T,\mathbf{v}';z)+f(\mathbf{w}',\mathbf{v}_T;z)-f(\mathbf{w}',\mathbf{v}'_T;z)|\mathcal{E}^c]\\\leq&\sqrt{2}G\mathbb{E}[\Delta_T|\mathcal{E}]+4\mathbb{P}[\mathcal{E}^c],\end{aligned}$$

where in the last step we have used (F.2) and the condition $|f(\cdot,\cdot,z)|\leq 1$. Using the union bound on the outcome $i_t=n$ we obtain that

$$\mathbb{P}[\mathcal{E}^c]\leq\sum_{t=1}^{t_0}\mathbb{P}[i_t=n]=\frac{t_0}{n}.$$

The proof is complete by combining the above two inequalities together. □

Lemma F.2 shows the monotonity of the gradient for weakly-convex-weakly-concave functions. Its proof is well known in the literature (Liu et al., 2020; Rockafellar, 1976).

Lemma F.2. *Let f be a ρ-weakly-convex-weakly-concave function. Then*

$$\Big\langle\begin{pmatrix}\mathbf{w}-\mathbf{w}'\\\mathbf{v}-\mathbf{v}'\end{pmatrix},\begin{pmatrix}\nabla_{\mathbf{w}}f(\mathbf{w},\mathbf{v})-\nabla_{\mathbf{w}}f(\mathbf{w}',\mathbf{v}')\\\nabla_{\mathbf{v}}f(\mathbf{w}',\mathbf{v}')-\nabla_{\mathbf{v}}f(\mathbf{w},\mathbf{v})\end{pmatrix}\Big\rangle\geq-\rho\Big\|\begin{pmatrix}\mathbf{w}-\mathbf{w}'\\\mathbf{v}-\mathbf{v}'\end{pmatrix}\Big\|_2^2. \tag{F.3}$$

We are now ready to prove Theorem 5.

Proof of Theorem 5. Note that the projection step is nonexpansive. We consider two cases at the t-th iteration. If $i_t\neq n$, then it follows from Lemma F.2 and the Lipschitz continuity of f that

$$\begin{aligned}&\Big\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\Big\|_2^2\leq\Big\|\begin{pmatrix}\mathbf{w}_t-\eta_t\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\mathbf{w}'_t+\eta_t\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})\\\mathbf{v}_t+\eta_t\nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\mathbf{v}'_t-\eta_t\nabla_{\mathbf{v}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})\end{pmatrix}\Big\|_2^2=\Big\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\Big\|_2^2\\&+\eta_t^2\Big\|\begin{pmatrix}\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})-\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\\\nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{v}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})\end{pmatrix}\Big\|_2^2-2\eta_t\Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix},\begin{pmatrix}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})-\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})\\\nabla_{\mathbf{v}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})-\nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\end{pmatrix}\Big\rangle\\\leq&(1+2\eta_t\rho)\Big\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\Big\|_2^2+8G^2\eta_t^2.\end{aligned} \tag{F.4}$$

If $i_t=n$, then it follows from the elementary inequality $(a+b)^2\leq(1+p)a^2+(1+1/p)b^2$ that

$$\begin{aligned}\Big\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\Big\|_2^2&\leq\Big\|\begin{pmatrix}\mathbf{w}_t-\eta_t\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_n)-\mathbf{w}'_t+\eta_t\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t;z'_n)\\\mathbf{v}_t+\eta_t\nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_n)-\mathbf{v}'_t-\eta_t\nabla_{\mathbf{v}}f(\mathbf{w}'_t,\mathbf{v}'_t;z'_n)\end{pmatrix}\Big\|_2^2\\&\leq(1+p)\Big\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\Big\|_2^2+(1+1/p)\eta_t^2\Big\|\begin{pmatrix}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_n)-\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t;z'_n)\\\nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_n)-\nabla_{\mathbf{v}}f(\mathbf{w}'_t,\mathbf{v}'_t;z'_n)\end{pmatrix}\Big\|_2^2.\end{aligned} \tag{F.5}$$

Note that the event $i_t\neq n$ happens with probability $1-1/n$ and the event $i_t=n$ happens with probability $1/n$. Therefore, we know

$$\begin{aligned}\mathbb{E}_{i_t}\Big[\Big\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\Big\|_2^2\Big]&\leq\frac{n-1}{n}\Big((1+2\eta_t\rho)\Big\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\Big\|_2^2+8G^2\eta_t^2\Big)+\frac{1+p}{n}\Big\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\Big\|_2^2+\frac{8(1+1/p)}{n}\eta_t^2G^2\\&\leq(1+2\eta_t\rho+p/n)\Big\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\\mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\Big\|_2^2+8\eta_t^2G^2(1+1/(np)).\end{aligned}$$

Let $t_0 \in \mathbb{N}$ and $\mathcal{E}$ be defined as in the proof of Lemma F.1. We apply the above equation recursively from $t = t_0 + 1$ to T, then

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_T - \mathbf{w}'_T \\ \mathbf{v}_T - \mathbf{v}'_T\end{pmatrix}\right\|_2^2 \,\middle|\, \mathcal{E}\right] \leq 8G^2\big(1 + 1/(np)\big)\sum_{t=t_0+1}^{T}\eta_t^2 \prod_{k=t+1}^{T}\big(1 + 2\eta_k\rho + p/n\big).$$

By the elementary inequality $1 + a \leq \exp(a)$ and $\eta_t = \frac{c}{t}$, we further derive

$$\begin{aligned}\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \,\middle|\, \mathcal{E}\right] &\leq 8G^2\big(1 + 1/(np)\big)\sum_{t=t_0+1}^{T}\frac{c^2}{t^2}\prod_{k=t+1}^{T}\exp\Big(\frac{2c\rho}{k} + \frac{p}{n}\Big) \\ &\leq 8G^2\big(1 + 1/(np)\big)\sum_{t=t_0+1}^{T}\frac{c^2}{t^2}\exp\Big(\sum_{k=t+1}^{T}\frac{2c\rho}{k} + \frac{pT}{n}\Big).\end{aligned}$$

By taking $p = n/T$ in the above inequality, we further derive

$$\begin{aligned}\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \,\middle|\, \mathcal{E}\right] &\leq 8eG^2\Big(1 + \frac{T}{n^2}\Big)\sum_{t=t_0+1}^{T}\frac{c^2}{t^2}\exp\Big(\sum_{k=t+1}^{T}\frac{2c\rho}{k}\Big) \\ &\leq 8eG^2\Big(1 + \frac{T}{n^2}\Big)\sum_{t=t_0+1}^{T}\frac{c^2}{t^2}\exp\Big(2c\rho\log\Big(\frac{T}{t}\Big)\Big) \\ &\leq 8c^2eG^2\Big(1 + \frac{T}{n^2}\Big)T^{2c\rho}\sum_{t=t_0+1}^{T}\frac{1}{t^{2c\rho+2}} \\ &\leq \frac{8c^2eG^2}{2c\rho + 1}\Big(1 + \frac{T}{n^2}\Big)\Big(\frac{T}{t_0}\Big)^{2c\rho}\frac{1}{t_0}.\end{aligned}$$

