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	$\mathbb{E}_A\left[\left\Vert\begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S')\end{pmatrix}\right\Vert_2\right]$ or $\left\Vert\begin{pmatrix} A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S') \\ A_{\mathbf{v}}(S) - A_{\mathbf{v}}(S')\end{pmatrix}\right\Vert_2$
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

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

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

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

The stated bound in Part (b) then follows.

B.3. Proof of Part (c)

In a similar way, one can show that

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

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

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

where we have used the elementary inequality $a+b\le\sqrt{2(a^2+b^2)}$. 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 by $S\backslash\{z_i\}$ 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 $\|Z\|_p := \big(\mathbb{E}[|Z|^p]\big)^{\frac{1}{p}}, p\ge 1$.

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

- $\big|\mathbb{E}_{S\setminus\{z_i\}}[g_i(S)]\big| \leq M$ *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 $z''_j \in \mathcal{Z}$*

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

Then, for any $p \geq 2$

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

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 $\|Z\|_p \leq \sqrt{p}a + pb$ for any $p \geq 2$. Then with probability at least $1 - \delta$*

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

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 $\mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_{S^{(i)}}; Z)] \geq \mathbb{E}_Z[f(A_{\mathbf{w}}(S^{(i)}), \mathbf{v}^*_S; Z)]$. It then follows that

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

According to (B.2), we know

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

As a combination of the above two inequalities, we derive

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

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

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

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

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

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 proves 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

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

(b) If Assumption 2 holds, then

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

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*

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

Proof of Lemma C.1. It is clear that

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

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

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

We can combine the above inequality with the Lipschitz continuity to derive Part (a). We refer 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

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

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

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

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

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

Applying this inequality recursively implies that

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

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

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

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

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

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

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

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

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

We can apply the above inequality recursively and derive

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

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

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

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

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

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 $1-\delta$

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

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

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

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

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

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

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

It then follows that

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

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

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

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 an 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 $|\xi_k-\mathbb{E}_{z_k}[\xi_k]|\leq b_k$ for each k. For $\delta\in(0,1)$, with probability at least $1-\delta$ we have*

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

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 $\|\xi_k\|\leq B$ and $\sum_{k=1}^{t}\mathbb{E}[\|\xi_k\|^2|\xi_1,\ldots,\xi_{k-1}]\leq\sigma_t^2$ for $\sigma_t\geq0$. Then, for any $0<\delta<1$, the following inequality holds with probability at least $1-\delta$*

$$\max_{1\leq j\leq t}\Big\|\sum_{k=1}^{j}\xi_k\Big\|\leq2\Big(\frac{B}{3}+\sigma_t\Big)\log\frac{2}{\delta}.$$

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 $\sup_{\mathbf{w}\in\mathcal{W}}\|\mathbf{w}\|_2\leq B_W$ and $\sup_{\mathbf{v}\in\mathcal{V}}\|\mathbf{v}\|_2\leq B_V$. Then the following inequality holds*

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

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

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

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

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

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

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

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

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

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

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

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

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

Note

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

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

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

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

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

It then follows that

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

In a similar way, we can show that

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

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

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

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

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 $\|\xi_t\|_2 \leq 2G$ and

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

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

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

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

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

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

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

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

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

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, $t_0 \geq 0$ 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 $\big(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T\big)$ defined in (4.2) we have*

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

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

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

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

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

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

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

Multiplying both sides by $t + t_0$ gives

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

Taking a summation of the above inequality further gives

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

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

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

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

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

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

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

In a similar way one can show

$$\sum_{t=1}^{T}\mathbb{E}_A\big[\sup_{\mathbf{v}\in\mathcal{V}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-F_S(\mathbf{w}_t,\mathbf{v}_t)\big]\leq 2\rho t_0 B_V^2+\frac{G^2\log(eT)}{2\rho}+B_VG\sqrt{T}.$$

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

$$\sum_{t=1}^{T}\mathbb{E}_A\big[F_S(\mathbf{w}_t,\mathbf{v}_t)-F_S(\mathbf{w},\bar{\mathbf{v}}_T)\big]\leq 2\rho t_0 B_W^2+\frac{G^2\log(eT)}{2\rho}.$$

In a similar way, one can show

$$\sum_{t=1}^{T}\mathbb{E}_A\big[F_S(\bar{\mathbf{w}}_T,\mathbf{v})-F_S(\mathbf{w}_t,\mathbf{v}_t)\big]\leq 2\rho t_0 B_V^2+\frac{G^2\log(eT)}{2\rho}.$$

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

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

It then follows from the convexity of a norm that

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

and therefore

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

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

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

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 from 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)[3] 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

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

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 the 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.

