Differentially Private SGDA for Minimax Problems

Abstract

Stochastic gradient descent ascent (SGDA) and its variants have been the workhorse for solving minimax problems. However, in contrast to the well-studied stochastic gradient descent (SGD) with differential privacy (DP) constraints, there is little work on understanding the generalization (utility) of SGDA with DP constraints. In this paper, we use the algorithmic stability approach to establish the generalization (utility) of DP-SGDA in different settings. In particular, for the convex-concave setting, we prove that the DP-SGDA can achieve an optimal utility rate in terms of the weak primal-dual population risk in both smooth and non-smooth cases. To our best knowledge, this is the first-ever-known result for DP-SGDA in the non-smooth case. We further provide its utility analysis in the nonconvex-strongly-concave setting which is the first-ever-known result in terms of the primal population risk. The convergence and generalization results for this nonconvex setting are new even in the non-private setting. Finally, numerical experiments are conducted to demonstrate the effectiveness of DP-SGDA for both convex and nonconvex cases.

1 INTRODUCTION

In recent years, there is a growing interest on studying the minimax problems which involve both minimization over the primal variable $\mathbf{w}$ and maximization over the dual variable $\mathbf{v}$. Notable examples include generative adversarial networks (GANs) [Goodfellow et al., 2014, Arjovsky et al., 2017], AUC maximization [Gao et al., 2013, Ying et al., 2016, Liu et al., 2020, Zhao et al., 2011], robust learning [Audibert and Catoni, 2011, Xu et al., 2009], adversarial training [Sinha et al., 2017], algorithmic fairness [Mohri et al., 2019, Li et al., 2019, Wang et al., 2020b, Martinez et al., 2020, Diana et al., 2021], and Markov Decision Process (MDP) [Puterman, 2014, Wang, 2017]. Details of these motivating examples are given in Appendix A.

The minimax problem can be formulated as

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} \Big\{ F(\mathbf{w},\mathbf{v}) := \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\mathbf{w},\mathbf{v};\mathbf{z})] \Big\}, \qquad (1)$$

where $\mathcal{W} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{V} \subseteq \mathbb{R}^{d_2}$ are two nonempty closed and convex domains and $\mathbf{z}$ is a random variable from some distribution $\mathcal{D}$ taking values in $\mathcal{Z}$. Since the distribution $\mathcal{D}$ is usually unknown and one has access only to an i.i.d. training dataset $S = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$, one resorts to solving its empirical minimax problem

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} \Big\{ F_S(\mathbf{w},\mathbf{v}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w},\mathbf{v};\mathbf{z}_i) \Big\}.$$

One popular optimization algorithm for solving this problem is SGDA. Specifically, at iteration t, upon receiving a random data point or mini-batch from S, it performs gradient descent over $\mathbf{w}$ with the stepsize $\eta_{\mathbf{w},t}$ and gradient ascent over $\mathbf{v}$ with the stepsize $\eta_{\mathbf{v},t}$.

As SGDA is conceptually simple and easy to implement, it is widely deployed in solving minimax problems, e.g., GANs [Goodfellow et al., 2014], adversarial learning [Sinha et al., 2017], and AUC maximization [Ying et al., 2016]. Its local convergence analysis for nonconvex-(strongly)-concave problems was established in Lin et al. [2020]. Other variants of SGDA were proposed and studied in [Luo et al., 2020, Nouiehed et al., 2019, Rafique et al., 2021, Yan et al., 2020].

On another front, collected data often contain sensitive information such as individual records from hospitals, online behavior from social media, and genomic data from cancer diagnosis. Differential privacy [Dwork et al., 2014] has emerged as a well-accepted mathematical definition of privacy which ensures that an attacker gets roughly the same information from the dataset regardless of whether an individual is present or not. Its related technologies have been

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

adopted by Google [Erlingsson et al., 2014], Apple [Ding et al., 2017], and the US Census Bureau [Abowd, 2016]. While SGD and SGDA have become the workhorse behind the remarkable progress of machine learning and AI, it is of pivotal importance for developing their counterparts with DP constraints.

Many studies analyze the privacy and utility of DP-SGD for the ERM problem that only involves the minimization over $\mathbf{w}$, [e.g. Bassily et al., 2019, 2020, Feldman et al., 2020, Song et al., 2013, Wang et al., 2021a, 2020a, 2019b, Wu et al., 2017, Zhou et al., 2020]. In contrast, there is little work on analysing the utility of minimax optimization algorithms with DP constraints except the recent work of Boob and Guzmán [2021]. However, Boob and Guzmán focus on the noisy stochastic extragradient method on convex-concave and smooth settings.

Studying the computational and statistical behavior of DP-SGDA is fundamental towards the understanding of stochastic optimization algorithm for minimax problem under the differential privacy constraint. In this paper, we propose novel convergence and stability analysis to establish the utility of DP-SGDA in empirical saddle point and population forms such as the weak primal-dual population risk and the primal population risk. We collect in Table 1 the notations and results of performance measures in this paper. In particular, our contributions can be summarized as follows.

• We analyze the privacy and utility of DP-SGDA under the convex-concave setting in terms of the weak primal-dual population risk, i.e., $\max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\big[F(A_{\mathbf{w}}(S),\mathbf{v})\big] - \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}\big[F(\mathbf{w},A_{\mathbf{v}}(S))\big]$, where $(A_{\mathbf{w}}(S),A_{\mathbf{v}}(S))$ is the output of DP-SGDA. Specifically, we show that it can guarantee (ϵ,δ)-DP and achieve the optimal rate $\mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big)$ for smooth and nonsmooth cases where $d = \max\{d_1, d_2\}$. To our best knowledge, this is the first-ever known result for DP-SGDA in the nonsmooth case.