Combining the above inequality and Lemma F.1 together, we obtain

$$\mathbb{E}_A[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}'_T, \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}'_T; z)] \leq \frac{4t_0}{n} + \frac{4\sqrt{e}cG^2}{\sqrt{2c\rho + 1}}\Big(1 + \frac{\sqrt{T}}{n}\Big)\Big(\frac{T}{t_0}\Big)^{c\rho}\frac{1}{\sqrt{t_0}}. \tag{F.6}$$

The right hand side is approximately minimized when

$$t_0 = \Big(\frac{\sqrt{e}cG^2}{\sqrt{2c\rho + 1}}\Big(1 + \frac{\sqrt{T}}{n}\Big)T^{c\rho}n\Big)^{\frac{2}{2c\rho+3}}.$$

Plugging it into the Eq. (F.6) we have (for simplicity we assume the above t_0 is an integer)

$$\mathbb{E}_A[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}'_T, \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}'_T; z)] \leq 8\Big(\frac{\sqrt{e}cG^2}{\sqrt{2c\rho + 1}}\Big(1 + \frac{\sqrt{T}}{n}\Big)T^{c\rho}\Big)^{\frac{2}{2c\rho+3}}\Big(\frac{1}{n}\Big)^{\frac{2c\rho+1}{2c\rho+3}}.$$

Since the above bound holds for all z, S, S' and $\mathbf{w}', \mathbf{v}'$, we immediately get the same upper bound on the weak stability. Finally the theorem holds by calling Theorem 1, Part (a). □

F.2. High-Probability Stability and Generalization Bounds

In this section, we give stability and generalization bounds of SGDA with nonconvex-nonconcave smooth objectives in high probability. The analysis requires a tail bound for a linear combination of independent Bernoulli random variables (Raghavan, 1988).

Lemma F.3. *Let $c_t \in (0, 1]$ and let $X_1, \cdots, X_T$ be independent Bernoulli random variables with the success rate of X_t being $p_t \in [0, 1]$. Denote $s = \sum_{t=1}^{T} c_t p_t$. Then, for all $a > 0$,*

$$\mathbb{P}\Big[\sum_{t=1}^{T} c_t X_t \geq (1 + a)s\Big] \leq \Big(\frac{e^a}{(1 + a)^{(1+a)}}\Big)^s.$$

In particular, for all $\delta \in (0,1)$ *such that* $\log(1/\delta) < s$ *with probability at least* $1-\delta$ *we have*

$$\sum_{t=1}^{T} c_t X_t \leq s + (e-1)\sqrt{\log(1/\delta)s}.$$

Theorem F.4. *Let* $\{\mathbf{w}_t, \mathbf{v}_t\}$ *be the sequence produced by* (4.1) *with* $\eta_t \leq \frac{c}{t}$ *for some* $c > 0$. *Assume Assumption 1, 2 hold and* $|f(\cdot,\cdot;z)| \leq 1$. *For any* $\delta \in (0,1)$, *if* $c \leq \frac{1}{(n\log(2/\delta)-1)L}$, *then with probability at least* $1-\delta$ *we have*

$$\big|F(\mathbf{w}_T,\mathbf{v}_T) - F_S(\mathbf{w}_T,\mathbf{v}_T)\big| = O\Big(T^{cL}\log(n)\log^{3/2}(1/\delta)n^{-1/2} + n^{-1/2}\log^{1/2}(1/\delta)\Big).$$

Proof. Let $S' = \{z_1,\ldots,z_{n-1},z_n'\}$ and $\{\mathbf{w}_t',\mathbf{v}_t'\}$ be the sequence produced by (4.1) w.r.t. S'. If $i_t \neq n$, it follows from the L-smoothness of f that

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}_{t+1}'\\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}'\end{pmatrix}\right\|_2 \leq \left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}_{t}'\\ \mathbf{v}_{t}-\mathbf{v}_{t}'\end{pmatrix}\right\|_2 + \eta_t\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}_t',\mathbf{v}_t';z_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t',\mathbf{v}_t';z_{i_t})\end{pmatrix}\right\|_2 \leq (1+L\eta_t)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}_{t}'\\ \mathbf{v}_{t}-\mathbf{v}_{t}'\end{pmatrix}\right\|_2.$$

If $i_t = n$, we have

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}_{t+1}'\\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}'\end{pmatrix}\right\|_2 \leq \left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}_{t}'\\ \mathbf{v}_{t}-\mathbf{v}_{t}'\end{pmatrix}\right\|_2 + 4\eta_t G.$$

We can combine the above two inequalities together and get

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}_{t+1}'\\ \mathbf{v}_{t+1}-\mathbf{v}_{t+1}'\end{pmatrix}\right\|_2 \leq (1+L\eta_t)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}_{t}'\\ \mathbf{v}_{t}-\mathbf{v}_{t}'\end{pmatrix}\right\|_2 + 4G\eta_t\mathbb{I}_{[i_t=n]}.$$

We apply the above inequality recursively from $t=1$ to T and get

$$\left\|\begin{pmatrix}\mathbf{w}_{T}-\mathbf{w}_{T}'\\ \mathbf{v}_{T}-\mathbf{v}_{T}'\end{pmatrix}\right\|_2 \leq 4G\sum_{t=1}^{T}\eta_t\mathbb{I}_{[i_t=n]}\prod_{k=t+1}^{T}\Big(1+L\eta_k\Big).$$

By the elementary inequality $1+a \leq \exp(a)$ and $\eta_t \leq \frac{c}{t}$, we further derive

$$\begin{aligned}\left\|\begin{pmatrix}\mathbf{w}_{T}-\mathbf{w}_{T}'\\ \mathbf{v}_{T}-\mathbf{v}_{T}'\end{pmatrix}\right\|_2 &\leq 4cG\sum_{t=1}^{T}\frac{\mathbb{I}_{[i_t=n]}}{t}\prod_{k=t+1}^{T}\exp\Big(\frac{cL}{k}\Big) = 4cG\sum_{t=1}^{T}\frac{\mathbb{I}_{[i_t=n]}}{t}\exp\Big(\sum_{k=t+1}^{T}\frac{cL}{k}\Big)\\ &\leq 4cG\sum_{t=1}^{T}\frac{\mathbb{I}_{[i_t=n]}}{t}\exp\Big(cL\log\Big(\frac{T}{t}\Big)\Big) \leq 4cGT^{cL}\sum_{t=1}^{T}\frac{\mathbb{I}_{[i_t=n]}}{t^{cL+1}}.\end{aligned}$$

