Appendix for "Differentially Private SGDA for Minimax Problems"

A Motivating Examples

We provide several examples that can be formulated as a stochastic minimax problem. All these examples have corresponding empirical minimax formulations.

AUC Maximization. Area Under the ROC Curve (AUC) is a widely used measure for binary classification. Optimizing AUC with square loss can be formulated as

$$\min_{\theta \in \Theta} \mathbb{E}_{\mathbf{z},\mathbf{z}'}[(1 - h(\theta; \mathbf{x}) + h(\theta; \mathbf{x}'))^2 | y = 1, y' = -1]$$

where $h : \Theta \times \mathbb{R}^d \to \mathbb{R}$ is the scoring function for the classifier. It has been shown this problem is equivalent to a minimax problem once auxiliary variables $a, b, \mathbf{v} \in \mathbb{R}$ are introduced [Ying et al., 2016].

$$\min_{\theta,a,b} \max_{\mathbf{v}} F(\theta, a, b, c) = \mathbb{E}_{\mathbf{z}}[f(\theta, a, b, \mathbf{v}; \mathbf{z})]$$

where $f = (1-p)(h(\theta;\mathbf{x}) - a)^2 \mathbb{I}[y = 1] + p(h(\theta;\mathbf{x}) - b)^2 \mathbb{I}[y = -1] + 2(1 + \mathbf{v})(ph(\theta;\mathbf{x})\mathbb{I}[y = -1] - (1 - p)h(\theta;\mathbf{x})\mathbb{I}[y = 1]) - p(1-p)\mathbf{v}^2$ and $p = \mathbb{P}[y = 1]$. Such problem is (non)convex-concave. In particular, Liu et al. [2020] showed that when h is a one hidden layer neural network the objective f satisfies the Polyak-Łojasiewicz condition. Differential privacy has been applied to learn private classifier by optimizing AUC [Wang et al., 2021]. The proposed privacy mechanisms there are objective perturbation and output perturbation.

Generative Adversarial Networks (GANs). GAN is introduced in Goodfellow et al. [2014] which can be regarded as a game between a generator network $G_{\mathbf{v}}$ and a discriminator network $D_{\mathbf{w}}$. The generator network produces synthetic data from random noise ξ, while the discriminator network discriminates between the true data and the synthetic data. In particular, a popular variant of GAN named as WGAN [Arjovsky et al., 2017] can be written as a minimax problem

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

Recently Sahiner et al. [2021] showed that WGAN with a two-layer discriminator and generator can be expressed as a convex-concave problem. An heuristic differentially private version of RMSProp were employed to train GANs by Xie et al. [2018]. Recently differential privacy has successfully applied to private synthetic data generation by GAN framework [Jordon et al., 2018, Beaulieu-Jones et al., 2019].

Markov Decision Process (MDP). Let $\mathcal{A}$ be a finite action space. For any $a \in \mathcal{A}$, $P(a) \in [0,1]^{n \times n}$ is the state-transition probability matrix and $\mathbf{r}(a) \in [0,1]^n$ is the vector of expected state-transition rewards. In the infinite-horizon average-reward Markov decision problem, one aims to find a stationary policy π to make an infinite sequence of actions and optimize the average-per-time-step reward $\bar{v}$. By classical theory of dynamics programming [Puterman, 2014], finding an optimal policy is equivalent as solving the fixed-point Bellman equation

$$\bar{v}^* + h_i^* = \max_{a \in \mathcal{A}} \Big\{ \sum_{j=1}^n (p_{ij}(a) h_i^* + p_{ij}(a) r_{ij}(a)) \Big\}, \quad \forall i$$

where $\mathbf{h} \in \mathbb{R}^n$ is the difference-of-value vector. Wang [2017] showed that this problem is equivalent to the minimax problem as follow

$$\min_{\mathbf{h} \in \mathcal{H}} \max_{\mu \in \mathcal{U}} \mu^\top ((P(a) - I)\mathbf{h} + \mathbf{r}(a))$$

where $\mathcal{H}$ and $\mathcal{U}$ are the feasible regions chosen according to the mixing time and stationary distribution. We refer to Zhang et al. [2021] for a discussion on the measure of population risk.

Robust Optimization and Fairness. Let $\mathcal{D}_1, \cdots, \mathcal{D}_m$ be m different distributions on some support. The aim is to minimize the worst population risks L parameterized by some $\mathbf{w}$ among multiple scenarios:

$$\min_{\mathbf{w} \in \mathcal{W}} L(\mathbf{w}) = \max_{1 \le i \le m} \big\{ \mathbb{E}_{\mathbf{z}_1 \sim \mathcal{D}_1}[\ell(\mathbf{w}; \mathbf{z}_1)], \cdots, \mathbb{E}_{\mathbf{z}_m \sim \mathcal{D}_m}[\ell(\mathbf{w}; \mathbf{z}_m)] \big\}$$

This problem can be reformulated as a zero-sum game between two players $\mathbf{w}$ and $\mathbf{v}$ as follow

$$\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{v}\in\Delta_m}\sum_{i=1}^{m} v_i\mathbb{E}_{\mathbf{z}_i\sim\mathcal{D}_i}[\ell(\mathbf{w};\mathbf{z}_i)] = \mathbb{E}\Big[\sum_{i=1}^{m} v_i\ell(\mathbf{w};\mathbf{z}_i)\Big]$$

where $\Delta_m = \big\{\mathbf{v}\in\mathbb{R}^m : v_i \geq 0, \sum_{i=1}^m v_i = 1\big\}$ denotes the m-dimensional simplex. Such robust optimization formulation has been recently proposed to address fairness among subgroups [Mohri et al., 2019] and federated learning on heterogeneous populations [Li et al., 2019].

B Proofs of Theorem 1 and Remark 1

In this section, we prove the privacy guarantee of DP-SGDA based on the privacy-amplification by the subsampling result, which is a direct application of Theorem 1 in Abadi et al. [2016]. First we introduce some necessary definitions.

Definition 1. *Given a function $g : \mathcal{Z}^n \to \mathbb{R}^d$, we say g has $\Delta(g)$ ℓ_2-sensitivity if for any neighboring datasets S, S' we have*

$$\|g(S) - g(S')\|_2 \leq \Delta(g).$$

Definition 2 ([Abadi et al., 2016])**.** *For an (randomized) algorithm A, and neighboring datasets S, S' the λ-th moment is given as*

$$\alpha_A(\lambda, S, S') = \log\mathbb{E}_{O\sim A(S)}\Big[\Big(\frac{\mathbb{P}[A(S)=O]}{\mathbb{P}[A(S')=O]}\Big)^{\lambda}\Big].$$

The moments accountant is then defined as

$$\alpha_A(\lambda) = \sup_{S,S'}\alpha_A(\lambda, S, S').$$

Lemma 1 ([Abadi et al., 2016])**.** *Consider a sequence of mechanisms $\{A_t\}_{t\in[T]}$ and the composite mechanism $A = (A_1, \cdots, A_T)$.*

a) [Composability] For any λ,

$$\alpha_A(\lambda) = \sum_{t=1}^{T}\alpha_{A_t}(\lambda).$$

b) [Tail bound] For any ϵ, the mechanism A is (ϵ, δ) differentially private for

$$\delta = \min_{\lambda}\alpha_A(\lambda) - \lambda\epsilon.$$

Lemma 2 ([Abadi et al., 2016])**.** *Consider a sequence of mechanisms $A_t = g_t(S_t) + \xi_t$ where $\xi \sim \mathcal{N}(0, \sigma^2 I)$. Here each function $g_t : \mathcal{Z}^m \to \mathbb{R}^d$ has ℓ_2-sensitivity of 1. And each S_t is a subsample of size m obtained by uniform sampling without replacement [1] from S, i.e. $S_t \sim (Unif(S))^m$, Then*

$$\alpha_A(\lambda) \leq \frac{m^2 n\lambda(\lambda+1)}{n^2(n-m)\sigma^2} + \mathcal{O}(\frac{m^3\lambda^3}{n^3\sigma^3}).$$

Theorem 1 (Theorem 1 restated)**.** *There exist constants c_1, c_2 and c_3 so that for any $\epsilon < c_1 T/n^2$, Algorithm 1 is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose*

$$\sigma_{\mathbf{w}} \geq \frac{c_2 G_{\mathbf{w}}\sqrt{T\log(1/\delta)}}{n\epsilon} \text{ and } \sigma_{\mathbf{v}} \geq \frac{c_3 G_{\mathbf{v}}\sqrt{T\log(1/\delta)}}{n\epsilon}.$$

[1] In our case we use uniform sampling on each iteration to construct I_t and therefore S_t, as opposed to the Poisson sampling in Abadi et al. [2016]. However, one can verify that similar moment estimates lead to our stated result [Wang et al., 2019]

Proof. Let $S = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ and $S' = \{\mathbf{z}'_1, \cdots, \mathbf{z}'_n\}$ be two neighboring datasets. At iteration t, we first focus on $A_t^{\mathbf{w}} = \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \xi_t$. Since $f(\cdot, \mathbf{v}; \mathbf{z})$ is $G_{\mathbf{w}}$-Lipschitz continuous, it implies for any neighboring datasets S, S',

$$\Big\| \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}'_{i_t^j}) \Big\|_2 \leq \frac{2G_{\mathbf{w}}}{m}.$$

Therefore we can define $g_t(S_t) = \frac{1}{2G_{\mathbf{w}}}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t, \mathbf{z}_{i_t^j})$ such that $\Delta(g_t) = 1$. By Lemma 1 b) and 2, the log moment of the composite mechanism $A^{\mathbf{w}} = (A_1^{\mathbf{w}}, \cdots, A_T^{\mathbf{w}})$ can be bounded as follows

$$\alpha_{A^{\mathbf{w}}}(\lambda) \leq \frac{m^2 T \lambda^2}{n^2 \tilde{\sigma}_{\mathbf{w}}^2}.$$

where $\tilde{\sigma}_{\mathbf{w}} = \sigma_{\mathbf{w}}/2G_{\mathbf{w}}$. Similarly, since $A_t^{\mathbf{v}} = \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t}) + \zeta_t$ has ℓ_2-sensitivity $2G_{\mathbf{v}}/m$, then the log moment of the final output $A = (A_1^{\mathbf{w}}, A_1^{\mathbf{v}}, \cdots, A_T^{\mathbf{w}}, A_T^{\mathbf{v}})$ can be bounded as follows

$$\alpha_A(\lambda) \leq \alpha_{A^{\mathbf{v}}}(\lambda) + \alpha_{A^{\mathbf{w}}}(\lambda) \leq \frac{m^2 T \lambda^2}{n^2 \tilde{\sigma}_{\mathbf{w}}^2} + \frac{m^2 T \lambda^2}{n^2 \tilde{\sigma}_{\mathbf{v}}^2}.$$

By Lemma 1 a), to guarantee A to be (ϵ, δ)-differentially private, it suffices that

$$\frac{\lambda^2 m^2 T}{n^2 \tilde{\sigma}_{\mathbf{w}}^2} \leq \frac{\lambda\epsilon}{4}, \frac{\lambda^2 m^2 T}{n^2 \tilde{\sigma}_{\mathbf{v}}^2} \leq \frac{\lambda\epsilon}{4}, \exp(-\frac{\lambda\epsilon}{4}) \leq \delta, \lambda \leq \tilde{\sigma}_{\mathbf{w}}^2 \log(\frac{n}{m\tilde{\sigma}_{\mathbf{w}}}) \text{ and } \lambda \leq \tilde{\sigma}_{\mathbf{v}}^2 \log(\frac{n}{m\tilde{\sigma}_{\mathbf{v}}})$$

It is now easy to verify that when $\epsilon = c_1 m^2 T/n^2$, we can satisfy all these conditions by setting

$$\tilde{\sigma}_{\mathbf{w}} \geq \frac{c_2\sqrt{T\log(1/\delta)}}{n\epsilon} \text{ and } \tilde{\sigma}_{\mathbf{v}} \geq \frac{c_3\sqrt{T\log(1/\delta)}}{n\epsilon}$$

for some explicit constants c_1, c_2 and c_3. The proof is complete. □

Proof of Remark 1. Without loss of generality, we consider with only one σ in the the proof of Theorem 1. Then algorithm A is guaranteed to be (ϵ, δ)-DP if one can find $\lambda > 0$ such that

$$\frac{\lambda^2 m^2 T}{n^2 \sigma^2} \leq \frac{\lambda\epsilon}{2}, \exp(-\frac{\lambda\epsilon}{2}) \leq \delta, \text{ and } \lambda \leq \sigma^2 \log(\frac{n}{m\sigma})$$

Given $\delta = \frac{1}{n^2}$, the second inequality can be reformulated as $\lambda \geq \frac{4\log(n)}{\epsilon}$. Therefore by choosing $\sigma^2 = \frac{8m^2 T\log(n)}{n^2\epsilon^2}$, the first inequality becomes $\lambda \leq \frac{4\log(n)}{\epsilon}$, indicating $\lambda = \frac{4\log(n)}{\epsilon}$. It suffices to show such choice of λ satisfies the third inequality, which is straightforward by the choice of m and $\epsilon \leq 1$. The proof is complete. □

C Proofs for the convex-concave setting in Section 3.1

Recall that the error decomposition (4) given in Section 3.1 that the weak PD risk can be decomposed as follows:

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

where the term $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle_S^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ is the generalization error and the term $\triangle_S^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ is the optimization error.
The proof of Theorem 2 involves the estimation of the optimization error and generalization error which are performed in the subsequent subsection, respectively.