[3] 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 $1-\delta$

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

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

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

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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)

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

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 with 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 $a > 0$,*

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

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

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

Theorem F.4. *Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be the sequence produced by (4.1) with $\eta_t \le \frac{c}{t}$ for some $c > 0$. Assume Assumption 1, 2 hold and $|f(\cdot,\cdot;z)| \le 1$. For any $\delta \in (0,1)$, if $c \le \frac{1}{(n\log(2/\delta)-1)L}$, 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 $i_t \ne n$, it follows from the L-smoothness of f that

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

If $i_t = n$, we have

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

We can combine the above two inequalities together and get

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

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

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

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

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

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

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

Note that

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

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

$$\left\|\begin{pmatrix}\mathbf{w}_T - \mathbf{w}_T' \\ \mathbf{v}_T - \mathbf{v}_T'\end{pmatrix}\right\|_2 \le 4cGT^{cL}\Big(\frac{cL+1}{cLn} + (e-1)\sqrt{\frac{(cL+1)\log(1/\delta)}{cLn}}\Big).$$

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

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

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

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

Therefore, we have

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

According to (4.1), we know

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

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

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

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

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

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

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

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 $\mathbf{w}$, $\mathbf{v}, \mathbf{v}'$ 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

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

The proof is complete. □

Corollary G.3. *Let $\beta_1,\rho>0$. Let Assumptions 1, 2, 3 with $\beta_1(S)\geq\beta_1,\beta_2(S)\geq\rho$ 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 sizes satisfy $\eta_{\mathbf{w},t}+\eta_{\mathbf{v},t}\leq c/t$). 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

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

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 $\mathbb{E}\big[F(\mathbf{w}^*,\hat{\mathbf{v}}^*_S)-F(\mathbf{w}^*,\mathbf{v}^*)\big]\leq 0$ since $(\mathbf{w}^*,\mathbf{v}^*)$ is a saddle point of F. Therefore by Lemma G.2 we have

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

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

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

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*

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

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

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

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

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

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

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

According to the Lipschitz continuity we know

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

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

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

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

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

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

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

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

I. Additional Experiments

In this section, we investigate the stability of SGDA on a nonconvex-nonconcave problem. We consider the vanilla GAN structure proposed in Goodfellow et al. (2014). The generator and the discriminator consist of 4 fully connected layers, and use the leaky rectified linear activation before the output layer. The generator uses the hyperbolic tangent activation at the output layer. The discriminator uses the sigmoid activation at the output layer. In order to make experiments more interpretable in terms of stability, we remove all forms of regularization such as the weight decay or dropout in the original paper. In order to truly implement SGDA, we generate only one noise for updating both the discriminator and the generator at each iteration. This differs from the common GAN training strategy, which uses different noises for updating the discriminator and the generator. We employ the `mnist` dataset (LeCun et al., 1998) and build neighboring datasets S and S' by removing a randomly chosen datum indexed by i from S and $i+1$ from S'. The algorithm is run based on the same trajectory for S and S' by fixing the random seed. We randomly pick 5 different i's and 5 different random seeds (total 25 runs). The step sizes for the discriminator and the generator are chosen as constants, i.e. $\eta = 0.0002$. We compute the Euclidean distance, i.e., Frobenius norm, between the parameters trained on the neighboring datasets. Note that we do not target at optimizing the test accuracy, but give an interpretable visualization to validate our theoretical findings. The results are given in Figure I.1.

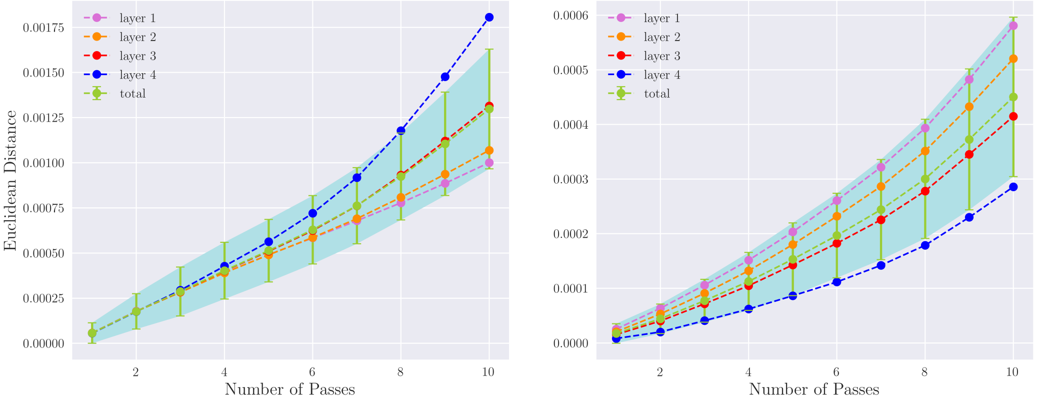

Figure I.1. The parameter distance versus the number of passes. Left: generator, right: discriminator. 'total' is the mean normalized Euclidean distance across all layers and the shaded area is the standard deviation.

It is clear that the parameter distances for both the generator and the discriminator continue to increase during the training process of SGDA, which is consistent with our analysis in Section F.1 and F.3.