By Lemma F.3, for any $\delta > 0$ such that $\log(2/\delta) < \sum_{t=1}^{T}\frac{1}{t^{cL+1}n}$, with probability at least $1-\delta/2$ we have

$$\left\|\begin{pmatrix}\mathbf{w}_{T}-\mathbf{w}_{T}'\\ \mathbf{v}_{T}-\mathbf{v}_{T}'\end{pmatrix}\right\|_2 \leq 4cGT^{cL}\Big(\sum_{t=1}^{T}\frac{1}{t^{cL+1}n} + (e-1)\sqrt{\log(1/\delta)\sum_{t=1}^{T}\frac{1}{t^{cL+1}n}}\Big). \tag{F.7}$$

Note that

$$\sum_{t=1}^{T}\frac{1}{t^{cL+1}} \leq 1 + \int_{t=1}^{T}\frac{\mathrm{d}t}{t^{cL+1}} \leq 1 + \frac{1}{cL}.$$

Plugging the above bound into Equation (F.7) , we know with probability at least $1-\delta/2$

$$\left\|\begin{pmatrix}\mathbf{w}_{T}-\mathbf{w}_{T}'\\ \mathbf{v}_{T}-\mathbf{v}_{T}'\end{pmatrix}\right\|_2 \leq 4cGT^{cL}\Big(\frac{cL+1}{cLn} + (e-1)\sqrt{\frac{(cL+1)\log(1/\delta)}{cLn}}\Big).$$

By the Lipschitz continuity of f, the above equation implies SGDA is ϵ-uniformly stable with probability at least $1-\delta/2$ and

$$\epsilon = O\Big(T^{cL}\sqrt{\log(1/\delta)}n^{-\frac{1}{2}}\Big).$$

This together with Part (e) of Theorem 1 implies the following inequality with probability at least $1-\delta$

$$\big|F(\mathbf{w},\mathbf{v}) - F_S(\mathbf{w}_T,\mathbf{v}_T)\big| = O\Big(T^{cL}\log(n)\log^{3/2}(1/\delta)n^{-1/2} + n^{-1/2}\log^{1/2}(1/\delta)\Big).$$

The proof is complete. □

F.3. Proof of Theorem 6

In this section, we prove Theorem 6 on generalization bounds under a regularity condition on the decay of weak-convexity-weak-concavity parameter along the optimization process.

Proof of Theorem 6. Let $S=\{z_1,\ldots,z_n\}$ and $S'=\{z'_1,\ldots,z'_n\}$ be two neighboring datasets. Without loss of generality, we assume $z_i = z'_i$ for $i\in[n-1]$. If $i_t\neq n$, then it follows from Assumption 2 that

$$\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t})\end{pmatrix}\right\|_2^2 \leq L^2\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2$$

If $i_t=n$, then it follows from Assumption 1 that

$$\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\end{pmatrix}\right\|_2^2 \leq 8G^2.$$

Therefore, we have

$$\mathbb{E}_{i_t}\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\end{pmatrix}\right\|_2^2 \leq \frac{(n-1)L^2}{n}\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{8G^2}{n}. \tag{F.8}$$

According to (4.1), we know

$$\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 \leq \left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \eta_t^2\left\|\begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\end{pmatrix}\right\|_2^2$$
$$- 2\eta_t\Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\end{pmatrix}\Big\rangle.$$

Taking a conditional expectation w.r.t. i_t gives

$$\begin{aligned}
&\mathbb{E}_{i_t}\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\\
&\leq \left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + L^2\eta_t^2\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{8G^2\eta_t^2}{n} - 2\eta_t\mathbb{E}_{i_t}\Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t})\\ \nabla_{\mathbf{v}} f(\mathbf{w}'_t,\mathbf{v}'_t;z'_{i_t}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t})\end{pmatrix}\Big\rangle\\
&= \left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + L^2\eta_t^2\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{8G^2\eta_t^2}{n} - 2\eta_t\Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} F_S(\mathbf{w}_t,\mathbf{v}_t) - \nabla_{\mathbf{w}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t)\\ \nabla_{\mathbf{v}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{v}} F_S(\mathbf{w}_t,\mathbf{v}_t)\end{pmatrix}\Big\rangle,
\end{aligned}$$

where we have used (F.8) in the first step and used the fact

$$\mathbb{E}_{i_t}\nabla f(\mathbf{w},\mathbf{v},z_{i_t}) = \nabla F_S(\mathbf{w},\mathbf{v}),\quad \mathbb{E}_{i_t}\nabla f(\mathbf{w},\mathbf{v},z'_{i_t}) = \nabla F_{S'}(\mathbf{w},\mathbf{v})$$

in the second step. According to (5.1), we know

$$\begin{aligned}
&\Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} F_S(\mathbf{w}_t,\mathbf{v}_t) - \nabla_{\mathbf{w}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t)\\ \nabla_{\mathbf{v}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{v}} F_S(\mathbf{w}_t,\mathbf{v}_t)\end{pmatrix}\Big\rangle\\
&= \Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} F_S(\mathbf{w}_t,\mathbf{v}_t) - \nabla_{\mathbf{w}} F_S(\mathbf{w}'_t,\mathbf{v}'_t)\\ \nabla_{\mathbf{v}} F_S(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{v}} F_S(\mathbf{w}_t,\mathbf{v}_t)\end{pmatrix}\Big\rangle + \Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} F_S(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{w}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t)\\ \nabla_{\mathbf{v}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{v}} F_S(\mathbf{w}'_t,\mathbf{v}'_t)\end{pmatrix}\Big\rangle\\
&\geq -\rho_t\left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \Big\langle\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}, \begin{pmatrix}\nabla_{\mathbf{w}} F_S(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{w}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t)\\ \nabla_{\mathbf{v}} F_{S'}(\mathbf{w}'_t,\mathbf{v}'_t) - \nabla_{\mathbf{v}} F_S(\mathbf{w}'_t,\mathbf{v}'_t)\end{pmatrix}\Big\rangle.
\end{aligned}$$

It follows from Assumption 1 that

$$\Big\langle \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} F_S(\mathbf{w}_t', \mathbf{v}_t') - \nabla_{\mathbf{w}} F_{S'}(\mathbf{w}_t', \mathbf{v}_t') \\ \nabla_{\mathbf{v}} F_{S'}(\mathbf{w}_t', \mathbf{v}_t') - \nabla_{\mathbf{v}} F_S(\mathbf{w}_t', \mathbf{v}_t') \end{pmatrix} \Big\rangle = \frac{1}{n} \Big\langle \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n) - \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') - \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n) \end{pmatrix} \Big\rangle$$

$$\geq -\frac{1}{n} \Big\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \Big\|_2 \Big\| \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n) - \nabla_{\mathbf{w}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') \\ \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n') - \nabla_{\mathbf{v}} f(\mathbf{w}_t', \mathbf{v}_t'; z_n) \end{pmatrix} \Big\|_2 \geq -\frac{2\sqrt{2}G}{n} \Big\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \Big\|_2.$$