C.1 Estimation of Optimization Error

We start by studying the optimization error for Algorithm 1. This is obtained as a direct corollary of Nemirovski et al. [2009], with the existence of the Gaussian noise's variance and the mini-batch. Recall that $d = \max\{d_1, d_2\}$.

Lemma 3. *Suppose **(A1)** holds, and F_S is convex-concave. Let the stepsizes $\eta_{\mathbf{w},t} = \eta_{\mathbf{v},t} = \eta$, $t \in [T]$ for some $\eta > 0$. Then Algorithm 1 satisfies*

$$\sup_{\mathbf{v}\in\mathcal{V}} \mathbb{E}_A[F_S(\bar{\mathbf{w}}_T, \mathbf{v})] - \inf_{\mathbf{w}\in\mathcal{W}} \mathbb{E}_A[F_S(\mathbf{w}, \bar{\mathbf{v}}_T)] \le \frac{\eta(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)}{2} + \frac{D_{\mathbf{w}}^2 + D_{\mathbf{v}}^2}{\eta T} + \frac{(D_{\mathbf{w}}G_{\mathbf{w}} + D_{\mathbf{v}}G_{\mathbf{v}})}{\sqrt{mT}} + \eta d(\sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{v}}^2).$$

Proof. According to the non-expansiveness of projection and update rule of Algorithm 1, for any $\mathbf{w} \in \mathcal{W}$, we have

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \le& \Big\|\mathbf{w}_t - \mathbf{w} - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \eta\xi_t\Big\|_2^2 \\
\le& \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 2\eta\Big\langle \mathbf{w} - \mathbf{w}_t, \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \xi_t\Big\rangle + \eta^2\Big\|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j})\Big\|_2^2 + \eta^2\|\xi_t\|_2^2 \\
&+ 2\eta^2\Big\langle \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}), \xi_t\Big\rangle \\
\le& \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 2\eta\langle \mathbf{w} - \mathbf{w}_t, \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle + 2\eta\Big\langle \mathbf{w} - \mathbf{w}_t, \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\Big\rangle \\
&+ \eta^2 G_{\mathbf{w}}^2 + \eta^2\|\xi_t\|_2^2 + 2\eta^2\Big\langle \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}), \xi_t\Big\rangle + 2\eta\langle \mathbf{w} - \mathbf{w}_t, \xi_t\rangle,
\end{aligned}$$

where in the last inequality we have used $f(\cdot, \mathbf{v}_t, \mathbf{z}_{i_t^j})$ is $G_{\mathbf{w}}$-Lipschitz continuous. According to the convexity of $F_S(\cdot, \mathbf{v}_t)$ we know

$$\begin{aligned}
2\eta(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)) \le& \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + 2\eta\Big\langle \mathbf{w} - \mathbf{w}_t, \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\Big\rangle \\
&+ \eta^2 G_{\mathbf{w}}^2 + \eta^2\|\xi_t\|_2^2 + 2\eta^2\Big\langle \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}), \xi_t\Big\rangle + 2\eta\langle \mathbf{w} - \mathbf{w}_t, \xi_t\rangle.
\end{aligned}$$

Taking a summation of the above inequality from $t = 1$ to T we derive

$$\begin{aligned}
2\eta\sum_{t=1}^T (F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \mathbf{v}_t)) \le& \|\mathbf{w}_1 - \mathbf{w}\|_2^2 + 2\eta\sum_{t=1}^T \Big\langle \mathbf{w} - \mathbf{w}_t, \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\Big\rangle \\
&+ T\eta^2 G_{\mathbf{w}}^2 + \eta^2\sum_{t=1}^T \|\xi_t\|_2^2 + 2\eta^2\sum_{t=1}^T \Big\langle \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}), \xi_t\Big\rangle + 2\eta\langle \mathbf{w} - \mathbf{w}_t, \xi_t\rangle.
\end{aligned}$$

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

$$\begin{aligned}
2\sum_{t=1}^T \eta(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}, \bar{\mathbf{v}}_T)) \le& 2D_{\mathbf{w}}^2 - 2\eta\sum_{t=1}^T \Big\langle \mathbf{w}_t, \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\Big\rangle \\
&+ 2D_{\mathbf{w}}\eta\Big\|\sum_{t=1}^T (\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\Big\|_2 \\
&+ T\eta^2 G_{\mathbf{w}}^2 + \eta^2\sum_{t=1}^T \|\xi_t\|_2^2 + 2\eta^2\sum_{t=1}^T \Big\langle \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}), \xi_t\Big\rangle + 2\eta\langle \mathbf{w} - \mathbf{w}_t, \xi_t\rangle. \quad (2)
\end{aligned}$$

We can take expectations on the randomness of A over both sides of(2) and get

$$2\eta\sum_{t=1}^{T}\mathbb{E}_A[F_S(\mathbf{w}_t,\mathbf{v}_t)-F_S(\mathbf{w},\bar{\mathbf{v}}_T)] \leq 2D_{\mathbf{w}}^2+2D_{\mathbf{w}}\eta\mathbb{E}_A\Big[\Big\|\sum_{t=1}^{T}\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\Big\|_2\Big] + T\eta^2G_{\mathbf{w}}^2+\eta^2 d_1\sigma_{\mathbf{w}}^2,$$

where we used that the variance $\mathbb{E}_A[\|\xi_t\|_2^2]=d_1\sigma_{\mathbf{w}}^2$, the unbiasedness $\mathbb{E}_A[\langle\mathbf{w}_t,\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\rangle]=0$, the independence $\mathbb{E}_A[\langle\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j}),\xi_t\rangle]=0$ and $\mathbb{E}_A[\langle\mathbf{w}-\mathbf{w}_t,\xi_t\rangle]=0$. Since the above inequality holds for all $\mathbf{w}$, we further get

$$2\eta\sum_{t=1}^{T}\mathbb{E}_A[F_S(\mathbf{w}_t,\mathbf{v}_t)]-\inf_{\mathbf{w}\in\mathcal{W}}\mathbb{E}_A[F_S(\mathbf{w},\bar{\mathbf{v}}_T)] \leq 2D_{\mathbf{w}}^2+2D_{\mathbf{w}}\eta\mathbb{E}_A\Big[\Big\|\sum_{t=1}^{T}\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\Big\|_2\Big] + T\eta^2G_{\mathbf{w}}^2+\eta^2 d_1\sigma_{\mathbf{w}}^2, \quad (3)$$

According to Jensen's inequality and $G_{\mathbf{w}}$-Lipschitz continuity we further derive

$$\begin{aligned}&\Big(\mathbb{E}_A\Big[\Big\|\sum_{t=1}^{T}(\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\Big\|_2)\Big]\Big)^2\\ \leq&\mathbb{E}_A\Big[\Big\|\sum_{t=1}^{T}(\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t))\Big\|_2^2\Big]=\sum_{t=1}^{T}\mathbb{E}_A\Big[\Big\|\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\Big\|_2^2\Big]\\ \leq&\frac{TG_{\mathbf{w}}^2}{m}.\end{aligned}$$

Plugging the above estimate into (3) we arrive

$$2\eta\sum_{t=1}^{T}\mathbb{E}_A[F_S(\mathbf{w}_t,\mathbf{v}_t)]-\inf_{\mathbf{w}\in\mathcal{W}}\mathbb{E}_A[F_S(\mathbf{w},\bar{\mathbf{v}}_T)]\leq 2D_{\mathbf{w}}^2+\frac{2D_{\mathbf{w}}\eta G_{\mathbf{w}}\sqrt{T}}{\sqrt{m}}+T\eta^2G_{\mathbf{w}}^2+T\eta^2d_1\sigma_{\mathbf{w}}^2.$$

By dividing $2\eta T$ on both sides we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_A[F_S(\mathbf{w}_t,\mathbf{v}_t)]-\inf_{\mathbf{w}\in\mathcal{W}}\mathbb{E}_A[F_S(\mathbf{w},\bar{\mathbf{v}}_T)]\leq\frac{D_{\mathbf{w}}^2}{\eta T}+\frac{D_{\mathbf{w}}G_{\mathbf{w}}}{\sqrt{mT}}+\frac{\eta G_{\mathbf{w}}^2}{2}+\frac{\eta d_1\sigma_{\mathbf{w}}^2}{2}. \quad (4)$$

In a similar way, we can show that

$$\frac{1}{T}\sum_{t=1}^{T}\sup_{\mathbf{v}\in\mathcal{V}}\mathbb{E}_A[F_S(\bar{\mathbf{w}}_T,\mathbf{v})]-\mathbb{E}_A[F_S(\mathbf{w}_t,\mathbf{v}_t)]\leq\frac{D_{\mathbf{v}}^2}{\eta T}+\frac{D_{\mathbf{v}}G_{\mathbf{v}}}{\sqrt{mT}}+\frac{\eta G_{\mathbf{v}}^2}{2}+\frac{\eta d_2\sigma_{\mathbf{v}}^2}{2}. \quad (5)$$

The stated bound then follows from (4) and (5) and the fact that $d=\max\{d_1,d_2\}$. □

C.2 Estimation of Generalization Error

Next we move on to the generalization error. Firstly, we introduce a lemma that bridges the generalization and the stability. We say the randomized algorithm A is *ε-weakly-stable* if, for any neighboring datasets S,S', there holds

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

Lemma 4. *[Lei et al., 2021] If A is ε-weakly-stable, then there holds*

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

We also need the following standard lemma before we prove the stability of DP-SGDA.

Lemma 5 ([Rockafellar, 1976])**.** *Let f be a convex-concave function. Then*

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

The stability analysis is given in the following lemma. This lemma is an extension of the uniform argument stability results in Lei et al. [2021] to the case of mini-batch DP-SGDA.