• We further study the utility of DP-SGDA in the nonconvex-strongly-concave case in terms of the primal population risk, i.e., $R(A_{\mathbf{w}}(S)) = \max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\big[F(A_{\mathbf{w}}(S),\mathbf{v})\big]$. In particular, under the Polyak-Łojasiewicz (PL) condition of F_S, we prove that the excess primal population risk, i.e., $R(A_{\mathbf{w}}(S)) - \min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})$, enjoys the rate $\mathcal{O}\Big(\frac{1}{n^{1/3}} + \frac{\sqrt{d\log(1/\delta)}}{n^{5/6}\epsilon}\Big)$ while guaranteeing (ϵ,δ)-DP. The key techniques involve the convergence analysis of $R_S(A_{\mathbf{w}}(S)) - \min_{\mathbf{w}} R_S(\mathbf{w})$ and the stability analysis for $A_{\mathbf{w}}(S)$ which are of interest in their own rights. As far as we are aware, these results are the first ones known for DP-SGDA in the nonconvex setting.

• We perform numerical experiments on three benchmark datasets which validate the effectiveness of DP-SGDA for both convex and non-convex cases.

1.1 MOTIVATING EXAMPLES

We give two examples of minimax problems under the DP constraint. See Appendix A for more examples and details.

AUC Maximization. Area Under the ROC Curve (AUC) is a widely used measure for binary classification. It has been shown optimizing AUC is equivalent to a minimax problem once auxiliary variables $a, b, v \in \mathbb{R}$ are introduced Ying et al. [2016].

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

Differential privacy has been applied to learn private classifier by optimizing AUC Wang et al. [2021b].

Generative Adversarial Networks. Originally proposed in Goodfellow et al. [2014], GAN in general can be written as a minimax problem between a generator network $G_{\mathbf{v}}$ and a discriminator network $D_{\mathbf{w}}$

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

DP-SGDA and its variants were employed to train differential private GANs by Xie et al. [2018]. Recently differential privacy has successfully applied to private data generation by GAN framework [Jordon et al., 2018, et al., 2019].

1.2 RELATED WORK

Below we briefly discuss some related work.

Convergence analysis for SGDA. It is a classical result that SGDA can achieve a convergence rate $\mathcal{O}(1/\sqrt{T})$ in the convex and concave case [e.g. Nedić and Ozdaglar, 2009, Nemirovski et al., 2009] where T is the number of iterations. For the nonconvex-(strongly)-concave case, the work of Lin et al. [2020] shows the local convergence of SGDA if the stepsizes $\eta_{\mathbf{w},t}$ and $\eta_{\mathbf{v},t}$ are chosen to be appropriately different. Other important studies consider variants of SGDA and prove their local convergence for the nonconvex case. Such algorithms include nested algorithms [Rafique et al., 2021] for weakly-convex-weakly-concave problems, multi-step GDA [Nouiehed et al., 2019] under the one-sided PL condition, epoch-wise SGDA [Yan et al., 2020], and stochastic recursive SGDA [Luo et al., 2020] for nonconvex-strongly-concave problems, to mention but a few.

Stability and generalization of non-private SGD and SGDA. The studies of Hardt et al. [2016], Charles and Papailiopoulos [2018], Kuzborskij and Lampert [2018] use uniform stability [Bousquet and Elisseeff, 2002] to derive the generalization of non-private SGD for the convex and smooth case while the convex and nonsmooth case was established by Bassily et al. [2020], Lei and Ying [2020]. The nonconvex case under the PL-condition was considered by Charles and Papailiopoulos [2018], Lei and Ying [2021]. The stability and generalization of SGDA for minimax problems were studied by Lei et al. [2021] in different forms for

Algorithm	Assumption	Measure	Rate	Complexity	Simplicity
NSEG	C-C, Lip, S	$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$	$\mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big)$	$\mathcal{O}(n^2)$	Single-loop
NISPP	C-C, Lip, S	$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$	$\mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big)$	$\mathcal{O}(n^{3/2}\log(n))$	Double-loop
DP-SGDA (Ours)	C-C, Lip, S	$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$	$\mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big)$	$\mathcal{O}(n^{3/2})$	Single-loop
	C-C, Lip	$\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$	$\mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big)$	$\mathcal{O}(n^{5/2})$	
	PL-SC, Lip, S	$R(A_{\mathbf{w}}(S)) - \min_{\mathbf{w}} R(\mathbf{w})$	$\mathcal{O}\Big(\frac{1}{n^{1/3}} + \frac{\sqrt{d\log(1/\delta)}}{n^{5/6}\epsilon}\Big)$	$\mathcal{O}(n^{3/2})$	

Table 1: *Summary of Results. DP-SGDA is Algorithm 1 in this paper. NSEG and NISPP are Algorithm 1 and 2 in Boob and Guzmán [2021], respectively. Here C-C means convexity and concavity, PL-SC means PL condition and strong concavity, Lip means Lipschitz continuity, S means the smoothness.* $\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ *is the weak PD population risk and* $R(A_{\mathbf{w}}(S)) - \min_{\mathbf{w}} R(\mathbf{w})$ *is the excess primal population risk.*

convex and nonconvex, smooth, and nonsmooth cases, and by Farnia and Ozdaglar [2021] with focus on the smooth cases.