We can combine the above three inequalities together and derive

$$\begin{aligned} \mathbb{E}_{i_t} \Big\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_{t+1}' \\ \mathbf{v}_{t+1} - \mathbf{v}_{t+1}' \end{pmatrix} \Big\|_2^2 &\leq \big(1 + 2\rho_t\eta_t + L^2\eta_t^2\big) \Big\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \Big\|_2^2 + \frac{8\eta_t^2 G^2}{n} + \frac{4\sqrt{2}G\eta_t}{n} \Big\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \Big\|_2 \\ &\leq \big(1 + 2\rho_t\eta_t + L^2\eta_t^2\big) \Big\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \Big\|_2^2 + \frac{8\eta_t^2 G^2}{n} + \eta_t^2 \Big\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}_t' \\ \mathbf{v}_t - \mathbf{v}_t' \end{pmatrix} \Big\|_2^2 + \frac{8G^2}{n^2}. \end{aligned}$$

Applying the above inequality recursively, we get

$$\mathbb{E}_A \Big\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_{t+1}' \\ \mathbf{v}_{t+1} - \mathbf{v}_{t+1}' \end{pmatrix} \Big\|_2^2 \leq \frac{8G^2}{n} \sum_{j=1}^{t} \Big(\eta_t^2 + \frac{1}{n}\Big) \prod_{k=j+1}^{t} \big(1 + 2\rho_k\eta_k + L^2\eta_k^2 + \eta_k^2\big).$$

By the elementary inequality $1 + a \leq \exp(a)$ we know

$$\mathbb{E}_A \Big\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_{t+1}' \\ \mathbf{v}_{t+1} - \mathbf{v}_{t+1}' \end{pmatrix} \Big\|_2^2 \leq \frac{8G^2}{n} \sum_{j=1}^{t} \Big(\eta_t^2 + \frac{1}{n}\Big) \exp\Big(\sum_{k=j+1}^{t} \big(2\rho_k\eta_k + (L^2+1)\eta_k^2\big)\Big).$$

It then follows from the Jensen's inequality that

$$\mathbb{E}_A \Big\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}_{t+1}' \\ \mathbf{v}_{t+1} - \mathbf{v}_{t+1}' \end{pmatrix} \Big\|_2 \leq \frac{2\sqrt{2}G}{\sqrt{n}} \left(\sum_{j=1}^{t} \Big(\eta_t^2 + \frac{1}{n}\Big) \exp\Big(\sum_{k=j+1}^{t} \big(2\rho_k\eta_k + (L^2+1)\eta_k^2\big)\Big) \right)^{\frac{1}{2}}.$$

The stated bound then follows from Part (a) of Theorem 1 and Assumption 1. The proof is complete. □

G. Stability and Generalization Bounds of AGDA on Nonconvex-Nonconcave Objectives

In this section, we give the proof on the stability and generalization bounds of AGDA for nonconvex-nonconcave functions. The next lemma is similar to Lemma F.1, which shows AGDA typically runs several iterations before encountering the different example between S and S'.

Lemma G.1. *Assume $|f(\cdot,\cdot,z)| \leq 1$ for any z and let Assumption 1 hold. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z_1, \ldots, z_{n-1}, z_n'\}$. Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ and $\{\mathbf{w}_t', \mathbf{v}_t'\}$ be the sequence produced by (5.2) w.r.t. S and S', respectively. Denote*

$$\Delta_t = \|\mathbf{w}_t - \mathbf{w}_t'\|_2 + \|\mathbf{v}_t - \mathbf{v}_t'\|_2. \tag{G.1}$$

Then for any $t_0 \in \mathbb{N}$ and any $\mathbf{w}', \mathbf{v}'$ we have

$$\mathbb{E}[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}_T', \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}_T'; z)] \leq \frac{8t_0}{n} + G\mathbb{E}[\Delta_T | \Delta_{t_0} = 0].$$

Proof. Let $\mathcal{E}$ denote the event that $\Delta_{t_0} = 0$. Then we have

$$\begin{aligned} &\mathbb{E}[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}_T', \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}_T'; z)] \\ =&\mathbb{P}[\mathcal{E}]\mathbb{E}[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}_T', \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}_T'; z) | \mathcal{E}] \\ &+ \mathbb{P}[\mathcal{E}^c]\mathbb{E}[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}_T', \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}_T'; z) | \mathcal{E}^c] \\ \leq& G\mathbb{E}[\Delta_T | \mathcal{E}] + 4\mathbb{P}[\mathcal{E}^c], \end{aligned} \tag{G.2}$$

where we have used (F.2) and the assumption $|f(\cdot,\cdot,z)|\leq 1$. Using the union bound on the outcome $i_t=n$ and $j_t=n$ we obtain that

$$\mathbb{P}[\mathcal{E}^c]\leq\sum_{t=1}^{t_0}\big(\mathbb{P}[i_t=n]+\mathbb{P}[j_t=n]\big)=\frac{2t_0}{n}.$$

The proof is complete by combining the above two inequalities together. □

Proof of Theorem 7. Since z_{i_t} and z_{j_t} are i.i.d, we can analyze the update of $\mathbf{w}$ and $\mathbf{v}$ separately. Note that the projection step is nonexpansive. We consider two cases at the t-th iteration. If $i_t\neq n$, then it follows from Assumption 2 that

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2\\ \leq&\|\mathbf{w}_t-\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t,z_{i_t})-\mathbf{w}'_t+\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t,z_{i_t})\|_2\\ \leq&\|\mathbf{w}_t-\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t,z_{i_t})-\mathbf{w}'_t+\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}_t,z_{i_t})\|_2+\|\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}_t,z_{i_t})-\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t,z_{i_t})\|_2\\ \leq&(1+L\eta_{\mathbf{w},t})\|\mathbf{w}_t-\mathbf{w}'_t\|_2+L\eta_{\mathbf{w},t}\|\mathbf{v}_t-\mathbf{v}'_t\|_2.\end{aligned}$$

If $i_t=n$, then it follows from Assumption 1 that

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2\leq&\|\mathbf{w}_t-\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t,z_{i_t})-\mathbf{w}'_t+\eta_{\mathbf{w},t}\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t,z_{i_t})\|_2\\ \leq&\|\mathbf{w}_t-\mathbf{w}'_t\|_2+2G\eta_{\mathbf{w},t}.\end{aligned}$$