Lemma 6. *Suppose the function F_S is convex-concave. Let the stepsizes $\eta_{\mathbf{w},t} = \eta_{\mathbf{v},t} = \eta$ for some $\eta > 0$.*

*a) Assume **(A1)** and **(A3)** hold, then Algorithm 1 satisfies*

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \leq \frac{4\sqrt{e(T + T^2/n)}(G_{\mathbf{w}} + G_{\mathbf{v}})^2 \eta \exp(L^2 T \eta^2/2)}{\sqrt{n}}.$$

*b) Assume **(A1)** holds, then Algorithm 1 satisfies*

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

Proof. Without loss of generality, let $S = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}, S' = \{\mathbf{z}'_1, \cdots, \mathbf{z}'_n\}$ be neighboring datasets differing by the last element, i.e. $\mathbf{z}_n \neq \mathbf{z}'_n$. Let $\{\mathbf{w}_t, \mathbf{v}_t\}, \{\mathbf{w}'_t, \mathbf{v}'_t\}$ be the sequence produced by Algorithm 1 w.r.t. S and S', respectively. We first prove Part a). In the case $n \notin I_t$, by the non-expansiveness of projection, we have

$$\begin{aligned}
&\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} \mathbf{w}_t - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) - \eta\xi_t - \mathbf{w}'_t + \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t^j}) + \eta\xi_t \\ \mathbf{v}_t + \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) + \eta\zeta_t - \mathbf{v}'_t - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t^j}) - \eta\zeta_t \end{pmatrix} \right\|_2^2 \\
&= \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + \frac{\eta}{m}\sum_{j=1}^m \left\langle \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix}, \begin{pmatrix} \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t^j}) \\ \nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t^j}) - \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) \end{pmatrix} \right\rangle \\
&+ \left\| \begin{pmatrix} \frac{\eta}{m}\sum_{j=1}^m (\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_n)) \\ \frac{\eta}{m}\sum_{j=1}^m (\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_n)) \end{pmatrix} \right\|_2^2 \\
&\leq (1 + L^2\eta^2) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2,
\end{aligned}$$

where the last inequality follows from Lemma 5 and the L-smoothness assumption. If $n \in I_t$, then it follows that

$$\begin{aligned}
&\left\| \begin{pmatrix} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} \mathbf{w}_t - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) - \eta\xi_t - \mathbf{w}'_t + \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t^j}) + \eta\xi_t \\ \mathbf{v}_t + \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) + \eta\zeta_t - \mathbf{v}'_t - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t^j}) - \eta\zeta_t \end{pmatrix} \right\|_2^2 \\
&\leq \frac{1}{m}\sum_{i_t^j \in I_t, i_t^j \neq n} \left\| \begin{pmatrix} \mathbf{w}_t - \eta\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) - \mathbf{w}'_t + \eta\nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t^j}) \\ \mathbf{v}_t + \eta\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) - \mathbf{v}'_t - \eta\nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_{i_t^j}) \end{pmatrix} \right\|_2^2 \\
&+ \frac{1}{m} \left\| \begin{pmatrix} \mathbf{w}_t - \eta\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \mathbf{w}'_t + \eta\nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_n) \\ \mathbf{v}_t + \eta\nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_n) - \mathbf{v}'_t - \eta\nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z'_n) \end{pmatrix} \right\|_2^2 \\
&\leq \frac{m-1}{m}(1 + L^2\eta^2) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 + \frac{1+p}{m} \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|_2^2 \\
&+ \frac{1 + 1/p}{m}\eta^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, \qquad (6)
\end{aligned}$$

where in the last inequality we used the elementary inequality $(a+b)^2 \le (1+p)a^2 + (1+1/p)b^2$ $(p > 0)$. Since I_t are drawn uniformly at random with replacement, the event $n \notin I_t$ happens with probability $1 - m/n$ and the event $n \in I_t$ happens with probability m/n. Therefore, we know

$$\begin{aligned}
\mathbb{E}_{i_t}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] &\le \frac{(n-m)(1+L^2\eta^2)}{n}\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{m(1+L^2\eta^2)}{n}\frac{m-1}{m}\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t\end{pmatrix}\right\|_2^2 \\
&\quad + \frac{m}{n}\frac{1+p}{m}\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{m}{n}\frac{4(1+1/p)}{m}\eta^2(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2) \\
&\le \Big(1 + L^2\eta^2 + p/n\Big)\left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t\end{pmatrix}\right\|_2^2 + \frac{4(1+1/p)}{n}\eta^2(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2).
\end{aligned}$$

Applying this inequality recursively, we derive

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \le \frac{4(1+1/p)}{n}(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)\sum_{k=1}^t \eta^2 \prod_{j=k+1}^t \Big(1 + L^2\eta^2 + p/n\Big).$$

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

$$\begin{aligned}
\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] &\le \frac{4(1+1/p)}{n}(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)\sum_{k=1}^t \eta^2 \prod_{j=k+1}^t \exp\Big(L^2\eta^2 + p/n\Big) \\
&= \frac{4(1+1/p)}{n}(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)\sum_{k=1}^t \eta^2 \exp\Big(L^2\sum_{j=k+1}^t \eta^2 + p(t-k)/n\Big) \\
&\le \frac{4(1+1/p)}{n}(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)\exp\Big(L^2\sum_{j=1}^t \eta^2 + pt/n\Big)\sum_{k=1}^t \eta^2.
\end{aligned}$$

By taking $p = n/t$ we get

$$\mathbb{E}_A\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \le \frac{4e(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)(1+t/n)}{n}\exp\Big(L^2\sum_{j=1}^t \eta^2\Big)\sum_{k=1}^t \eta^2.$$

Now by the Lipschitz continuity and Jensen's inequality we ave

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

According to Lemma 4 we know

$$\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle_S^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \le \frac{4\sqrt{e(T + T^2/n)}(G_{\mathbf{w}} + G_{\mathbf{v}})^2\eta\exp(L^2T\eta^2/2)}{\sqrt{n}}.$$

Next we focus on Part b). We consider two cases at the t-th iteration. If $n \notin I_t$, then analogous to the discussions in Lei et al. [2021] we can show

$$\begin{aligned}
\left\|\begin{pmatrix}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 &\le \left\|\begin{pmatrix}\mathbf{w}_t - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) - \eta\xi_t - \mathbf{w}'_t + \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t^j}) + \eta\xi_t \\ \mathbf{v}_t + \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t^j}) + \eta\zeta_t - \mathbf{v}'_t - \frac{\eta}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}'_t, \mathbf{v}'_t; z_{i_t^j}) - \eta\zeta_t\end{pmatrix}\right\|_2^2 \\
&\le \left\|\begin{pmatrix}\mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t\end{pmatrix}\right\|_2^2 + 4(G_{\mathbf{w}}^2 + G_{\mathbf{v}}^2)\eta^2. \qquad (7)
\end{aligned}$$

Combining the preceding inequality with (6) and using the probability of $n \notin I_t$, we derive

$$
\begin{aligned}
&\mathbb{E}_{i_t}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \le \frac{n-1}{n}\left(\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2 + 4(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2\right)\\
&+\frac{1+p}{n}\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2 + \frac{4(1+1/p)}{n}(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2\\
&=(1+p/n)\left\|\begin{pmatrix}\mathbf{w}_{t}-\mathbf{w}'_{t}\\ \mathbf{v}_{t}-\mathbf{v}'_{t}\end{pmatrix}\right\|_2^2 + 4(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2(1+1/(np)).
\end{aligned}
$$

Applying this inequality recursively implies that

$$
\begin{aligned}
&\mathbb{E}_{A}\left[\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2\right] \le 4(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2\big(1+1/(np)\big)\sum_{k=1}^{t}\Big(1+\frac{p}{n}\Big)^{t-k}\\
&=4(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2\Big(1+\frac{1}{np}\Big)\frac{n}{p}\Big(\Big(1+\frac{p}{n}\Big)^t-1\Big) = 4(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2\Big(\frac{n}{p}+\frac{1}{p^2}\Big)\Big(\Big(1+\frac{p}{n}\Big)^t-1\Big).
\end{aligned}
$$

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

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

Now by the Lipschitz continuity and Jensen's inequality we ave

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

According to Lemma 4 we know

$$
\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) \le 32(G_{\mathbf{w}}+G_{\mathbf{v}})^2\eta^2\Big(\sqrt{T}+\frac{T}{n}\Big).
$$

□

C.3 Proof of Theorem 2

Finally we are ready to present the proof of Theorem 2.

Theorem 2 (Theorem 2 restated)**.** *Suppose the function F_S is convex-concave. Let the stepsizes $\eta_{\mathbf{w},t} = \eta_{\mathbf{v},t} = \eta$, $t = [T]$ for some $\eta > 0$.*

*a) Assume **(A1)** and **(A3)** hold. If we choose $T \asymp n$ and $\eta \asymp 1/\Big(\sqrt{L}\max\{\sqrt{n},\sqrt{d\log(1/\delta)}/\epsilon\}\Big)$, then Algorithm 1 satisfies*

$$
\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) = \mathcal{O}\Big(\max\{G_{\mathbf{w}}^2+G_{\mathbf{v}}^2,(G_{\mathbf{w}}+G_{\mathbf{v}})^2,D_{\mathbf{w}}^2+D_{\mathbf{v}}^2,D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}\}\max\Big\{\frac{1}{\sqrt{n}},\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big\}\Big).
$$

*b) Assume **(A1)** holds. If we choose $T \asymp n^2$ and $\eta \asymp 1/\Big(n\max\{\sqrt{n},\sqrt{d\log(1/\delta)}/\epsilon\}\Big)$, then Algorithm 1 satisfies*

$$
\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T) = \mathcal{O}\Big(\max\{G_{\mathbf{w}}^2+G_{\mathbf{v}}^2,(G_{\mathbf{w}}+G_{\mathbf{v}})^2,D_{\mathbf{w}}^2+D_{\mathbf{v}}^2,D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}\}\max\Big\{\frac{1}{\sqrt{n}},\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big\}\Big).
$$

Proof of Theorem 2. We first focus on Part a). According to Part a) of Lemma 6 we know

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)-\triangle_S^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)\le\frac{4\sqrt{e(T+T^2/n)}(G_{\mathbf{w}}+G_{\mathbf{v}})^2\eta\exp(L^2T\eta^2/2)}{\sqrt{n}}$$

and by Lemma 3 we know

$$\triangle_S^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)\le\frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)}{2}+\frac{D_{\mathbf{w}}^2+D_{\mathbf{v}}^2}{2\eta T}+\frac{D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}}{\sqrt{mT}}+\eta d(\sigma_{\mathbf{w}}^2+\sigma_{\mathbf{v}}^2).$$

Combining the above two quantities we have

$$\begin{aligned}\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)\le&\frac{4\sqrt{e(T+T^2/n)}(G_{\mathbf{w}}+G_{\mathbf{v}})^2\eta\exp(L^2T\eta^2/2)}{\sqrt{n}}+\frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)}{2}+\frac{D_{\mathbf{w}}^2+D_{\mathbf{v}}^2}{2\eta T}\\&+\frac{D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}}{\sqrt{mT}}+\eta d(\sigma_{\mathbf{w}}^2+\sigma_{\mathbf{v}}^2).\end{aligned}\tag{8}$$

Furthermore, by Theorem 1, we know

$$\sigma_{\mathbf{w}}^2=\mathcal{O}\Big(\frac{G_{\mathbf{w}}^2T\log(1/\delta)}{n^2\epsilon^2}\Big),\quad\sigma_{\mathbf{v}}^2=\mathcal{O}\Big(\frac{G_{\mathbf{v}}^2T\log(1/\delta)}{n^2\epsilon^2}\Big).$$

Plugging it back into (8) we have

$$\begin{aligned}\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)=\mathcal{O}\Big(&\frac{\sqrt{(T+T^2/n)}(G_{\mathbf{w}}+G_{\mathbf{v}})^2\eta\exp(L^2T\eta^2)}{\sqrt{n}}\\&+\frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)}{2}+\frac{D_{\mathbf{w}}^2+D_{\mathbf{v}}^2}{2\eta T}+\frac{D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}}{\sqrt{mT}}+\frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)Td\log(1/\delta)}{n^2\epsilon^2}\Big).\end{aligned}$$