DP-SGD and DP-SGDA. DP-SGD was shown to attain the optimal excess population risk $\mathcal{O}(1/\sqrt{n}+\sqrt{d\log(1/\delta)}/n\epsilon)$ in Bassily et al. [2019, 2020], Wang et al. [2021a, 2020a] for the convex case. For nonconvex objectives, Wang et al. [2019a] studied the DP Gradient Langevin Dynamics, and Zhang et al. [2021b] studied a multi-stage type of DP-SGD assuming the weakly-quasi-convexity and PL condition. In Xie et al. [2018], Zhang et al. [2018], DP-SGDA and its variants together with clipping techniques were employed to train differentially private GANs which showed promising results in applications. However, no utility analysis was given there. Boob and Guzmán [2021] focused on the noisy stochastic extragradient method with DP constraints for minimax problems in the convex-concave and smooth settings and provided its utility analysis using variational inequality (VI) and stability approaches.

2 PROBLEM FORMULATION

In this section, we introduce necessary assumptions, notations and the DP-SGDA algorithm.

2.1 ASSUMPTIONS AND NOTATIONS

Firstly, we introduce necessary assumptions and notations. A function $h : \mathcal{W} \to \mathbb{R}$ is said to be convex if, for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, there holds $h(\mathbf{w}) \geq h(\mathbf{w}') + \langle \nabla h(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \rangle$ where ∇ is the gradient operator and $\langle \cdot, \cdot \rangle$ is the inner product. Let $\|\cdot\|_2$ denote the Euclidean norm. We say h is ρ-strongly-convex if $h - \frac{\rho}{2}\|\mathbf{w}\|_2^2$ is convex, h is concave if $-h$ is convex, and ρ-strongly-concave if $-h - \frac{\rho}{2}\|\mathbf{w}\|_2^2$ is convex. Let $[n] := \{1, 2, \ldots, n\}$.

Definition 1. *Given a function* $h : \mathcal{W} \times \mathcal{V} \to \mathbb{R}$. *We say* h *is convex-concave if for any* $\mathbf{v} \in \mathcal{V}$, *the function* $\mathbf{w} \mapsto h(\mathbf{w}, \mathbf{v})$ *is convex and for any* $\mathbf{w} \in \mathcal{W}$, *the function* $\mathbf{v} \mapsto h(\mathbf{w}, \mathbf{v})$ *is concave.*

Assumption 1 (A1). *The function* f *is said to be Lipschitz continuous if there exist* $G_{\mathbf{w}}, G_{\mathbf{v}} > 0$ *such that, for any* $\mathbf{w}, \mathbf{w}' \in \mathcal{W}, \mathbf{v}, \mathbf{v}' \in \mathcal{V}$ *and* $\mathbf{z} \in \mathcal{Z}$, $\|f(\mathbf{w}, \mathbf{v}; \mathbf{z}) - f(\mathbf{w}', \mathbf{v}; \mathbf{z})\|_2 \leq G_{\mathbf{w}}\|\mathbf{w} - \mathbf{w}'\|_2$, *and* $\|f(\mathbf{w}, \mathbf{v}; \mathbf{z}) - f(\mathbf{w}, \mathbf{v}'; \mathbf{z})\|_2 \leq G_{\mathbf{v}}\|\mathbf{v} - \mathbf{v}'\|_2$. *And denote* $G = \max\{G_{\mathbf{w}}, G_{\mathbf{v}}\}$.

Assumption 2 (A2). *For randomly drawn* $j \in [n]$, *the gradients* $\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{v}; \mathbf{z}_j)$ *and* $\nabla_{\mathbf{v}} f(\mathbf{w}, \mathbf{v}; \mathbf{z}_j)$ *have bounded variances* $B_{\mathbf{w}}$ *and* $B_{\mathbf{v}}$ *respectively. And let* $B = \max\{B_{\mathbf{w}}, B_{\mathbf{v}}\}$.

Assumption 3 (A3). *The function* f *is said to be smooth if it is continuously differentiable and there exists a constant* $L > 0$ *such that for any* $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, $\mathbf{v}, \mathbf{v}' \in \mathcal{V}$ *and* $\mathbf{z} \in \mathcal{Z}$,

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

We also require the Polyak-Łojasiewicz (PL) condition.

Definition 2 (Polyak [1964])**.** *A function* $h : \mathcal{W} \to \mathbb{R}$ *satisfies the PL condition if there exist a constant* $\mu > 0$ *such that, for any* $\mathbf{w} \in \mathcal{W}$, $\frac{1}{2}\|\nabla h(\mathbf{w})\|_2^2 \geq \mu(h(\mathbf{w}) - \min_{\mathbf{w}' \in \mathcal{W}} h(\mathbf{w}'))$.

We refer to Karimi et al. [2016] for a nice discussion of this condition and other general conditions that allow the global convergence of gradient descent.

2.2 DP-SGDA ALGORITHM

We now move on to the definition of differential privacy and the description of DP-SGDA. Differential privacy was introduced by Dwork et al. [2006, 2014]. We say that two datasets S, S' are neighboring datasets if they differ by at most one example.