According to the distribution of i_t, we have

$$\begin{aligned}\mathbb{E}_A[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2]\leq&\frac{n-1}{n}\mathbb{E}_A\Big[(1+\eta_{\mathbf{w},t}L)\|\mathbf{w}_t-\mathbf{w}'_t\|_2+L\eta_{\mathbf{w},t}\|\mathbf{v}_t-\mathbf{v}'_t\|_2\Big]+\frac{1}{n}(\|\mathbf{w}_t-\mathbf{w}'_t\|_2+2\eta_{\mathbf{w},t}G)\\ \leq&(1+\eta_{\mathbf{w},t}L)\mathbb{E}_A[\|\mathbf{w}_t-\mathbf{w}'_t\|_2]+L\eta_{\mathbf{w},t}\mathbb{E}_A\big[\|\mathbf{v}_t-\mathbf{v}'_t\|_2\big]+\frac{2\eta_{\mathbf{w},t}G}{n}.\end{aligned}\tag{G.3}$$

Similarly, for $\mathbf{v}$ we also have

$$\mathbb{E}_A[\|\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\|_2]\leq(1+\eta_{\mathbf{v},t}L)\mathbb{E}_A[\|\mathbf{v}_t-\mathbf{v}'_t\|_2]+L\eta_{\mathbf{v},t}\mathbb{E}_A\big[\|\mathbf{w}_t-\mathbf{w}'_t\|_2\big]+\frac{2\eta_{\mathbf{v},t}G}{n}.\tag{G.4}$$

Combining (G.3) and (G.4) we have

$$\mathbb{E}_A[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2+\|\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\|_2]\leq(1+(\eta_{\mathbf{w},t}+\eta_{\mathbf{v},t})L)\mathbb{E}_A\big[\|\mathbf{w}_t-\mathbf{w}'_t\|_2+\|\mathbf{v}_t-\mathbf{v}'_t\|_2\big]+\frac{2(\eta_{\mathbf{w},t}+\eta_{\mathbf{v},t})G}{n}.$$

Recalling the event $\mathcal{E}$ that $\Delta_{t_0}=0$, we apply the above equation recursively from $t=t_0+1$ to T, then

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2+\|\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\|_2\big|\Delta_{t_0}=0\big]\leq\frac{2G}{n}\sum_{t=t_0+1}^{T}(\eta_{\mathbf{w},t}+\eta_{\mathbf{v},t})\prod_{k=t+1}^{T}(1+(\eta_{\mathbf{w},k}+\eta_{\mathbf{v},k})L).$$

By the elementary inequality $1+x\leq\exp(x)$ and $\eta_{\mathbf{w},t}+\eta_{\mathbf{v},t}\leq\frac{c}{t}$, we have

$$\begin{aligned}&\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\|_2+\|\mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\|_2\big|\Delta_{t_0}=0\big]\\ \leq&\frac{2cG}{n}\sum_{t=t_0+1}^{T}\frac{1}{t}\prod_{k=t+1}^{T}\exp\Big(\frac{cL}{k}\Big)=\frac{2cG}{n}\sum_{t=t_0+1}^{T}\frac{1}{t}\exp\Big(\sum_{k=t+1}^{T}\frac{cL}{k}\Big)\\ \leq&\frac{2cG}{n}\sum_{t=t_0+1}^{T}\frac{1}{t}\exp\Big(cL\log\Big(\frac{T}{t}\Big)\Big)\leq\frac{2cGT^{cL}}{n}\sum_{t=t_0+1}^{T}\frac{1}{t^{cL+1}}\leq\frac{2G}{Ln}\Big(\frac{T}{t_0}\Big)^{cL}.\end{aligned}$$

By Lemma G.1 we have

$$\mathbb{E}[f(\mathbf{w}_T,\mathbf{v}';z)-f(\mathbf{w}'_T,\mathbf{v}';z)+f(\mathbf{w}',\mathbf{v}_T;z)-f(\mathbf{w}',\mathbf{v}'_T;z)]\leq\frac{8t_0}{n}+\frac{2G^2}{Ln}\Big(\frac{T}{t_0}\Big)^{cL}.\tag{G.5}$$

The right hand side of the above inequality is approximately minimized when

$$t_0 = \Big(\frac{G^2}{4L}\Big)^{\frac{1}{cL+1}} T^{\frac{cL}{cL+1}}.$$

Plugging it into Eq. (G.5) we have (for simplicity we assume the above t_0 is an integer)

$$\mathbb{E}[f(\mathbf{w}_T, \mathbf{v}'; z) - f(\mathbf{w}'_T, \mathbf{v}'; z) + f(\mathbf{w}', \mathbf{v}_T; z) - f(\mathbf{w}', \mathbf{v}'_T; z)] \leq 16\Big(\frac{G^2}{4L}\Big)^{\frac{1}{cL+1}} n^{-1} T^{\frac{cL}{cL+1}}.$$

Since the above bound holds for all z, S, S' and $\mathbf{w}', \mathbf{v}'$, we immediately get the same upper bound on the weak stability. Finally the theorem holds by calling Theorem 1, Part (a). □

We require an assumption on the existence of saddle point to address the optimization error of AGDA (Yang et al., 2020).

Assumption 4 (Existence of Saddle Point)**.** Assume for any S, F_S has at least one saddle point. Assume for any $\mathbf{v}$, $\min_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v})$ has a nonempty solution set and a finite optimal value. Assume for any $\mathbf{w}$, $\max_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v})$ has a nonempty solution set and a finite optimal value.

The following lemma establishes the generalization bound for the empirical maximizer of a strongly concave objective. It is a direct extension of the stability analysis in Shalev-Shwartz et al. (2010) for strongly convex objectives.

Lemma G.2. *Assume that for any $\mathbf{w}$ and S, the function $\mathbf{v} \mapsto F_S(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave. Suppose for any $\mathbf{w}$ and $\mathbf{v}, \mathbf{v}'$ and for any z we have*

$$\big|f(\mathbf{w}, \mathbf{v}; z) - f(\mathbf{w}, \mathbf{v}'; z)\big| \leq G\|\mathbf{v} - \mathbf{v}'\|_2. \tag{G.6}$$

Fix any $\mathbf{w}$. Denote $\hat{\mathbf{v}}_S^* = \arg\max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}, \mathbf{v})$. Then

$$\mathbb{E}[F_S(\mathbf{w}, \hat{\mathbf{v}}_S^*) - F(\mathbf{w}, \hat{\mathbf{v}}_S^*)] \leq \frac{4G^2}{\rho n}.$$