By picking $T\asymp n$ and $\eta\asymp1/\Big(L\max\{\sqrt{n},\sqrt{d\log(1/\delta)}/\epsilon\}\Big)$ we have $\exp(L^2T\eta^2)=\mathcal{O}\Big(\min\{1,\frac{n\epsilon^2}{d\log(1/\delta)}\}\Big)=\mathcal{O}(1)$ and

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)=\mathcal{O}\Big(\max\{G_{\mathbf{w}}^2+G_{\mathbf{v}}^2,(G_{\mathbf{w}}+G_{\mathbf{v}})^2,D_{\mathbf{w}}^2+D_{\mathbf{v}}^2,D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}\}\max\Big\{\frac{1}{\sqrt{n}},\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big\}\Big).$$

We now turn to Part b). According to Lemma 6 Part b) we know

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)-\triangle_S^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)\le4\sqrt{2}\eta(G_{\mathbf{w}}+G_{\mathbf{v}})^2\Big(\sqrt{T}+\frac{T}{n}\Big).$$

Similar to Part a) we have

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)=\mathcal{O}\Big(\eta(G_{\mathbf{w}}+G_{\mathbf{v}})^2\Big(\sqrt{T}+\frac{T}{n}\Big)+\frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)}{2}+\frac{D_{\mathbf{w}}^2+D_{\mathbf{v}}^2}{2\eta T}+\frac{D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}}{\sqrt{mT}}+\frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)Td\log(1/\delta)}{n^2\epsilon^2}\Big).$$

By picking $T\asymp n^2$ and $\eta\asymp1/\Big(n\max\{\sqrt{n},\sqrt{d\log(1/\delta)}/\epsilon\}\Big)$ we have

$$\triangle^w(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)=\mathcal{O}\Big(\max\{G_{\mathbf{w}}^2+G_{\mathbf{v}}^2,(G_{\mathbf{w}}+G_{\mathbf{v}})^2,D_{\mathbf{w}}^2+D_{\mathbf{v}}^2,D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}\}\max\Big\{\frac{1}{\sqrt{n}},\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big\}\Big).$$

The proof is complete. □

D Proofs for the nonconvex-strongly-concave setting in Section 3.2

In this section, we will provide the proofs for the theorems in Section 3.2. Recall that we define $R_S^* = \min_{\mathbf{w}\in\mathcal{W}} R_S(\mathbf{w})$, and $R^* = \min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})$. Then, for any $\mathbf{w}^* \in \arg\min_{\mathbf{w}} R(\mathbf{w})$ we have the error decomposition:

$$\begin{aligned}\mathbb{E}[R(\mathbf{w}_T) - R^*] =& \mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] + \mathbb{E}[R_S(\mathbf{w}_T) - R_S^*] + \mathbb{E}[R_S^* - R_S(\mathbf{w}^*)] + \mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)] \\ \leq& \mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] + \mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)] + \mathbb{E}[R_S(\mathbf{w}_T) - R_S^*].\end{aligned}$$

The term $\mathbb{E}[R_S(\mathbf{w}_T) - R_S^*]$ is the *optimization error* which characterizes the discrepancy between the primal empirical risk of an output of Algorithm 1 and the least possible one. The term $\mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] + \mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)]$ is called the *generalization error* which measures the discrepancy between the primal population risk and the empirical one. The estimations for these two errors are described as follows.

D.1 Proof of Theorem 3

To prove Theorem 3, i.e., optimization error, we introduce several necessary lemmas. The first lemma is an application of Danskin's Theorem.

Lemma 7 ([Lin et al., 2020])**.** *Assume **(A3)** holds and $F_S(\mathbf{w},\cdot)$ is ρ-strongly concave. Assume $\mathcal{V}$ is a convex and bounded set. Then the function $R_S(\mathbf{w})$ is $L + L^2/\rho$-smooth and $\nabla R_S(\mathbf{w}) = \nabla_{\mathbf{w}} F_S(\mathbf{w}, \hat{\mathbf{v}}_S(\mathbf{w}))$, where $\hat{\mathbf{v}}_S(\mathbf{w}) = \arg\max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w},\mathbf{v})$. And $\hat{\mathbf{v}}_S(\mathbf{w})$ is L/ρ Lipschitz continuous.*

The second lemma shows that R_S also satisfies the PL condition whenever F_S does.

Lemma 8. *Assume **(A3)** holds. Assume $F_S(\cdot,\mathbf{v})$ satisfies PL condition with constant μ and $F_S(\mathbf{w},\cdot)$ is ρ-strongly concave. Then the function $R_S(\mathbf{w})$ satisfies the PL condition with μ.*

Proof. From Lemma 7, $\|\nabla R_S(\mathbf{w})\|_2^2 = \|\nabla_{\mathbf{w}} F_S(\mathbf{w}, \hat{\mathbf{v}}_S(\mathbf{w}))\|_2^2$. Since F_S satisfies PL condition with constant μ, we get

$$\|\nabla R_S(\mathbf{w})\|_2^2 \geq 2\mu\big(F_S(\mathbf{w}, \hat{\mathbf{v}}_S(\mathbf{w})) - \min_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}', \hat{\mathbf{v}}_S(\mathbf{w}))\big). \tag{9}$$

Also, since $F_S(\mathbf{w}', \hat{\mathbf{v}}_S(\mathbf{w})) \leq \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}',\mathbf{v})$, we have

$$\min_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}', \hat{\mathbf{v}}_S(\mathbf{w})) \leq \min_{\mathbf{w}'\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}',\mathbf{v}) = \min_{\mathbf{w}'\in\mathcal{W}} R_S(\mathbf{w}') \tag{10}$$

Combining equation (9) and (10), we have

$$\|\nabla R_S(\mathbf{w})\|_2^2 \geq 2\mu\big(R_S(\mathbf{w}) - \min_{\mathbf{w}'\in\mathcal{W}} R_S(\mathbf{w}')\big).$$

The proof is complete. ☐

Now we present two key lemmas for the convergence analysis. The next lemma characterizes the descent behavior of $R_S(\mathbf{w}_t)$.

Lemma 9. *Assume **(A2)** and **(A3)** hold. Assume $F_S(\cdot,\mathbf{v})$ satisfies the μ-PL condition and $F_S(\mathbf{w},\cdot)$ is ρ-strongly concave. For Algorithm 1, the iterates $\{\mathbf{w}_t,\mathbf{v}_t\}_{t\in[T]}$ satisfies the following inequality*

$$\begin{aligned}\mathbb{E}[R_S(\mathbf{w}_{t+1}) - R_S^*] \leq& (1-\mu\eta_{\mathbf{w},t})\mathbb{E}[R_S(\mathbf{w}_t) - R_S^*] + \frac{L^2\eta_{\mathbf{w},t}}{2}\mathbb{E}[\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \mathbf{v}_t\|_2^2] \\ &+ \frac{(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{2}\Big(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2\Big).\end{aligned}$$

Proof. Because R_S is $L + L^2/\rho$-smooth by Lemma 7, we have

$$\begin{aligned}R_S(\mathbf{w}_{t+1}) - R_S^* \leq& R_S(\mathbf{w}_t) - R_S^* + \langle \nabla R_S(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{L+L^2/\rho}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ =& R_S(\mathbf{w}_t) - R_S^* - \eta_{\mathbf{w},t}\langle \nabla R_S(\mathbf{w}_t), \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j}) + \xi_t\rangle \\ &+ \frac{(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{2}\|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j}) + \xi_t\|_2^2.\end{aligned}$$

We denote $\mathbb{E}_t$ as the conditional expectation of given $\mathbf{w}_t$ and $\mathbf{v}_t$. Taking this conditional expectation of both sides, we get

$$\begin{aligned}\mathbb{E}_t[R_S(\mathbf{w}_{t+1}) - R_S^*] =& R_S(\mathbf{w}_t) - R_S^* - \eta_{\mathbf{w},t}\langle \nabla R_S(\mathbf{w}_t), \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle \\ &+ \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2} \|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) + \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t) - \xi_t\|_2^2 \\ \le& R_S(\mathbf{w}_t) - R_S^* - \eta_{\mathbf{w},t}\langle \nabla R_S(\mathbf{w}_t), \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle \\ &+ \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2} \|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2} (\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\ \le& R_S(\mathbf{w}_t) - R_S^* - \frac{\eta_{\mathbf{w},t}}{2}\|\nabla R_S(\mathbf{w}_t)\|_2^2 + \frac{\eta_{\mathbf{w},t}}{2}\|\nabla R_S(\mathbf{w}_t) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\ &+ \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2} (\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2),\end{aligned}$$

where in first inequality since $\mathbb{E}_t[\|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2] = \frac{1}{m}\sum_{j=1}^m \mathbb{E}_t[\|\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2] \le \frac{B_{\mathbf{w}}^2}{m}$ and $\mathbb{E}_t[\|\xi_t\|_2^2] = d_1\sigma_{\mathbf{w}}^2 \le d\sigma_{\mathbf{w}}^2$, and the last inequality we use $\eta_{\mathbf{w}} \le 1/(L + L^2/\rho)$. Because R_S satisfies PL condition with μ by Lemma 8, we have

$$\begin{aligned}\mathbb{E}_t[R_S(\mathbf{w}_{t+1}) - R_S^*] \le& (1 - \mu\eta_{\mathbf{w},t})(R_S(\mathbf{w}_t) - R_S^*) + \frac{\eta_{\mathbf{w},t}}{2}\|\nabla R_S(\mathbf{w}_t) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 \\ &+ \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2} (\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\ \le& (1 - \mu\eta_{\mathbf{w},t})(R_S(\mathbf{w}_t) - R_S^*) + \frac{L^2\eta_{\mathbf{w},t}}{2}\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2} (\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2),\end{aligned}$$

where the second we use F_S is L-smooth. Now taking expectation of both sides yields the claimed bound. The proof is complete. □

The next lemma characterizes the descent behavior of $\mathbf{v}_t$.

Lemma 10. *Assume **(A2)** and **(A3)** hold. Assume $F_S(\cdot, \mathbf{v})$ satisfies PL condition with constant μ and $F_S(\mathbf{w}, \cdot)$ is ρ-strongly concave. Let $\hat{\mathbf{v}}_S(\mathbf{w}) = \arg\max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}, \mathbf{v})$. For Algorithm 1 and any $\epsilon > 0$, the iterates $\{\mathbf{w}_t, \mathbf{v}_t\}$ satisfies the following inequality*

$$\begin{aligned}\mathbb{E}[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_{t+1})\|_2^2] \le& ((1+\frac{1}{\epsilon})2L^4/\rho\eta_{\mathbf{w},t}^2 + (1+\epsilon)(1-\rho\eta_{\mathbf{v},t}))\mathbb{E}[\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] + (1+\frac{1}{\epsilon})\eta_{\mathbf{w},t}^2 L^2/\rho^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\ &+ (1+\frac{1}{\epsilon})4L^2/\rho^2(L + L^2/\rho)\eta_{\mathbf{w},t}^2\mathbb{E}[R_S(\mathbf{w}_t) - R_S^*] + (1+\epsilon)\eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2).\end{aligned}$$

Proof. By Young's inequality, we have

$$\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_{t+1})\|_2^2 \le (1+\epsilon)\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + (1+\frac{1}{\epsilon})\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \hat{\mathbf{v}}_S(\mathbf{w}_{t+1})\|_2^2.$$