Definition 3 (Differential Privacy)**.** *A (randomized) algorithm* A *is called* (ϵ, δ)*-differentially private (DP) if, for all neighboring datasets* S, S' *and for all events* O *in the output space of* A, *the following holds*

$$\mathbb{P}[A(S) \in O] \leq e^{\epsilon}\mathbb{P}[A(S') \in O] + \delta.$$

Algorithm 1 Differentially Private Stochastic Gradient Descent Ascent (DP-SGDA) Method

Inputs: data $S = \{\mathbf{z}_i : i \in [n]\}$, privacy budget ϵ, δ, number of iterations T, learning rates $\{\eta_{\mathbf{w},t}, \eta_{\mathbf{v},t}\}_{t=1}^T$, and initialize $(\mathbf{w}_0, \mathbf{v}_0)$
Compute noise parameters $\sigma_{\mathbf{w}}$ and $\sigma_{\mathbf{v}}$ based on Eq. (3)
for $t = 1$ to T **do**
  Sample a mini-batch $I_t = \{i_t^1, \cdots, i_t^m \in [n]\}$ uniformly with replacement
  Sample independent noises $\xi_t \sim \mathcal{N}(0, \sigma_{\mathbf{w}}^2 I_{d_1})$ and $\zeta_t \sim \mathcal{N}(0, \sigma_{\mathbf{v}}^2 I_{d_2})$
  $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\Big(\mathbf{w}_t - \eta_{\mathbf{w},t}\Big(\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \xi_t\Big)\Big)$
  $\mathbf{v}_{t+1} = \Pi_{\mathcal{V}}\Big(\mathbf{v}_t + \eta_{\mathbf{v},t}\Big(\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}) + \zeta_t\Big)\Big)$
end for
Outputs: $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = \frac{1}{T}\sum_{t=1}^T (\mathbf{w}_t, \mathbf{v}_t)$ or $(\mathbf{w}_T, \mathbf{v}_T)$

Our aim is to design a randomized algorithm satisfying (ϵ, δ)-DP which solves the empirical minimax problem:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} \Big\{F_S(\mathbf{w}, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}, \mathbf{v}; \mathbf{z}_i)\Big\}. \quad (2)$$

Notice that in the standard ERM problem, which involves the minimization only with respect to $\mathbf{w}$, DP-SGD [Wu et al., 2017, Song et al., 2013, Bassily et al., 2019, Wang et al., 2020a, e.g.] uses the gradient perturbation at each iteration. Specifically, at each iteration of this algorithm, a randomized gradient estimated from a random subset (mini-batch) of S is perturbed by a Gaussian noise and then the model parameter is updated based on this noisy gradient.

Following the same spirit, DP-SGDA [e.g. Xie et al., 2018, Zhang et al., 2018] adds Gaussian noises per iteration to the randomized gradient mapping $(g_{\mathbf{w},t}, g_{\mathbf{v},t}) = (\frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}), \frac{1}{m}\sum_{j=1}^m \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t^j}))$ where the index of example $\mathbf{z}_{i_t^j}$ is from the mini-batch I_t. Then, the primal variable $\mathbf{w}$ is updated by gradient descent based on the noisy gradient $g_{\mathbf{w},t} + \xi_t$ and the dual variable $\mathbf{v}$ is updated by gradient ascent based on the noisy gradient $g_{\mathbf{v},t} + \zeta_t$. The pseudo-code for DP-SGDA is given in Algorithm 1. The noise levels $\sigma_{\mathbf{w}}, \sigma_{\mathbf{v}}$ are given by (3) which will be specified soon in Section 3 in order to guarantee (ϵ, δ)-DP. The notations $\Pi_{\mathcal{W}}(\cdot)$ and $\Pi_{\mathcal{V}}(\cdot)$ denote the projections to $\mathcal{W}$ and $\mathcal{V}$, respectively. From now on, the notation A denotes the DP-SGDA algorithm and its output is denoted by $A(S) = (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$.

2.3 MEASURES OF UTILITY

Since the model $A(S)$ is only trained based on the training data S, its empirical behavior as measured by F_S may not generalize well on test data. Our goal is to investigate the statistical behavior of $A(S)$ on the test data in terms of some population risk. However, unlike the standard statistical learning theory (SLT) setting where there is only a minimization of $\mathbf{w}$, we have different measures of population risk due to the minimax structure [Zhang et al., 2021a, Lei et al., 2021]. Let $\mathbb{E}[\cdot]$ denote the expectation with respect to the randomness of algorithm A and data S. We are particularly interested in the following metrics.

Definition 4 (Weak Primal-Dual (PD) Risk)**.** *The weak primal-dual population risk of $A(S)$, denoted by $\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$, is defined as*

$$\max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\big[F(A_{\mathbf{w}}(S), \mathbf{v})\big] - \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}\big[F(\mathbf{w}, A_{\mathbf{v}}(S))\big].$$

The corresponding weak PD empirical risk, denoted by $\triangle_S^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$, is defined as

$$\max_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\big[F_S(A_{\mathbf{w}}(S), \mathbf{v})\big] - \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}\big[F_S(\mathbf{w}, A_{\mathbf{v}}(S))\big].$$

Definition 5 (Primal Risk)**.** *The primal population risk of $A(S)$ is given by $R(A_{\mathbf{w}}(S)) = \max_{\mathbf{v}\in\mathcal{V}} F(A_{\mathbf{w}}(S), \mathbf{v})$ and empirical risk is defined by $R_S(A_{\mathbf{w}}(S)) = \max_{\mathbf{v}\in\mathcal{V}} F_S(A_{\mathbf{w}}(S), \mathbf{v})$, respectively. The excess primal population risk is defined as*

$$\mathbb{E}\big[R(A_{\mathbf{w}}(S)) - \min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})\big].$$