Proof. Let $S' = \{z'_1, \ldots, z'_n\}$ be drawn independently from ρ. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$. Denote $\hat{\mathbf{v}}^*_{S^{(i)}} = \arg\max_{\mathbf{v}\in\mathcal{V}} F_{S^{(i)}}(\mathbf{w}, \mathbf{v})$. Then

$$\begin{aligned}
F_S(\mathbf{w}, \hat{\mathbf{v}}_S^*) - F_S(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}) =& \frac{1}{n}\sum_{j\neq i}\Big(f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z_j) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_j)\Big) + \frac{1}{n}\Big(f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i)\Big)\\
=& \frac{1}{n}\Big(f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z'_i) - f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z'_i)\Big) + \frac{1}{n}\Big(f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i)\Big)\\
&+ F_{S^{(i)}}(\mathbf{w}, \hat{\mathbf{v}}_S^*) - F_{S^{(i)}}(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}})\\
\leq& \frac{1}{n}\Big(f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z'_i) - f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z'_i)\Big) + \frac{1}{n}\Big(f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i)\Big)\\
\leq& \frac{2G}{n}\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}^*_{S^{(i)}}\big\|_2,
\end{aligned} \tag{G.7}$$

where the first inequality follows from the fact that $\hat{\mathbf{v}}^*_{S^{(i)}}$ is the maximizer of $F_{S^{(i)}}(\mathbf{w}, \cdot)$ and the second inequality follows from (G.6). Since F_S is strongly-concave and $\hat{\mathbf{v}}_S^*$ maximizes $F_S(\mathbf{w}, \cdot)$, we know

$$\frac{\rho}{2}\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}^*_{S^{(i)}}\big\|_2^2 \leq F_S(\mathbf{w}, \hat{\mathbf{v}}_S^*) - F_S(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}).$$

Combining it with (G.7) we get $\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}^*_{S^{(i)}}\big\|_2 \leq 4G/(\rho n)$. By (G.6), the following inequality holds for any z

$$\big|f(\mathbf{w}, \hat{\mathbf{v}}_S^*; z) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z)\big| \leq \frac{4G^2}{\rho n}.$$

Since z_i and z'_i are i.i.d., we have

$$\mathbb{E}\big[F(\mathbf{w}, \hat{\mathbf{v}}_S^*)\big] = \mathbb{E}\big[F(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}})\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i)\big],$$

where the last identity holds since z_i is independent of $\hat{\mathbf{v}}^*_{S^{(i)}}$. Therefore

$$\mathbb{E}\big[F_S(\mathbf{w},\hat{\mathbf{v}}_S^*) - F(\mathbf{w},\hat{\mathbf{v}}_S^*)\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(\mathbf{w},\hat{\mathbf{v}}_S^*;z_i) - f(\mathbf{w},\hat{\mathbf{v}}^*_{S^{(i)}};z_i)\big] \le \frac{4G^2}{\rho n}.$$

The proof is complete. □

Corollary G.3. *Let $\beta_1,\rho>0$. Let Assumptions 1, 2, 3 with $\beta_1(S)\ge\beta_1,\beta_2(S)\ge\rho$ and 4 hold. Assume for any $\mathbf{w}$ and any S, the functions $\mathbf{v}\mapsto F(\mathbf{w},\mathbf{v})$ and $\mathbf{v}\mapsto F_S(\mathbf{w},\mathbf{v})$ are ρ-strongly concave. Let $\{\mathbf{w}_t,\mathbf{v}_t\}$ be the sequence produced by (5.2) with $\eta_{\mathbf{w},t}\asymp 1/(\beta_1 t)$ and $\eta_{\mathbf{v},t}\asymp 1/(\beta_1\rho^2 t)$. Then for $T\asymp\big(\frac{n}{\beta_1^2\rho^3}\big)^{\frac{cL+1}{2cL+1}}$, we have*

$$\mathbb{E}\big[R(\mathbf{w}_T) - R(\mathbf{w}^*)\big] = O\Big(n^{-\frac{cL+1}{2cL+1}}\beta_1^{-\frac{2cL}{2cL+1}}\rho^{-\frac{5cL+1}{2cL+1}}\Big),$$

where $c\asymp 1/(\beta_1\rho^2)$.

Proof. We have the error decomposition

$$R(\mathbf{w}_T) - R(\mathbf{w}^*) = \big(R(\mathbf{w}_T) - R_S(\mathbf{w}_T)\big) + \big(R_S(\mathbf{w}_T) - R_S(\mathbf{w}^*)\big) + \big(R_S(\mathbf{w}^*) - R(\mathbf{w}^*)\big). \tag{G.8}$$

First we consider the term $R(\mathbf{w}_T) - R_S(\mathbf{w}_T)$. Analogous to the proof of Theorem 7 (i.e., the only difference is to replace the conditional expectation of function values in (G.2) with the conditional expectation of $\mathbb{E}[\|\mathbf{w}_T-\mathbf{w}'_T\|_2+\|\mathbf{v}_T-\mathbf{v}'_T\|_2]$), one can show that AGDA is $O\big(n^{-1}T^{\frac{cL}{cL+1}}\big)$-argument stable (note the step size satisfy $\eta_{\mathbf{w},t}+\eta_{\mathbf{v},t}\le c/t$). This together with Part (b) of Theorem 1 implies that

$$\mathbb{E}\big[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)\big] = O\big((\rho n)^{-1}T^{\frac{cL}{cL+1}}\big). \tag{G.9}$$

For the term $R_S(\mathbf{w}_T) - R_S(\mathbf{w}^*)$, the optimization error bounds in Yang et al. (2020) show that

$$\mathbb{E}\big[R_S(\mathbf{w}_T) - R_S(\mathbf{w}^*)\big] = O\Big(\frac{1}{\beta_1^2\rho^4 T}\Big). \tag{G.10}$$

Finally, for the term $R_S(\mathbf{w}^*) - R(\mathbf{w}^*)$, we further decompose it as

$$\mathbb{E}\big[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)\big] = \mathbb{E}\big[F_S(\mathbf{w}^*,\hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*,\mathbf{v}^*)\big] = \mathbb{E}\big[F_S(\mathbf{w}^*,\hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*,\hat{\mathbf{v}}_S^*)\big] + \mathbb{E}\big[F(\mathbf{w}^*,\hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*,\mathbf{v}^*)\big],$$

where $\hat{\mathbf{v}}_S^* = \arg\max_{\mathbf{v}} F_S(\mathbf{w}^*,\mathbf{v})$. The second term $\mathbb{E}\big[F(\mathbf{w}^*,\hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*,\mathbf{v}^*)\big]\le 0$ since $(\mathbf{w}^*,\mathbf{v}^*)$ is a saddle point of F. Therefore by Lemma G.2 we have

$$\mathbb{E}\big[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)\big] \le \mathbb{E}\big[F_S(\mathbf{w}^*,\hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*,\hat{\mathbf{v}}_S^*)\big] = O\Big(\frac{1}{\rho n}\Big).$$

We can plug the above inequality, (G.9), (G.10) into (G.8), and get

$$\mathbb{E}\big[R(\mathbf{w}_T) - R(\mathbf{w}^*)\big] = O\big((\rho n)^{-1}T^{\frac{cL}{cL+1}}\big) + O\Big(\frac{1}{\beta_1^2\rho^4 T}\Big) + O\Big(\frac{1}{\rho n}\Big).$$

We can choose $T\asymp\big(\frac{n}{\beta_1^2\rho^3}\big)^{\frac{cL+1}{2cL+1}}$ to get the stated excess primal population risk bounds. The proof is complete. □

H. Proof of Theorem 9

To prove Theorem 9, we first introduce a lemma on relating the difference of function values to gradients.