For the term $\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \hat{\mathbf{v}}_S(\mathbf{w}_{t+1})\|_2^2$, since $\hat{\mathbf{v}}_S(\cdot)$ is L/ρ-Lipschitz by Lemma 7, taking conditional expectation, we have

$$\begin{aligned}\mathbb{E}_t[\|\hat{\mathbf{v}}_S(\mathbf{w}_{t+1}) - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] &\le L^2/\rho^2\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2] = L^2/\rho^2\eta_{\mathbf{w},t}^2\mathbb{E}_t[\|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \xi_t\|_2^2] \\ &\le L^2/\rho^2\eta_{\mathbf{w},t}^2\|\nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\ &\le 2L^2/\rho^2\eta_{\mathbf{w},t}^2\|\nabla R_S(\mathbf{w}_t) - \nabla_{\mathbf{w}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + 2L^2/\rho^2\eta_{\mathbf{w},t}^2\|\nabla R_S(\mathbf{w}_t)\|_2^2 + L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\ &\le 2L^4/\rho^2\eta_{\mathbf{w},t}^2\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + 2L^2/\rho^2\eta_{\mathbf{w},t}^2\|\nabla R_S(\mathbf{w}_t)\|_2^2 + L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2),\end{aligned}$$

where the last step uses the fact that F_S is L-smooth. Because R_S is $L+L^2/\rho$-smooth by Lemma 7 we have $\frac{1}{2(L+L^2\rho)}\|\nabla R_S(\mathbf{w}_t)\|_2^2 \leq R_S(\mathbf{w}_t) - R_S^*$. Therefore

$$\mathbb{E}_t[\|\hat{\mathbf{v}}_S(\mathbf{w}_{t+1}) - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] \leq 2L^4/\rho^2\eta_{\mathbf{w},t}^2\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + 4L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2(R_S(\mathbf{w}_t) - R_S(\mathbf{w}^*)) + L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2). \tag{11}$$

For the term $\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2$, by the contraction of projection, we have

$$\mathbb{E}_t[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] \leq \mathbb{E}_t[\|\mathbf{v}_t + \eta_{\mathbf{v},t}(\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \zeta_t) - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2]$$

$$\leq \|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + 2\eta_{\mathbf{v},t}\mathbb{E}_t[\langle \mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t), \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j})\rangle] + \eta_{\mathbf{v},t}^2\mathbb{E}_t[\|\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \zeta_t\|_2^2]$$

$$\leq \|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + 2\eta_{\mathbf{v},t}\langle \mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t), \nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\rangle + \eta_{\mathbf{v},t}^2\|\nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2)$$

$$\leq (1-\rho\eta_{\mathbf{v},t})\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + 2\eta_{\mathbf{v},t}(F_S(\mathbf{w}_t, \mathbf{v}_t) - F_S(\mathbf{w}_t, \hat{\mathbf{v}}_S(\mathbf{w}_t))) + \eta_{\mathbf{v},t}^2\|\nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2),$$

where the third inequality we use the $F_S(\mathbf{w}, \cdot)$ is ρ-strongly concave. Since F_S is L-smooth, by choosing $\eta_{\mathbf{v},t} \leq 1/L$, we have

$$\mathbb{E}_t[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] \leq (1-\rho\eta_{\mathbf{v},t})\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 - \frac{\eta_{\mathbf{v},t}}{L}\|\nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \eta_{\mathbf{v},t}^2\|\nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2)$$

$$\leq (1-\rho\eta_{\mathbf{v},t})\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + \eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2). \tag{12}$$

Combining (12) and (11) we have

$$\mathbb{E}_t[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_{t+1})\|_2^2] \leq ((1+\frac{1}{\epsilon})2L^4/\rho^2\eta_{\mathbf{w},t}^2 + (1+\epsilon)(1-\rho\eta_{\mathbf{v},t}))\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + (1+\frac{1}{\epsilon})\eta_{\mathbf{w},t}^2 L^2/\rho^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) + (1+\frac{1}{\epsilon})4L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2(R_S(\mathbf{w}_t) - R_S(\mathbf{w}^*)) + (1+\epsilon)\eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2).$$

Taking expectation on both sides yields the desired bound. The proof is complete. ☐

Lemma 11. *Assume **(A2)** and **(A3)** hold. Assume $F_S(\cdot, \mathbf{v})$ satisfies PL condition with constant μ and $F_S(\mathbf{w}, \cdot)$ is ρ-strongly concave. Define $a_t = \mathbb{E}[R_S(\mathbf{w}_t) - R_S(\mathbf{w}^*)]$ and $b_t = \mathbb{E}[\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \mathbf{v}_t\|_2^2]$. For Algorithm 1, if $\eta_{\mathbf{w},t} \leq 1/(L+L^2/\rho)$ and $\eta_{\mathbf{v},t} \leq 1/L$, then for any non-increasing sequence $\{\lambda_t > 0\}$ and $\epsilon > 0$, the iterates $\{\mathbf{w}_t, \mathbf{v}_t\}_{t\in[T]}$ satisfy the following inequality*

$$a_{t+1} + \lambda_{t+1}b_{t+1} \leq k_{1,t}a_t + k_{2,t}\lambda_t b_t + \frac{(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{2}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) + 2(1+\frac{1}{\epsilon})\lambda_t L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) + \lambda_t(1+\epsilon)\eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2),$$

where

$$k_{1,t} = (1-\mu\eta_{\mathbf{w},t}) + \lambda_t(1+\frac{1}{\epsilon})4L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2,$$

$$k_{2,t} = \frac{L^2\eta_{\mathbf{w},t}}{2\lambda_t} + (1+\epsilon)(1-\rho\eta_{\mathbf{v},t}) + (1+\frac{1}{\epsilon})2L^4/\rho^2\eta_{\mathbf{w},t}^2.$$

Proof. Combining Lemma 9 and Lemma 10, we have for any $\lambda_{t+1} > 0$, we have

$$\begin{aligned}
a_{t+1} + \lambda_{t+1} b_{t+1} \leq & ((1-\mu\eta_{\mathbf{w},t}) + \lambda_{t+1}(1+\frac{1}{\epsilon})4L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2)a_t \\
& + (\frac{L^2\eta_{\mathbf{w},t}}{2} + \lambda_{t+1}(1+\epsilon)(1-\rho\eta_{\mathbf{v},t}) + \lambda_{t+1}(1+\frac{1}{\epsilon})2L^4/\rho^2\eta_{\mathbf{w},t}^2)b_t \\
& + \frac{(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{2}(\frac{B_{\mathbf{w}}^2}{m}+d\sigma_{\mathbf{w}}^2)+2(1+\frac{1}{\epsilon})\lambda_{t+1}L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m}+d\sigma_{\mathbf{w}}^2)+\lambda_{t+1}(1+\epsilon)\eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m}+d\sigma_{\mathbf{v}}^2) \\
\leq & ((1-\mu\eta_{\mathbf{w},t}) + \lambda_t(1+\frac{1}{\epsilon})4L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2)a_t \\
& + (\frac{L^2\eta_{\mathbf{w},t}}{2} + \lambda_t(1+\epsilon)(1-\rho\eta_{\mathbf{v},t}) + \lambda_t(1+\frac{1}{\epsilon})2L^4/\rho^2\eta_{\mathbf{w},t}^2)b_t \\
& + \frac{(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{2}(\frac{B_{\mathbf{w}}^2}{m}+d\sigma_{\mathbf{w}}^2)+2(1+\frac{1}{\epsilon})\lambda_t L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m}+d\sigma_{\mathbf{w}}^2)+\lambda_t(1+\epsilon)\eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m}+d\sigma_{\mathbf{v}}^2) \\
= & ((1-\mu\eta_{\mathbf{w},t}) + \lambda_t(1+\frac{1}{\epsilon})4L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2)a_t \\
& + \lambda_t(\frac{L^2\eta_{\mathbf{w},t}}{2\lambda_t} + (1+\epsilon)(1-\rho\eta_{\mathbf{v},t}) + (1+\frac{1}{\epsilon})2L^4/\rho^2\eta_{\mathbf{w},t}^2)b_t \\
& + \frac{(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{2}(\frac{B_{\mathbf{w}}^2}{m}+d\sigma_{\mathbf{w}}^2)+2(1+\frac{1}{\epsilon})\lambda_t L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m}+d\sigma_{\mathbf{w}}^2)+\lambda_t(1+\epsilon)\eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m}+d\sigma_{\mathbf{v}}^2).
\end{aligned}$$

where the first inequality we used $\lambda_{t+1} \leq \lambda_t$. The proof is completed. □

We are now ready to state the convergence theorem of Algorithm 1.

Theorem 3 (Theorem 3 restated)**.** *Assume **(A2)** and **(A3)** hold. Assume $F_S(\cdot, \mathbf{v})$ satisfies PL condition with constant μ and $F_S(\mathbf{w}, \cdot)$ is ρ-strongly concave. Assume $\mu \leq 2L^2$ and Let $\kappa = \frac{L}{\rho}$. For Algorithm 1, if $\eta_{\mathbf{w},t} = \mathcal{O}(\frac{1}{\mu t})$ and $\eta_{\mathbf{v},t} = \mathcal{O}(\frac{\kappa^2 \max\{1, \sqrt{\kappa/\mu}\}}{\mu t^{2/3}})$, then the iterates $\{\mathbf{w}_t, \mathbf{v}_t\}_{t\in[T]}$ satisfy the following inequality*

$$\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] = \mathcal{O}(\min\Big\{\frac{1}{L}, \frac{1}{\mu}\Big\}(\frac{B_{\mathbf{w}}^2/m + d\sigma_{\mathbf{w}}^2}{T^{2/3}}) + \max\Big\{1, \sqrt{\frac{L\kappa}{\mu}}\Big\}\frac{L\kappa^3}{\mu^2}(\frac{B_{\mathbf{v}}^2/m + d\sigma_{\mathbf{v}}^2}{T^{2/3}})). \tag{13}$$

Furthermore, if $\sigma_{\mathbf{w}}, \sigma_{\mathbf{v}}$ are given by (3)*, we have*

$$\begin{aligned}
&\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] \\
=&\mathcal{O}(\min\Big\{\frac{1}{L}, \frac{1}{\mu}\Big\}(\frac{B_{\mathbf{w}}^2}{mT^{2/3}} + \frac{G_{\mathbf{w}}^2 dT^{1/3}\log(1/\delta)}{n^2\epsilon^2}) + \max\Big\{1, \sqrt{\frac{L\kappa}{\mu}}\Big\}\frac{L\kappa^3}{\mu^2}(\frac{B_{\mathbf{v}}^2}{mT^{2/3}} + \frac{G_{\mathbf{v}}^2 dT^{1/3}\log(1/\delta)}{n^2\epsilon^2})).
\end{aligned} \tag{14}$$

Proof. Since $\eta_{\mathbf{v},t} \leq 1/L$, we can pick $\epsilon = \frac{\rho\eta_{\mathbf{v},t}}{2(1-\rho\eta_{\mathbf{v},t})}$. Then we have $(1+\epsilon)(1-\rho\eta_{\mathbf{v},t}) = 1 - \frac{\rho\eta_{\mathbf{v},t}}{2}$ and $1 + \frac{1}{\epsilon} \leq \frac{2}{\rho\eta_{\mathbf{v},t}}$. Therefore Lemma 11 can be simplified as

$$\begin{aligned}
k_{1,t} \leq & (1-\mu\eta_{\mathbf{w},t}) + \lambda_t \frac{8L^2/\rho^2(L+L^2/\rho)\eta_{\mathbf{w},t}^2}{\rho\eta_{\mathbf{v},t}}, \\
k_{2,t} \leq & \frac{L^2\eta_{\mathbf{w},t}}{2\lambda_t} + 1 - \frac{\rho\eta_{\mathbf{v},t}}{2} + \frac{4L^4/\rho^2\eta_{\mathbf{w},t}^2}{\rho\eta_{\mathbf{v},t}}.
\end{aligned}$$