The corresponding excess primal empirical risk is then

$$\mathbb{E}\big[R_S(A_{\mathbf{w}}(S)) - \min_{\mathbf{w}\in\mathcal{W}} R_S(\mathbf{w})\big].$$

Meanwhile, the strong PD risk defined as $\triangle^s(\mathbf{w}, \mathbf{v}) = \mathbb{E}\big[\sup_{\mathbf{v}'\in\mathcal{V}} F(\mathbf{w}, \mathbf{v}') - \inf_{\mathbf{w}'\in\mathcal{W}} F(\mathbf{w}', \mathbf{v})\big]$. We have $\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \triangle^s(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ by applying Jensen's inequality. However, when F is strongly-convex-strongly-concave, the point distance from the model $(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$ to the true saddle point $(\mathbf{w}^*, \mathbf{v}^*) \in \arg\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F(\mathbf{w}, \mathbf{v})$ can be bounded by the weak PD population risk, i.e. $\mathbb{E}[\|A_{\mathbf{w}}(S) - \mathbf{w}^*\|_2^2 + \|A_{\mathbf{v}}(S) - \mathbf{v}^*\|_2^2] \leq \mathcal{O}(\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)))$. For certain problems, it is suffices to bound the weak PD risk, such as the learning problem for Markov decision process in Appendix A. The primal risk is more meaningful when one is concerned about the risk with respect to the primal variable, such as the AUC maximization problem.

3 MAIN RESULTS

In this section, we present our main theoretical results for DP-SGDA. For the privacy guarantee, we leverage the moments accountant method [Abadi et al., 2016], which implies tight privacy loss for adaptive Gaussian mechanisms with amplification by subsampling. Below we summarize a specific version of this method that suffices for our purpose.

Theorem 1. *Let **(A1)** hold true. Then, there exist constants c_1, c_2 and c_3 so that given the mini-batch size m and total iterations T, for any $\epsilon < c_1 m^2 T/n^2$, Algorithm 1 is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose*

$$\sigma_{\mathbf{w}} = \frac{c_2 G_{\mathbf{w}} \sqrt{T \log(1/\delta)}}{n\epsilon}, \ \sigma_{\mathbf{v}} = \frac{c_3 G_{\mathbf{v}} \sqrt{T \log(1/\delta)}}{n\epsilon}. \quad (3)$$

The proof of Theorem 1 is given in Appendix B.

Remark 1. *In practice, given privacy budget ϵ, δ and parameters m, T, the constant c_2 and hence σ can be found by grid search [Abadi et al., 2016]. Here we provide a set of parameters that satisfies the condition in that reference and our Theorem 1. That is, by choosing $\epsilon \leq 1, \delta \leq 1/n^2$ and $m = \max(1, n\sqrt{\epsilon/(4T)})$, then we have explicit values for the variances as $\sigma_{\mathbf{w}} = \frac{8G_{\mathbf{w}}\sqrt{T\log(1/\delta)}}{n\epsilon}, \sigma_{\mathbf{v}} = \frac{8G_{\mathbf{v}}\sqrt{T\log(1/\delta)}}{n\epsilon}$.*

Remark 2. *Our Algorithm 1 allows the application of independent noises ξ_t, ζ_t with different $\sigma_{\mathbf{w}}, \sigma_{\mathbf{v}}$, respectively. In Boob and Guzmán [2021], a uniform σ is used (Theorem 5.4 or 7.4 there) for both primal and dual variables. In many examples, the primal and dual gradients $\nabla_{\mathbf{w}} f(\mathbf{w}_t, \mathbf{v}_t, \mathbf{z}_{i_t^j}), \nabla_{\mathbf{v}} f(\mathbf{w}_t, \mathbf{v}_t, \mathbf{z}_{i_t^j})$ enjoy different Lipschitz constants (ℓ_2-sensitivity). Therefore, our treatment leads to a more delicate way of calibrating the variances of the Gaussian noises. As we shall see in the experiments in Section 4, this treatment enables Algorithm 1 to achieve better performance.*

In the subsequent subsections, we present our main contribution of this paper, i.e., the utility bounds of DP-SGDA for the convex-concave and nonconvex-strongly-concave cases, respectively.

3.1 CONVEX-CONCAVE CASE

In this subsection, we present the utility bound of DP-SGDA for the convex-concave case in terms of the weak PD risk of the output $(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ of Algorithm 1.

Theorem 2. *Assume the function f is convex-concave. Assume $\mathcal{W}$ and $\mathcal{V}$ are bounded so that $\max_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_2 \leq D_{\mathbf{w}}$, $\max_{\mathbf{v}\in\mathcal{V}} \|\mathbf{v}\|_2 \leq D_{\mathbf{v}}$. And let $D = \max\{D_{\mathbf{w}}, D_{\mathbf{v}}\}$. Let the stepsizes $\eta_{\mathbf{w},t} = \eta_{\mathbf{v},t} = \eta$ for all $t \in [T]$ with some $\eta > 0$. Under one of the condition*

*a) Assumption **(A1)** and **(A3)** hold true and we choose $T \asymp n$ and $\eta \asymp 1/\big(\max\{\sqrt{n}, \sqrt{d\log(1/\delta)}/\epsilon\}\big)$,*

*b) or Assumption **(A1)** holds true and 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

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

Its detailed proof can be found in Appendix C. The proof mainly relies on the concept of stability [Bousquet and Elisseeff, 2002, Charles and Papailiopoulos, 2018, Hardt et al., 2016, Kuzborskij and Lampert, 2018]. Specifically, the weak PD population risk can be decomposed as follows:

$$\begin{aligned}\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) =& \triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \\ &+ \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T), \end{aligned} \quad (4)$$

where the term $\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ is the generalization error and $\triangle^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)$ is the optimization error.