Lemma H.1. *Let Assumption 3 hold. For any $\mathbf{u}=(\mathbf{w},\mathbf{v})$ and any stationary point $\mathbf{u}_{(S)}=(\mathbf{w}_{(S)},\mathbf{v}_{(S)})$ of F_S, we have*

$$-\frac{\|\nabla_{\mathbf{v}}F_S(\mathbf{w},\mathbf{v})\|_2^2}{2\beta_2(S)} \le F_S(\mathbf{u}) - F_S(\mathbf{u}_{(S)}) \le \frac{\|\nabla_{\mathbf{w}}F_S(\mathbf{w},\mathbf{v})\|_2^2}{2\beta_1(S)}.$$

Proof. Since $\mathbf{u}_{(S)}$ is a stationary point, it is also a saddle point under the PL condition (Yang et al., 2020) which means that

$$F_S(\mathbf{w}_{(S)}, \mathbf{v}') \leq F_S(\mathbf{w}_{(S)}, \mathbf{v}_{(S)}) \leq F_S(\mathbf{w}', \mathbf{v}_{(S)}), \quad \forall \mathbf{w}' \in \mathcal{W}, \mathbf{v}' \in \mathcal{V}.$$

It then follows that

$$\begin{aligned} F_S(\mathbf{u}) - F_S(\mathbf{u}_{(S)}) &= F_S(\mathbf{w}, \mathbf{v}) - F_S(\mathbf{w}_{(S)}, \mathbf{v}) + F_S(\mathbf{w}_{(S)}, \mathbf{v}) - F_S(\mathbf{w}_{(S)}, \mathbf{v}_{(S)}) \\ &\leq F_S(\mathbf{w}, \mathbf{v}) - F_S(\mathbf{w}_{(S)}, \mathbf{v}) \leq F_S(\mathbf{w}, \mathbf{v}) - \inf_{\mathbf{w}' \in \mathcal{W}} F_S(\mathbf{w}', \mathbf{v}) \leq \frac{1}{2\beta_1(S)} \|\nabla_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v})\|_2^2, \end{aligned}$$

where in the last inequality we have used Assumption 3. In a similar way, we know

$$\begin{aligned} F_S(\mathbf{u}) - F_S(\mathbf{u}_{(S)}) &= F_S(\mathbf{w}, \mathbf{v}) - F_S(\mathbf{w}, \mathbf{v}_{(S)}) + F_S(\mathbf{w}, \mathbf{v}_{(S)}) - F_S(\mathbf{w}_{(S)}, \mathbf{v}_{(S)}) \\ &\geq F_S(\mathbf{w}, \mathbf{v}) - F_S(\mathbf{w}, \mathbf{v}_{(S)}) \geq F_S(\mathbf{w}, \mathbf{v}) - \sup_{\mathbf{v}'} F_S(\mathbf{w}, \mathbf{v}') \geq -\frac{1}{2\beta_2(S)} \|\nabla_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v})\|_2^2. \end{aligned}$$

The proof is complete. □

Proof of Theorem 9. Let $S' = \{z_1', \ldots, z_n'\}$ be drawn independently from ρ. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}$. Let $\mathbf{u}_S = (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ and $\mathbf{u}_S^{(S)}$ be the projection of $\mathbf{u}_S$ onto the set of stationary points of F_S. For each $i \in [n]$, we denote $\mathbf{u}_i = (A_{\mathbf{w}}(S^{(i)}), A_{\mathbf{v}}(S^{(i)}))$ and $\mathbf{u}_i^{(i)}$ the projection of $\mathbf{u}_i$ onto the set of stationary points of $F_{S^{(i)}}$. Then $\nabla F_{S^{(i)}}(\mathbf{u}_i^{(i)}) = 0$.

We decompose $f(\mathbf{u}_i; z_i) - f(\mathbf{u}_S; z_i)$ as follows

$$f(\mathbf{u}_i; z_i) - f(\mathbf{u}_S; z_i) = \big(f(\mathbf{u}_i; z_i) - f(\mathbf{u}_i^{(i)}; z_i)\big) + \big(f(\mathbf{u}_i^{(i)}; z_i) - f(\mathbf{u}_S^{(S)}; z_i)\big) + \big(f(\mathbf{u}_S^{(S)}; z_i) - f(\mathbf{u}_S; z_i)\big). \tag{H.1}$$

We now address the above three terms separately.

We first address $f(\mathbf{u}_i^{(i)}; z_i) - f(\mathbf{u}_S^{(S)}; z_i)$. According to the definition of $F_S, S, S^{(i)}$, we know

$$f(\mathbf{u}_i^{(i)}; z_i) = nF_S(\mathbf{u}_i^{(i)}) - nF_{S^{(i)}}(\mathbf{u}_i^{(i)}) + f(\mathbf{u}_i^{(i)}; z_i').$$

Since z_i and z_i' follow from the same distribution, we know $\mathbb{E}[f(\mathbf{u}_i^{(i)}; z_i')] = \mathbb{E}[f(\mathbf{u}_S^{(S)}; z_i)]$ and further get

$$\mathbb{E}\big[f(\mathbf{u}_i^{(i)}; z_i)\big] = n\mathbb{E}\big[F_S(\mathbf{u}_i^{(i)})\big] - n\mathbb{E}\big[F_{S^{(i)}}(\mathbf{u}_i^{(i)})\big] + \mathbb{E}\big[f(\mathbf{u}_S^{(S)}; z_i)\big].$$

It then follows that

$$\mathbb{E}\big[f(\mathbf{u}_i^{(i)}; z_i) - f(\mathbf{u}_S^{(S)}; z_i)\big] = n\mathbb{E}\big[F_S(\mathbf{u}_i^{(i)}) - F_{S^{(i)}}(\mathbf{u}_i^{(i)})\big] = n\mathbb{E}\Big[F_S(\mathbf{u}_i^{(i)}) - F_S(\mathbf{u}_S^{(S)})\Big], \tag{H.2}$$

where we have used the following identity due to the symmetry between z_i and z_i': $\mathbb{E}[F_{S^{(i)}}(\mathbf{u}_i^{(i)})] = \mathbb{E}\Big[F_S(\mathbf{u}_S^{(S)})\Big]$. By the PL condition of F_S, it then follows from (H.2) and Lemma H.1 that

$$\mathbb{E}\big[f(\mathbf{u}^{(i)}; z_i) - f(\mathbf{u}_S^{(S)}; z_i)\big] \leq \frac{n}{2}\mathbb{E}\big[\frac{1}{\beta_1(S)}\|\nabla_{\mathbf{w}} F_S(\mathbf{u}_i^{(i)})\|_2^2\big]. \tag{H.3}$$