If we choose $\lambda_t = \frac{4L^2\eta_{\mathbf{w},t}}{\rho\eta_{\mathbf{v},t}}$ and $\eta_{\mathbf{w},t} \leq \min\{\frac{\sqrt{\mu}}{8\kappa^2\sqrt{L+L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}\eta_{\mathbf{v},t}$, then further we have $k_{1,t} \leq 1 - \frac{\mu\eta_{\mathbf{w},t}}{2}$

and $k_{2,t} \le 1 - \frac{\rho\eta_{\mathbf{v},t}}{4}$. By Lemma 11 we have

$$\begin{aligned}
a_{t+1} + \lambda_{t+1} b_{t+1} \le & (1 - \min\{\frac{\mu}{2}, L^2\}\eta_{\mathbf{w},t})(a_t + \lambda_t b_t) + \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\
& + \frac{16L^4/\rho^3\eta_{\mathbf{w},t}^3}{\rho\eta_{\mathbf{v},t}^2}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) + \frac{4L^2(2-\rho\eta_{\mathbf{v},t})\eta_{\mathbf{w},t}\eta_{\mathbf{v},t}}{2\rho(1-\rho\eta_{\mathbf{v},t})}(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2) \\
\le & (1 - \frac{\mu\eta_{\mathbf{w},t}}{2})(a_t + \lambda_t b_t) + \frac{(L + L^2/\rho)\eta_{\mathbf{w},t}^2}{2}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\
& + \frac{16L^4/\rho^3\eta_{\mathbf{w},t}^3}{\rho\eta_{\mathbf{v},t}^2}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) + \frac{4L^2(2-\rho\eta_{\mathbf{v},t})\eta_{\mathbf{w},t}\eta_{\mathbf{v},t}}{2\rho(1-\rho\eta_{\mathbf{v},t})}(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2),
\end{aligned}$$

where we used $\mu \le 2L^2$. Taking $\eta_{\mathbf{w},t} = \frac{2}{\mu t}$ and $\eta_{\mathbf{v},t} = \max\{8\kappa^2\sqrt{(L + L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}\frac{2}{\mu t^{2/3}}$ and multiplying the preceding inequality with t on both sides, there holds

$$\begin{aligned}
& t(a_{t+1} + \lambda_{t+1} b_{t+1}) \le (t-1)(a_t + \lambda_t b_t) + \frac{2(L + L^2/\rho)}{\mu^2 t}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) \\
& + \frac{32L^4/\rho^3 \min\{\frac{\sqrt{\mu}}{8\kappa^2\sqrt{L+L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}^2}{\mu\rho t^{2/3}}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2) + \frac{16L^2 \max\{8\kappa^2\sqrt{(L + L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}}{2\mu^2\rho t^{2/3}}(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2).
\end{aligned}$$

Applying the preceding inequality inductively from $t = 1$ to T, we have

$$\begin{aligned}
T(a_{T+1} + \lambda_{T+1} b_{T+1}) \le & \frac{2(L + L^2/\rho)}{\mu^2}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2)\log(T) + \frac{32L^4/\rho^3 \min\{\frac{\sqrt{\mu}}{8\kappa^2\sqrt{L+L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}^2}{\mu\rho}(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2)T^{1/3} \\
& + \frac{16L^2 \max\{8\kappa^2\sqrt{(L + L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}}{2\mu^2\rho}(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2)T^{1/3}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] \le & a_{T+1} + \lambda_{T+1} b_{T+1} \\
\le & \frac{2(L + L^2/\rho)(B_{\mathbf{w}}^2/m + d\sigma_{\mathbf{w}}^2)}{\mu^2}\frac{\log(T)}{T} + \frac{32(B_{\mathbf{w}}^2/m + d\sigma_{\mathbf{w}}^2)L^4/\rho^3 \min\{\frac{\sqrt{\mu}}{8\kappa^2\sqrt{L+L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}^2}{\mu\rho}\frac{1}{T^{2/3}} \\
& + \frac{16(B_{\mathbf{v}}^2/m + d\sigma_{\mathbf{v}}^2)L^2 \max\{8\kappa^2\sqrt{(L+L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}}{2\mu^2\rho}\frac{1}{T^{2/3}}. \qquad (15)
\end{aligned}$$

Therefore, the estimation (13) follows from the fact that $\kappa = L/\rho$.
The result in Theorem 3 follows by observing $\max\left\{1, \sqrt{\frac{L\kappa}{\mu}}\right\}\frac{L\kappa^3}{\mu^2} \ge \min\left\{\frac{1}{L}, \frac{1}{\mu}\right\}$. Substituting the values of $\sigma_{\mathbf{w}}, \sigma_{\mathbf{v}}$, i.e., $\sigma_{\mathbf{w}} = \frac{c_2 G_{\mathbf{w}}\sqrt{T\log(\frac{1}{\delta})}}{n\epsilon}$ and $\sigma_{\mathbf{v}} = \frac{c_3 G_{\mathbf{v}}\sqrt{T\log(\frac{1}{\delta})}}{n\epsilon}$, into (13) yields the desired estimation (14). □

D.2 Proof of Theorem 4 (Generalization Error)

We first focus on to the generalization error $\mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)]$. Firstly, we introduce a lemma that bridges the generalization and the uniform argument stability. We modify the lemma so that it satisfies our needs.

Lemma 12 ([Lei et al., 2021])**.** *Let A be a randomized algorithm and $\epsilon > 0$. If for all neighboring datasets S, S', there holds*

$$\mathbb{E}_A[\|A_{\mathbf{w}}(S) - A_{\mathbf{w}}(S')\|_2] \le \varepsilon.$$

Furthermore, if the function $F(\mathbf{w}, \cdot)$ is ρ-strongly-concave and Assumptions 1, ***(A3)*** *hold, then the primal generalization error satisfies*

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

The next proposition states the set of saddle points is unique with respect to the variable $\mathbf{v}$ when $F_S(\mathbf{w},\cdot)$ is strongly concave.

Proposition 1. *Assume $F_S(\mathbf{w},\cdot)$ is ρ-strongly concave with $\rho>0$. Let $(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S)$ and $(\hat{\mathbf{w}}'_S,\hat{\mathbf{v}}'_S)$ be two saddle points of F_S. Then we have $\hat{\mathbf{v}}_S=\hat{\mathbf{v}}'_S$.*

Proof. Given $\hat{\mathbf{w}}_S$, by the strong concavity, we have

$$F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S)\geq F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}'_S)+\langle\nabla_{\mathbf{v}}F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S),\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\rangle+\frac{\rho}{2}\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\|_2^2.$$

Since $(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S)$ is a saddle point of F_S, it implies $\hat{\mathbf{v}}_S$ attains maximum of $F_S(\hat{\mathbf{w}}_S,\cdot)$. By the first order optimality we know $\langle\nabla_{\mathbf{v}}F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S),\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\rangle\geq 0$ and therefore

$$F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S)\geq F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}'_S)+\frac{\rho}{2}\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\|_2^2\geq F_S(\hat{\mathbf{w}}'_S,\hat{\mathbf{v}}'_S)+\frac{\rho}{2}\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\|_2^2, \tag{16}$$

where in the second inequality we used $(\hat{\mathbf{w}}'_S,\hat{\mathbf{v}}'_S)$ is also a saddle point of F_S. Similarly, given $\hat{\mathbf{w}}'_S$ we can show

$$F_S(\hat{\mathbf{w}}'_S,\hat{\mathbf{v}}'_S)\geq F_S(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S)+\frac{\rho}{2}\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\|_2^2. \tag{17}$$

Adding (16) and (17) together implies that $\rho\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}'_S\|_2^2\leq 0$. This implies $\hat{\mathbf{v}}_S=\hat{\mathbf{v}}'_S$ which completes the proof. ☐

Recall that $\pi_S:\mathcal{W}\to\mathcal{W}$ is the projection onto the set of saddle points $\Omega_S=\{\hat{\mathbf{w}}_S:(\hat{\mathbf{w}}_S,\hat{\mathbf{v}}_S\in\arg\min\max F_S(\mathbf{w},\mathbf{v})\}$. i.e. $\pi_S(\mathbf{w})=\arg\min_{\hat{\mathbf{w}}_S\in\Omega_S}\frac{1}{2}\|\mathbf{w}-\hat{\mathbf{w}}_S\|_2^2$. Proposition 1 makes sure the projection is well-defined. The next lemma shows that PL condition implies quadratic growth (QG) condition. The proof follows straightforward from Karimi et al. [2016] and we omit it for brevity.

Lemma 13. *Suppose the function $F_S(\cdot,\mathbf{v})$ satisfies μ-PL condition. Then F_S satisfies the QG condition with respect to $\mathbf{w}$ with constant 4μ, i.e.*

$$F_S(\mathbf{w},\mathbf{v})-F_S(\pi_S(\mathbf{w}),\mathbf{v})\geq 2\mu\|\mathbf{w}-\pi_S(\mathbf{w})\|_2^2,\quad\forall\mathbf{v}\in\mathcal{V}$$

With the help of Assumption 4 and the preceding lemmas, we can derive the uniform argument stability.

Lemma 14. *Assume **(A1)**, **(A3)** and **(A4)** hold. Assume $F_S(\cdot,\mathbf{v})$ satisfies PL condition with constant μ and $F_S(\mathbf{w},\cdot)$ is ρ-strongly concave. Let A be a randomized algorithm. If for any S, $\mathbb{E}[\|A_{\mathbf{w}}(S)-\pi_S(A_{\mathbf{w}}(S))\|_2]=\mathcal{O}(\varepsilon_A)$, then we have*

$$\mathbb{E}[\|A_{\mathbf{w}}(S)-A_{\mathbf{w}}(S')\|_2]\leq\mathcal{O}(\varepsilon_A)+\frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2}+\frac{G_{\mathbf{v}}^2}{\rho\mu}}.$$

Proof. Let $(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_S)\in\arg\min_{\mathbf{w}}\max_{\mathbf{v}}F_S(\mathbf{w},\mathbf{v})$ and $(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_{S'})$ defined in the similar way. By triangle inequality we have

$$\begin{aligned}\mathbb{E}[\|A_{\mathbf{w}}(S)-A_{\mathbf{w}}(S')\|_2]\leq&\mathbb{E}[\|A_{\mathbf{w}}(S)-\pi_S(A_{\mathbf{w}}(S))\|_2]+\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2+\mathbb{E}[\|A_{\mathbf{w}}(S')-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2]\\=&\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2+\mathcal{O}(\varepsilon_A).\end{aligned}$$

Since $\pi_S(A_{\mathbf{w}}(S))\in\arg\min_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\hat{\mathbf{v}}_S)$ and by Assumption **(A4)** we know that $\pi_S(A_{\mathbf{w}}(S))$ is the closest optimal point of F_S to $\pi_{S'}(A_{\mathbf{w}}(S'))$. And since $\hat{\mathbf{v}}_S$ is fixed, by Lemma 13, we have

$$2\mu\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2\leq F_S(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S)-F_S(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_S).$$

Similarly, we have

$$2\mu\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2\leq F_{S'}(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'})-F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_{S'}).$$

Summing up the above two inequalities we have

$$\begin{aligned}4\mu\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2 \leq & F_S(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S)-F_S(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_S)\\ &+F_{S'}(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'})-F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_{S'}).\end{aligned} \tag{18}$$

On the other hand, by the ρ-strong concavity of $F_S(\cdot,\mathbf{v})$ and $\hat{\mathbf{v}}_S=\arg\max_{\mathbf{v}\in\mathcal{V}}F_S(\pi_S(A_{\mathbf{w}}(S)),\mathbf{v})$, we have

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

Similarly, we have

$$\frac{\rho}{2}\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}_{S'}\|_2^2\leq F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_{S'})-F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S).$$

Summing up the above two inequalities we have

$$\begin{aligned}\rho\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}_{S'}\|_2^2\leq & F_S(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_S)-F_S(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'})\\ &+F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_{S'})-F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S).\end{aligned} \tag{19}$$

Summing up (18) and (19) rearranging terms, we have

$$\begin{aligned}&4\mu\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2+\rho\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}_{S'}\|_2^2\\ \leq & F_S(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S)-F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S)+F_{S'}(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'})-F_S(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'})\\ = & \frac{1}{n}\big(f(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S;\mathbf{z})-f(\pi_{S'}(A_{\mathbf{w}}(S')),\hat{\mathbf{v}}_S;\mathbf{z}')+f(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'};\mathbf{z}')-f(\pi_S(A_{\mathbf{w}}(S)),\hat{\mathbf{v}}_{S'};\mathbf{z})\big)\\ \leq & \frac{2G_{\mathbf{w}}}{n}\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2+\frac{2G_{\mathbf{v}}}{n}\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}_{S'}\|_2\\ \leq & \frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{\mu}+\frac{4G_{\mathbf{v}}^2}{\rho}}\times\sqrt{4\mu\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2+\rho\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}_{S'}\|_2^2},\end{aligned}$$

where the second inequality is due to Lipschitz continuity of f, the third inequality is due to Cauchy-Schwartz inequality. Therefore

$$2\sqrt{\mu}\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2\leq\sqrt{4\mu\|\pi_S(A_{\mathbf{w}}(S))-\pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2+\rho\|\hat{\mathbf{v}}_S-\hat{\mathbf{v}}_{S'}\|_2^2}\leq\frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{\mu}+\frac{4G_{\mathbf{v}}^2}{\rho}}.$$

The proof is complete. □

We are now ready to present the generalization error of Algorithm 1 in terms of $\mathbf{w}_T$.