The estimation for the optimization error can be conducted by standard techniques [e.g. Nemirovski et al., 2009]. We give a self-contained proof in Appendix C.1. The generalization error is estimated using a concept of weak stability [Lei et al., 2021]. Specifically, we say the randomized algorithm A is *ε-weakly-stable* if, for any neighboring sets S, S' differing at one single datum, there holds

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

We know from Lei et al. [2021] that ε-weak-stability implies $\triangle^w(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) - \triangle^w_S(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \varepsilon$.

In Appendix C.2, we prove the weak stability of DP-SGDA (i.e. Algorithm 1) for both smooth and nonsmooth cases. Putting the estimations for the optimization error and generalization error into (4) can yield the bound in Theorem 2. We end this subsection with some remarks.

Remark 3. *The same optimal utility bound $\mathcal{O}\Big(\max\Big\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big\}\Big)$ was claimed in Boob and Guzmán [2021][1]. Yet our results also possess two theoretical gains compared to theirs. Firstly, when the smoothness assumption holds, Part a) in our Theorem 2 shows the optimal utility with $T = \mathcal{O}(n)$ iterations and $\mathcal{O}(n^{3/2})$ gradient computations by Remark 1, while their single-looped algorithm (Algorithm 1 there) requires $\mathcal{O}(n^2)$ gradient computations in their Theorem 5.4. They further improved the gradient complexity to $\mathcal{O}(n^{3/2}\log(n))$ in Theorem 7.4, which, however, requires an extra subroutine algorithm (inner-loop) (Algorithm 2 there). Secondly, we also derive the same optimal bound with only Lipschitz continuous assumption for the nonsmooth case which was not addressed in Boob and Guzmán [2021].*

3.2 NONCONVEX-STRONGLY-CONCAVE CASE

We proceed to the case when f is non-convex-strongly-concave. In this case, we can present utility bounds of DP-SGDA in terms of the primal excess risk, i.e., $R(\mathbf{w}_T) - \min_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})$, where $\mathbf{w}_T$ is the last iterate of Algorithm

[1] The bound was claimed in terms of the strong PD risk, which seems to be stronger than ours. However, their proof is not rigorous. See Appendix C.4 for detail.

1. Generally speaking, a saddle point may not always exist without the convexity assumption. Since our goal in this paper is to find global optima, we assume that the saddle point of the empirical minimax problem exists, i.e., there exists $(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S)$ such that, for any $\mathbf{w} \in \mathcal{W}$ and $\mathbf{v} \in \mathcal{V}$,

$$F_S(\hat{\mathbf{w}}_S, \mathbf{v}) \leq F_S(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S) \leq F_S(\mathbf{w}, \hat{\mathbf{v}}_S).$$

To estimate the primal excess risk, 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:

$$\begin{aligned}\mathbb{E}[R(\mathbf{w}_T)-R^*] &= \mathbb{E}[R(\mathbf{w}_T)-R_S(\mathbf{w}_T)]+\mathbb{E}[R_S(\mathbf{w}_T)-R_S^*] \\ &\quad + \mathbb{E}[R_S^*-R_S(\mathbf{w}^*)]+\mathbb{E}[R_S(\mathbf{w}^*)-R(\mathbf{w}^*)] \\ &\leq \mathbb{E}[R(\mathbf{w}_T)-R_S(\mathbf{w}_T)]+\mathbb{E}[R_S(\mathbf{w}^*)-R(\mathbf{w}^*)] \\ &\quad + \mathbb{E}[R_S(\mathbf{w}_T)-R_S^*], \end{aligned} \tag{5}$$

where the last inequality follows from the fact that $R_S^* - R_S(\mathbf{w}^*) \leq 0$ since $R_S^* = \min_{\mathbf{w}\in\mathcal{W}} R_S(\mathbf{w})$. 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.

Optimization Error. The next theorem characterizes the primal empirical risk of DP-SGDA under the PL-SC assumption.

Theorem 3. *Assume Assumptions **(A1)** and **(A2)** hold true, and the function $F_S(\mathbf{w}, \cdot)$ is ρ-strongly concave and $F_S(\cdot, \mathbf{v})$ satisfies μ-PL condition. Assume $\mathcal{V}$ is bounded. Let $\kappa = L/\rho$. If we choose $\eta_{\mathbf{w},t} \asymp \frac{1}{\mu t}$ and $\eta_{\mathbf{v},t} \asymp \frac{\kappa^{2.5}}{\mu^{1.5} t^{2/3}}$, then*

$$\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] = \mathcal{O}\Big(\frac{\kappa^{3.5}}{\mu^{2.5}}\Big(\frac{1/m + d(\sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{v}}^2)}{T^{2/3}}\Big)\Big).$$

We provide the proof of Theorem 3 in Appendix D.1. In the non-private setting, i.e. $\sigma_{\mathbf{w}} = \sigma_{\mathbf{v}} = 0$, Theorem 3 implies that the convergence rate in terms of the primal empirical risk is of the order $\mathcal{O}(\frac{\kappa^{3.5}}{\mu^{2.5} T^{2/3}})$, which is a new result even in the non-private case as far as we are aware of.