According to the definition of $\mathbf{u}_i^{(i)}$ we know $\nabla_{\mathbf{w}} F_{S^{(i)}}(\mathbf{u}_i^{(i)}) = 0$ and therefore ($(a+b)^2 \leq 2a^2 + 2b^2$)

$$\begin{aligned} \|\nabla_{\mathbf{w}} F_S(\mathbf{u}_i^{(i)})\|_2^2 &= \Big\|\nabla_{\mathbf{w}} F_{S^{(i)}}(\mathbf{u}_i^{(i)}) - \frac{1}{n}\nabla_{\mathbf{w}} f(\mathbf{u}_i^{(i)}; z_i') + \frac{1}{n}\nabla_{\mathbf{w}} f(\mathbf{u}_i^{(i)}; z_i)\Big\|_2^2 \\ &\leq \frac{2}{n^2}\|\nabla_{\mathbf{w}} f(\mathbf{u}_i^{(i)}; z_i')\|_2^2 + \frac{2}{n^2}\|\nabla_{\mathbf{w}} f(\mathbf{u}_i^{(i)}; z_i)\|_2^2 \leq \frac{4G^2}{n^2}, \end{aligned} \tag{H.4}$$

where we have used Assumption 1. This together with (H.3) gives

$$\mathbb{E}\big[f(\mathbf{u}^{(i)}; z_i) - f(\mathbf{u}_S^{(S)}; z_i)\big] \leq \frac{2G^2}{n}\mathbb{E}\Big[\frac{1}{\beta_1(S)}\Big]. \tag{H.5}$$

We then address $f(\mathbf{u}_i;z_i)-f(\mathbf{u}_i^{(i)};z_i)$. Since $\mathbf{u}_i$ and $\mathbf{u}_i^{(i)}$ are independent of z_i, we know

$$\mathbb{E}\big[f(\mathbf{u}_i;z_i)-f(\mathbf{u}_i^{(i)};z_i)\big]=\mathbb{E}\big[F(\mathbf{u}_i)-F(\mathbf{u}_i^{(i)})\big]=\mathbb{E}\big[F(\mathbf{u}_S)-F(\mathbf{u}_S^{(S)})\big], \tag{H.6}$$

where we have used the symmetry between z_i and z_i'.

Finally, we address $f(\mathbf{u}_S^{(S)};z_i)-f(\mathbf{u}_S;z_i)$. By the definition of $\mathbf{u}_S^{(S)}$ we know

$$\sum_{i=1}^n\big(f(\mathbf{u}_S^{(S)};z_i)-f(\mathbf{u}_S;z_i)\big)=n\big(F_S(\mathbf{u}_S^{(S)})-F_S(\mathbf{u}_S)\big). \tag{H.7}$$

Plugging (H.5), (H.6) and the above inequality back into (H.1), we derive

$$\sum_{i=1}^n\mathbb{E}\big[f(\mathbf{u}_i;z_i)-f(\mathbf{u}_S;z_i)\big]\leq\mathbb{E}\Big[\frac{2G^2}{\beta_1(S)}\Big]+n\mathbb{E}\big[F(\mathbf{u}_S)-F(\mathbf{u}_S^{(S)})\big]+n\mathbb{E}\big[F_S(\mathbf{u}_S^{(S)})-F_S(\mathbf{u}_S)\big].$$

Since z_i and z_i' are drawn from the same distribution, we know

$$\begin{aligned}\mathbb{E}\big[F(\mathbf{u}_S)-F_S(\mathbf{u}_S)\big]&=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big[F(\mathbf{u}_i)-F_S(\mathbf{u}_S)\big]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big[f(\mathbf{u}_i;z_i)-f(\mathbf{u}_S;z_i)\big]\\&\leq\frac{2G^2}{n}\mathbb{E}\Big[\frac{1}{\beta_1(S)}\Big]+\mathbb{E}\big[F(\mathbf{u}_S)-F(\mathbf{u}_S^{(S)})\big]+\mathbb{E}\big[F_S(\mathbf{u}_S^{(S)})-F_S(\mathbf{u}_S)\big],\end{aligned} \tag{H.8}$$

where the second identity holds since z_i is independent of $\mathbf{u}_i$. It then follows that

$$\mathbb{E}\big[F(\mathbf{u}_S^{(S)})-F_S(\mathbf{u}_S^{(S)})\big]\leq\frac{2G^2}{n}\mathbb{E}\Big[\frac{1}{\beta_1(S)}\Big]. \tag{H.9}$$

According to the Lipschitz continuity we know

$$\big|F(\mathbf{u}_S)-F(\mathbf{u}_S^{(S)})\big|\leq G\|\mathbf{u}_S-\mathbf{u}_S^{(S)}\|_2\quad\text{and}\quad\big|F_S(\mathbf{u}_S)-F_S(\mathbf{u}_S^{(S)})\big|\leq G\|\mathbf{u}_S-\mathbf{u}_S^{(S)}\|_2.$$

Plugging the above inequality back into (H.8), we derive the following inequality

$$\mathbb{E}\big[F(\mathbf{u}_S)-F_S(\mathbf{u}_S)\big]\leq\frac{2G^2}{n}\mathbb{E}\Big[\frac{1}{\beta_1(S)}\Big]+2G\mathbb{E}\big[\|\mathbf{u}_S-\mathbf{u}_S^{(S)}\|_2\big]. \tag{H.10}$$

By Lemma H.1 and (H.2), we can also have

$$\mathbb{E}\big[f(\mathbf{u}_i^{(i)};z_i)-f(\mathbf{u}_S^{(S)};z_i)\big]\geq-\frac{n}{2}\mathbb{E}\big[\frac{1}{\beta_2(S)}\|\nabla_{\mathbf{v}}F_S(\mathbf{u}_i^{(i)})\|_2^2\big].$$

Using this inequality, one can analyze analogously to (H.10) and derive the following inequality

$$\mathbb{E}\big[F(\mathbf{u}_S)-F_S(\mathbf{u}_S)\big]\geq-\frac{2G^2}{n}\mathbb{E}\Big[\frac{1}{\beta_2(S)}\Big]-2G\mathbb{E}\big[\|\mathbf{u}_S-\mathbf{u}_S^{(S)}\|_2\big].$$

The stated inequality follows from the above inequality and (H.10). The proof is complete. □