Theorem 4. *Assume **(A1)**, **(A3)** and **(A4)** hold. Assume $F_S(\cdot,\mathbf{v})$ satisfies PL condition with constant μ and $f(\mathbf{w},\cdot;\mathbf{z})$ is ρ-strongly concave. For Algorithm 1, the iterates $\{\mathbf{w}_t,\mathbf{v}_t\}$ satisfies the following inequality*

$$\mathbb{E}[R(\mathbf{w}_T)-R_S(\mathbf{w}_T)]\leq(1+\frac{L}{\rho})G_{\mathbf{w}}\Big(\sqrt{\frac{\varepsilon_T}{2\mu}}+\frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2}+\frac{G_{\mathbf{v}}^2}{\rho\mu}}\Big).$$

Proof. Since R_S satisfies μ-PL, by Lemma 13 and Theorem 3, we have

$$\mathbb{E}[\|\mathbf{w}_T-\pi(\mathbf{w}_T)\|_2]\leq\sqrt{\mathbb{E}[\|\mathbf{w}_T-\pi(\mathbf{w}_T)\|_2^2]}\leq\sqrt{\mathbb{E}[\frac{1}{2\mu}(R_S(\mathbf{w}_T)-R_S^*)]}\leq\sqrt{\frac{\varepsilon_T}{2\mu}}.$$

By Lemma 14, we have

$$\mathbb{E}[\|\mathbf{w}_T-\mathbf{w}_T'\|_2]\leq\sqrt{\frac{\varepsilon_T}{2\mu}}+\frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2}+\frac{G_{\mathbf{v}}^2}{\rho\mu}}.$$

By Part b) of Lemma 12, we have

$$\mathbb{E}[R(\mathbf{w}_T)-R_S(\mathbf{w}_T)]\leq(1+\frac{L}{\rho})G_{\mathbf{w}}\Big(\sqrt{\frac{\varepsilon_T}{2\mu}}+\frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2}+\frac{G_{\mathbf{v}}^2}{\rho\mu}}\Big).$$

The proof is complete. □

The next theorem establishes the generalization bound for the empirical maximizer of a strongly concave objective, i.e. $\mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)]$. The proof follows from Shalev-Shwartz et al. [2009].

Theorem 5. *Assume **(A1)** holds. Assume $F_S(\mathbf{w}, \cdot)$ is ρ-strongly concave. Assume that for any $\mathbf{w}$ and S, the function $\mathbf{v} \mapsto F_S(\mathbf{w}, \mathbf{v})$ is ρ-strongly-concave. Then*

$$\mathbb{E}\big[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)\big] \le \frac{4G_{\mathbf{v}}^2}{\rho n}.$$

Proof. We decompose the term $\mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)]$ as

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

where $\hat{\mathbf{v}}_S^* = \arg\max_{\mathbf{v}} F_S(\mathbf{w}^*, \mathbf{v})$. The second term $\mathbb{E}\big[F(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*, \mathbf{v}^*)\big] \le 0$ since $(\mathbf{w}^*, \mathbf{v}^*)$ is a saddle point of F. Hence it suffices to bound $\mathbb{E}\big[F_S(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*, \hat{\mathbf{v}}_S^*)\big]$. 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

$$\begin{aligned}
F_S(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F_S(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*) =& \frac{1}{n}\sum_{j\neq i}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_j) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_j)\Big) + \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i)\Big) \\
=& \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i') - f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i')\Big) + \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i)\Big) \\
&+ F_{S^{(i)}}(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F_{S^{(i)}}(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*) \\
\le& \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i') - f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i')\Big) + \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i)\Big) \\
\le& \frac{2G_{\mathbf{v}}}{n}\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}_{S^{(i)}}^*\big\|_2,
\end{aligned} \tag{20}$$

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 the Lipschitz continuity. Since F_S is strongly-concave and $\hat{\mathbf{v}}_S^*$ maximizes $F_S(\mathbf{w}^*, \cdot)$, we know

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

Combining it with (20) we get $\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}_{S^{(i)}}^*\big\|_2 \le 4G_{\mathbf{v}}/(\rho n)$. By Lipschitz continuity, the following inequality holds for any z

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

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

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

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

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

The proof is complete. ☐

Theorem 6 (Theorem 4 restated)**.** *Assume the function $f(\mathbf{w}, \cdot; \mathbf{z})$ is ρ-strongly concave and $F_S(\cdot, \mathbf{v})$ satisfies μ-PL condition. Suppose **(A1)** and **(A3)** hold. If $\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] \le \varepsilon_T$, then*

$$\mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] \le (1+\kappa)G_{\mathbf{w}}\Big(\sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2} + \frac{G_{\mathbf{v}}^2}{\rho\mu}}\Big),$$

and

$$\mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)] \leq \frac{4G_{\mathbf{v}}^2}{\rho n}.$$

Proof. It follows directly from Theorem 4 and 5. □

D.3 Proof of Theorem 5

Theorem 7 (Theorem 5 restated). *Assume **(A1)**, **(A3)** and **(A4)** hold. Assume $F_S(\cdot,\mathbf{v})$ satisfies PL condition with constant μ and $f(\mathbf{w},\cdot;\mathbf{z})$ is ρ-strongly concave. For SGDA, if $\mathbb{E}[R_S(\mathbf{w}_T)-R_S^*] = \mathcal{O}(\varepsilon_T)$, then iterates $\{\mathbf{w}_t,\mathbf{v}_t\}$ satisfies the following inequality*

$$\mathbb{E}[R(\mathbf{w}_T) - R^*] = \mathcal{O}(\varepsilon_T + (1+\frac{L}{\rho})G_{\mathbf{w}}\Big(\sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2} + \frac{G_{\mathbf{v}}^2}{\rho\mu}}\Big) + \frac{4G_{\mathbf{v}}^2}{\rho n}).$$

Furthermore, if we choose $T=\mathcal{O}(n)$, $\eta_{\mathbf{w},t} = \mathcal{O}(\frac{1}{\mu t})$ and $\eta_{\mathbf{v},t} = \mathcal{O}(\frac{\kappa^2\max\{1,\sqrt{\kappa/\mu}\}}{\mu t^{2/3}})$, then

$$\mathbb{E}[R(\mathbf{w}_T) - R^*] = \mathcal{O}\big(\frac{\kappa^{2.75}}{\mu^{1.75}}(\frac{1}{n^{1/3}} + \frac{\sqrt{d\log(1/\delta)}}{n^{5/6}\epsilon})\big).$$

Proof. For any $\mathbf{w}^* \in \arg\min_{\mathbf{w}} R(\mathbf{w})$, recall that we have the error decomposition (5), which is

$$\begin{aligned}\mathbb{E}[R(\mathbf{w}_T) - R^*] =& \mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] + \mathbb{E}[R_S(\mathbf{w}_T) - R_S^*] + \mathbb{E}[R_S^* - R_S(\mathbf{w}^*)] + \mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)] \\ \leq& \mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] + \mathbb{E}[R_S(\mathbf{w}_T) - R_S^*] + \mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)],\end{aligned}$$

where the inequality is by $R_S^* - R_S(\mathbf{w}^*) \leq 0$. By Theorem 4, we have

$$\mathbb{E}[R(\mathbf{w}_T) - R_S(\mathbf{w}_T)] \leq (1+\frac{L}{\rho})G_{\mathbf{w}}\Big(\sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2} + \frac{G_{\mathbf{v}}^2}{\rho\mu}}\Big).$$

And by Theorem 5, we have

$$\mathbb{E}[R_S(\mathbf{w}^*) - R(\mathbf{w}^*)] \leq \frac{4G_{\mathbf{v}}^2}{\rho n}.$$

We can plug the above two inequalities into (5), and get

$$\mathbb{E}[R(\mathbf{w}_T) - R^*] = \mathcal{O}(\varepsilon_T + (1+\frac{L}{\rho})G_{\mathbf{w}}\Big(\sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n}\sqrt{\frac{G_{\mathbf{w}}^2}{4\mu^2} + \frac{G_{\mathbf{v}}^2}{\rho\mu}}\Big) + \frac{4G_{\mathbf{v}}^2}{\rho n}).$$

Now by the choice of $\eta_{\mathbf{w},t},\eta_{\mathbf{v},t}$, and Theorem 3 , we have $\varepsilon_T = \mathcal{O}(\frac{\kappa^{3.5}}{\mu^{2.5}}\frac{1/m + d(\sigma_{\mathbf{w}}^2+\sigma_{\mathbf{v}}^2)}{T^{2/3}})$. Assume m is a constant. Plugging ε_T into the preceding inequality and letting $T=\mathcal{O}(n)$ yields the second statement. □

E Additional Experimental Details

E.1 Source Code

For the purpose of double-blind peer-review, the source code is accessible in the supplementary file.

E.2 Computing Infrastructure Description

All algorithms are implemented in Python 3.6 and trained and tested on an Intel(R) Xeon(R) CPU W5590 @3.33GHz with 48GB of RAM and an NVIDIA Quadro RTX 6000 GPU with 24GB memory. The PyTorch version is 1.6.0.

E.3 Description of Datasets

In experiments, we use three benchmark datasets. Specifically, ijcnn1 dataset from LIBSVM repsitory, MNIST dataset and Fashion-MNIST dataset are from LeCun et al. [1998], and Xiao et al. [2017]. The details of these datasets are shown in Table 3. For the ijcnn1 dataset, we normalize the features into [0,1]. For MNIST and Fashion-MNIST datasets, we first normalize the features of them into [0,1] then normalize them according to the mean and standard deviation.

Dataset	#Classes	#Training Samples	#Testing Samples	#Features
ijcnn1	2	39,992	9,998	22
MNIST	10	60,000	10,000	784
Fashion-MNIST	10	60,000	10,000	784

Table 1: *Statistical information of each dataset for AUC optimization.*

E.4 Training Settings

The training settings for NSEG and DP-SGDA on all datasets are shown in Table 2.