In Lin et al. [2020], the local convergence of SGDA in the non-private case was proved in terms of the metric $\mathbb{E}_\tau[\|\nabla R_S(\mathbf{w}_\tau)\|_2^2]$ where τ is chosen uniformly at random from the set $\{1, 2, \ldots, T\}$. Our analysis is much more involved since it proves the global convergence of the last iterate $\mathbf{w}_T$. Our main idea is to prove the coupled recursive inequalities for two terms, i.e., $a_t = R_S(\mathbf{w}_t) - R_S^*$ and $b_t = \|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2$ where $\hat{\mathbf{v}}_S(\mathbf{w}_t) = \arg\max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}_t, \mathbf{v})$, and then carefully derive the the convergence rate for $a_t + \lambda_t b_t$ by choosing λ_t appropriately. The convergence rate and its proof can be of interest in their own right. One can find more detailed arguments in Appendix D.1.

Generalization Error. We present the bound for the generalization error which is proved again using the stability approach.

We begin with a discussion of the saddle points. While the saddle point $(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S)$ may not be unique, $\hat{\mathbf{v}}_S$ must be unique if $F_S(\mathbf{w}, \mathbf{v})$ is strongly-concave in $\mathbf{v}$ (see Proposition 1 in Appendix D). Therefore, we can define $\pi_S(\mathbf{w})$ the projection of $\mathbf{w}$ to the set of saddle points, as $\Omega_S = \big\{\hat{\mathbf{w}}_S : (\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S) \in \arg\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}, \mathbf{v})\big\} = \big\{\hat{\mathbf{w}}_S : \hat{\mathbf{w}}_S \in \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w}, \hat{\mathbf{v}}_S)\big\}$.

Recall that $\mathbf{w}_T$ is the iterate of DP-SGDA at time T based on the training data S. Likewise, we denote by $\mathbf{w}'_T$ based on the training set S' which differs from S at one single datum. Due to the possibly multiple saddle points, we need the following critical assumption for estimating the generalization error.

Assumption 4 (**A4**). *For the (randomized) algorithm DP-SGDA, assume that $\pi_{S'}(\pi_S(\mathbf{w}_T)) = \pi_{S'}(\mathbf{w}'_T)$ for any neighboring sets S and S'.*

Assumption **(A4)** was introduced in Charles and Papailiopoulos [2018] for studying the stability of SGD in the non-convex case which only involves the minimization over $\mathbf{w}$. In our case, **(A4)** holds true whether the saddle point is unique (e.g., F_S is strongly-convex and strongly-concave) or the two sets of saddle points based on S and S', i.e. Ω_S and $\Omega_{S'}$ do not change too much. Since our algorithm satisfies (ϵ, δ)-DP it means that the distributions of $\mathbf{w}_T$ and $\mathbf{w}'_T$ generated from two neighboring sets S and S' are "close", which indicates $\sup_{S,S'} \|\pi_{S'}(\pi_S(\mathbf{w}_T)) - \pi_{S'}(\mathbf{w}'_T)\|_2$ can be small. Proving such statement serves as an interesting open problem.

Now we can state the results on the generalization error.

Theorem 4 (Generalization Error)**.** *Assume Assumptions **(A1)**, **(A3)** and **(A4)** hold true, and assume the function $f(\mathbf{w}, \cdot; \mathbf{z})$ is ρ-strongly concave and $F_S(\cdot, \mathbf{v})$ satisfies μ-PL condition. Let $\kappa = L/\rho$. If $\mathbb{E}[R_S(\mathbf{w}_T) - R_S^*] \leq \varepsilon_T$, then*

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

and

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

The proof of Theorem 4 is provided in Appendix D.2.

Remark 4. *The generalization error bounds given in Theorem 4 indicate that if the optimization error $\mathbb{E}[R_S(\mathbf{w}_T) - R_S^*]$ is small then the generalization error will be small. This is consistent with the observation in the stability and*

generalization analysis of SGD [Charles and Papailiopoulos, 2018, Hardt et al., 2016, Lei and Ying, 2021] for the minimization problems in the sense of "optimization can help generalization".

We can derive the following utility bound for DP-SGDA by combining the results in Theorems 4 and 3.

Theorem 5. *Under the same assumptions of Theorem 4, if we choose* $T \asymp n$, $\eta_{\mathbf{w},t} \asymp \frac{1}{\mu t}$ *and* $\eta_{\mathbf{v},t} \asymp \frac{\kappa^{2.5}}{\mu^{1.5} t^{2/3}}$, *then*

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

The proof can be found in Appendix D.3.

4 EXPERIMENTS

In this section, we evaluate the performance of DP-SGDA by taking AUC maximization as an example. Due to space limitation, we present the most significant information and results of our experiments while more detailed information and additional results are given in Appendix E and F.

4.1 EXPERIMENTAL SETTINGS

Baseline Model. We perform experiments on the problem of AUC maximization with the least square loss to evaluate the DP-SGDA algorithm in linear and non-linear settings (two-layer multilayer perceptron (MLP)). In this case, AUC maximization 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. As shown in Ying et al. [2016], it is equivalent to a minimax problem:

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

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 $p = \mathbb{P}[y = 1]$.

When h is a linear function, the AUC learning objective above is convex-strongly-concave. On the other hand, when h is a MLP function, it becomes a nonconvex-strongly-concave minimax problem. In addition, following Liu et al. [2020], we use Leaky ReLU as an activation function for MLP. It was shown in their paper the empirical AUC objective satisfies the PL condition with this choice of h. Without a special statement, we set 256 as the number of hidden units in MLP and 64 as the mini-batch size during the training.