Methods	Datasets	Batch Size	Learning Rate				Epochs		Projection Size	
			Ori		DP		Ori	DP	Ori	DP
			w	**v**	**w**	**v**				
NSEG	ijcnn1	64	300	300	350	350	1000	15	100	100
	MNIST	64	11	11	5	5	100	15	2	2
	Fashion-MNIST	64	11	11	5	5	100	15	3	3
DP-SGDA (Linear)	ijcnn1	64	300	300	350	350	100	15	10	10
	MNIST	64	11	11	5	5	100	15	2	2
	Fashion-MNIST	64	11	11	5	5	100	15	3	3
DP-SGDA (MLP)	ijcnn1	64	3000	3001	500	501	10	10	100	100
	MNIST	64	900	1000	100	210	10	10	2	2
	Fashion-MNIST	64	900	1000	100	210	10	10	2	2

Table 2: *Training settings for each model and each dataset.*

E.5 DP-SGDA for AUC Maximization

In this section, we provide details of using DP-SGDA to learn AUC maximization problem. AUC maximization with square loss can be reformulated as

$$
\begin{aligned}
F(\theta, a, b, \mathbf{v}) = \mathbb{E}_{\mathbf{z}}[(1-p)(h(\theta;\mathbf{x})-a)^2\mathbb{I}[y=1] + p(h(\theta;\mathbf{x})-b)^2\mathbb{I}[y=-1] \\
+ 2(1+\mathbf{v})(ph(\theta;\mathbf{x})\mathbb{I}[y=-1] - (1-p)h(\theta;\mathbf{x})\mathbb{I}[y=1])] - p(1-p)\mathbf{v}^2]
\end{aligned}
$$

where $\mathbf{z} = (\mathbf{x}, y)$ and $p = \mathbb{P}[y=1]$. The empirical risk formulation is given as

$$
\begin{aligned}
F_S(\theta, a, b, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{n_+}(h(\theta;\mathbf{x}_i)-a)^2\mathbb{I}[y_i=1] + \frac{1}{n_-}(h(\theta;\mathbf{x}_i)-b)^2\mathbb{I}[y_i=-1] \\
+ 2(1+\mathbf{v})\Big(\frac{1}{n_-}h(\theta;\mathbf{x}_i)\mathbb{I}[y_i=-1] - \frac{1}{n_+}h(\theta;\mathbf{x}_i)\mathbb{I}[y_i=1]\Big) - \frac{1}{n}\mathbf{v}^2\Big\}
\end{aligned}
$$

Algorithm 1 DP-SGDA for AUC Maximization

Inputs: Private dataset $S = \{\mathbf{z}_i : i \in [n]\}$, privacy budget ϵ, δ, number of iterations T, learning rates $\{\gamma_t, \lambda_t\}_{t=1}^T$, initial points $(\theta_0, a_0, b_0, \mathbf{v}_0)$
Compute $n_+ = \sum_{i=1}^n \mathbb{I}[y_i = 1]$ and $n_- = \sum_{i=1}^n \mathbb{I}[y_i = -1]$
Compute noise parameters σ_1 and σ_2 based on Eq. (3)
for $t = 1$ to T **do**
  Randomly select a batch S_t
  For each $j \in I_t$, compute gradient $\nabla_\theta f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j), \nabla_a f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j), \nabla_b f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j)$ and $\nabla_c f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j)$ based on Eq. (21)
  Sample independent noises $\xi_t \sim \mathcal{N}(0, \sigma_1^2 I_{d+2})$ and $\zeta_t \sim \mathcal{N}(0, \sigma_2^2)$
  Update

$$\begin{aligned}\begin{pmatrix}\theta_{t+1}\\ a_{t+1}\\ b_{t+1}\end{pmatrix} =& \Pi\left\{\begin{pmatrix}\theta_t\\ a_t\\ b_t\end{pmatrix} - \gamma_t\Big(\frac{1}{m}\sum_{j\in I_t}\begin{pmatrix}\nabla_\theta f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j)\\ \nabla_a f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j)\\ \nabla_b f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j)\end{pmatrix} + \xi_t\Big)\right\}\\ \mathbf{v}_{t+1} =& \Pi\Big\{\mathbf{v}_t + \lambda_t(\frac{1}{m}\sum_{j\in I_t}\nabla_{\mathbf{v}} f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j) + \zeta_t)\Big\}\end{aligned}$$

end for
Outputs: $(\theta_T, a_T, b_T, \mathbf{v}_T)$ or $(\bar{\theta}_T, \bar{a}_T, \bar{b}_T, \bar{\mathbf{v}}_T)$

For any subset S_t of size m, let I_t denote the set of indices in S_t, the gradients of any $j \in I_t$ are given by

$$\begin{aligned}\nabla_\theta f(\theta, a, b, \mathbf{v}; \mathbf{z}_j) =& \frac{2}{n_+}(h(\theta; \mathbf{x}_j) - a)\nabla h(\theta; \mathbf{x}_j)\mathbb{I}[y_j = 1] + \frac{2}{n_-}(h(\theta; \mathbf{x}_j) - b)\nabla h(\theta; \mathbf{x}_j)\mathbb{I}[y_j = -1]\\ &+ 2(1 + \mathbf{v})\Big(\frac{1}{n_-}\nabla h(\theta; \mathbf{x}_j)\mathbb{I}[y_j = -1] - \frac{1}{n_+}\nabla h(\theta; \mathbf{x}_j)\mathbb{I}[y_j = 1]\Big)\\ \nabla_a f(\theta, a, b, \mathbf{v}; \mathbf{z}_j) =& \frac{2}{n_+}(a - h(\theta; \mathbf{x}_j))\mathbb{I}[y_j = 1], \quad \nabla_b f(\theta, a, b, \mathbf{v}; \mathbf{z}_j) = \frac{2}{n_-}(b - h(\theta; \mathbf{x}_j))\mathbb{I}[y_j = -1]\\ \nabla_{\mathbf{v}} f(\theta, a, b, \mathbf{v}; \mathbf{z}_j) =& 2\Big(\frac{1}{n_-}h(\theta; \mathbf{x}_j)\mathbb{I}[y_j = -1] - \frac{1}{n_+}h(\theta; \mathbf{x}_j)\mathbb{I}[y_j = 1]\Big) - \frac{2}{n}\mathbf{v}\end{aligned} \tag{21}$$

The pseudo-code can be found in Algorithm 1.

F Additional Experimental Results

We show the details of NSEG and DP-SGDA (Linear and MLP settings) performance with using five different $\epsilon \in \{0.1, 0.5, 1, 5, 10\}$ and three different $\delta \in \{1e-4, 1e-5, 1e-6\}$ in Table 3. From Table 3, we can find that the performance will be decreased when decrease the value of δ in the same ϵ settings. The reason is that the small δ is corresponding to a large value of σ based on Theorem 1. A large σ means a large noise will be added to the gradients during the training updates. Therefore, the AUC performance will be decreased as δ decreasing. On the other hand, we can find that our DP-SGDA(Linear) outperforms NSEG under the same settings. This is because the NSEG method will add a larger noise than DP-SGDA into the gradients in the training and we have discussed this detail in the Section 4.2.
We also compare the σ values from NSEG and DP-SGDA methods on all datasets in Figure 1 (a) with setting δ=1e-5 and (b) δ=1e-4. From the figure, it is clear that the σ from NSEG is larger than ours in all ϵ settings. This implies the noise generated from NSEG is also larger than ours.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer*

Dataset		ijcnn1			MNIST			Fashion-MNIST		
Algorithm		Linear		MLP	Linear		MLP	Linear		MLP
		NSEG	DP-SGDA	DP-SGDA	NSEG	DP-SGDA	DP-SGDA	NSEG	DP-SGDA	DP-SGDA
Original		92.191	92.448	96.609	93.306	93.349	99.546	96.552	96.523	98.020
δ=1e-4	ϵ=0.1	90.231	91.229	94.020	91.285	91.962	98.300	95.490	95.637	96.312
	ϵ=0.5	90.352	91.366	96.108	91.328	92.067	98.703	95.533	95.829	97.098
	ϵ=1	90.358	91.376	96.316	91.331	92.073	98.722	95.536	95.840	97.143
	ϵ=5	90.363	91.385	96.326	91.334	92.079	98.746	95.539	95.849	97.208
	ϵ=10	90.363	91.387	96.329	91.335	92.080	98.750	95.539	95.850	97.219
δ=1e-5	ϵ=0.1	90.168	91.169	93.274	91.266	91.910	98.092	95.468	95.535	95.989
	ϵ=0.5	90.349	91.362	96.029	91.326	92.063	98.675	95.531	95.823	97.031
	ϵ=1	90.357	91.373	96.209	91.330	92.071	98.714	95.535	95.837	97.122
	ϵ=5	90.363	91.384	96.300	91.334	92.079	98.743	95.538	95.848	97.200
	ϵ=10	90.363	91.386	96.301	91.334	92.080	98.747	95.539	95.850	97.213
δ=1e-6	ϵ=0.1	90.106	91.110	92.763	91.247	91.858	97.878	95.446	95.468	95.692
	ϵ=0.5	90.346	91.357	95.840	91.324	92.058	98.656	95.530	95.816	96.988
	ϵ=1	90.355	91.371	96.167	91.330	92.070	98.705	95.534	95.834	97.102
	ϵ=5	90.363	91.383	96.294	91.334	92.078	98.742	95.538	95.848	97.198
	ϵ=10	90.363	91.386	96.297	91.334	92.080	98.747	95.539	95.850	97.213

Table 3: *Comparison of AUC performance in NSEG and DP-SGDA (Linear and MLP settings) on three datasets with different ϵ and different δ. The "Original" means no noise ($\epsilon = \infty$) is added in the algorithms.*

and communications security, pages 308–318, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017.

Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *ICLR*, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *ICML*, 2021.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, pages 6083–6093. PMLR, 2020.

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic auc maximization with deep neural networks. In *ICLR*, 2020.

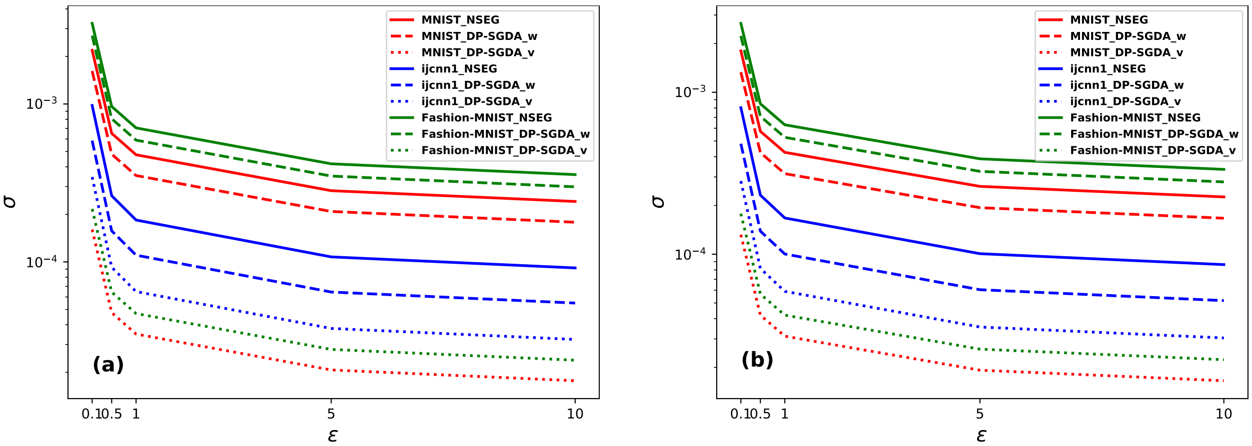

Figure 1: *Comparison of σ in NSEG and DP-SGDA (with Linear setting) on three datasets with different ϵ and (a) δ=1e-5 and (b) δ=1e-4.*

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Arda Sahiner, Tolga Ergen, Batu Ozturkler, Burak Bartan, John Pauly, Morteza Mardani, and Mert Pilanci. Hidden convexity of wasserstein gans: Interpretable generative models with closed-form solutions. *arXiv preprint arXiv:2107.05680*, 2021.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Puyu Wang, Zhenhuan Yang, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private empirical risk minimization for auc maximization. *Neurocomputing*, 461:419–437, 2021.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016.

Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR, 2021.