Datasets and Evaluation Metrics. Our experiments are based on three popular datasets, namely ijcnn1 [Chang and Lin, 2011], MNIST [LeCun et al., 1998], and Fashion-MNIST [Xiao et al., 2017] that have been used in previous studies. For MNIST and Fashion-MNIST, following Gao et al. [2013], Ying et al. [2016], we transform their classes into binary classes by randomly partitioning the data into two groups, each with an equal number of classes. For ijcnn1, we randomly split its original training set into new training (80%) and testing (20%) sets. For MNIST and Fashion-MNIST, we use their original training set and testing set. For each method, the reported performance is obtained by averaging the AUC scores on the test set according to 5 random seeds (for initial $\mathbf{w}$ and $\mathbf{v}$, sampling and noise generation).

Privacy Budget Settings. In the experiments, we set up five privacy levels from small to large: $\epsilon \in \{0.1, 0.5, 1, 5, 10\}$. We also consider three different δ from $\{1\mathrm{e}{-4}, 1\mathrm{e}{-5}, 1\mathrm{e}{-6}\}$. Due to space limitation, we only report the performance when $\delta = 1\mathrm{e}{-6}$. More results can be found in Appendix F. To estimate the Lipschitz constants $G_{\mathbf{w}}$ and $G_{\mathbf{v}}$ (in Theorem 1), we first run the algorithms without adding noise. Then we calculate the maximum gradient norms of AUC loss w.r.t $\mathbf{w}$ and $\mathbf{v}$ and assign them as $G_{\mathbf{w}}$ and $G_{\mathbf{v}}$, respectively. According to these parameters, we calculate the noise parameter σ by applying autodp[2], which is widely used in the existing works Wang et al. [2019b].

Compared Algorithms. Boob and Guzmán [2021] is the only existing paper that considers differential privacy in the convex-concave minimax problem. Therefore, we use their single-loop NSEG algorithm as our baseline method on the AUC optimization under the linear setting.

4.2 RESULTS

General AUC Performance vs Privacy. The general performance of all algorithms under linear and MLP settings of AUC optimization is shown in Table 2. Since the standard deviation of the AUC performance is around $[0, 0.1\%]$ and the difference between different algortihms is very small, we only report the average AUC performance. First, without adding noise into gradients, we can find the NSEG method and our DP-SGDA method have similar performance under the linear case. Furthermore, we can find the performance of the DP-SGDA with MLP model can outperform linear models on all datasets. This is because the non-linear model can learn more information among features than linear models. Second, by adding noise into the gradients, we can find the AUC performance of all models is decreased on all datasets. However, by increasing the privacy budget ϵ, the AUC performance is increased. The reason is that ϵ and σ have opposite trends according to equation (3). The relation between ϵ and AUC score also verifies our Theorem 2 and Theorem 5. Third, to verify our statement in Remark 2, we compare the σ values from NSEG and DP-SGDA on all datasets in Figure 1(a). From the figure, it is clear that the σ from NSEG is larger than ours in all ϵ settings since it is calibrated based on the gradients' sensitivity from both

[2] https://github.com/yuxiangw/autodp

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
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 2: *Comparison of AUC performance in NSEG and DP-SGDA (Linear and MLP settings) on three datasets with different ϵ and δ=1e-6. The "Original" means no noise ($\epsilon = \infty$) is added in the algorithms.*

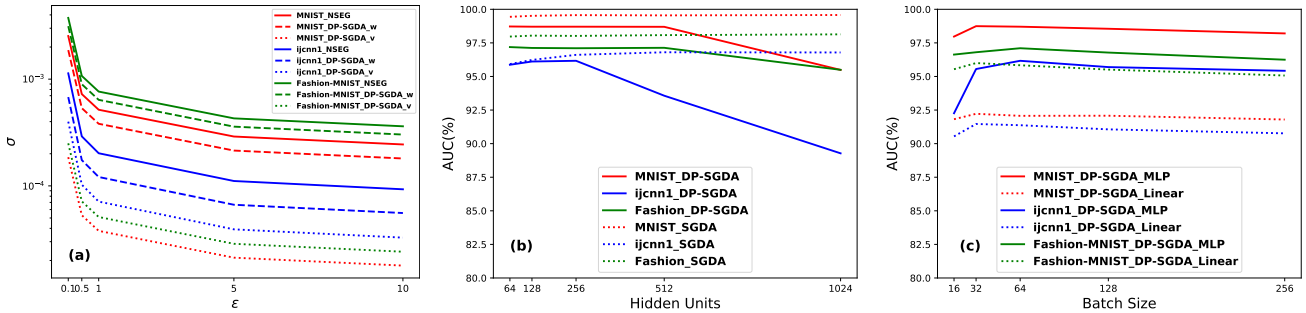

Figure 1: *(a) Comparison of σ for NSEG and DP-SGDA (Linear setting) on three datasets with different ϵ and δ=1e-6. (b)Comparison of AUC performance for SGDA and DP-SGDA in MLP settings on three datasets with different hidden units and ϵ=1 and δ=1e-6. (c) Comparison of AUC performance for DP-SGDA (Linear and MLP settings) on three datasets with different batch size and ϵ=1 and δ=1e-6.*

$\mathbf{w}$ and $\mathbf{v}$. In fact, the sensitivity w.r.t. $\mathbf{v}$ is small as it is a one-dimensional variable for AUC maximization. Therefore, NSEG leads to overestimate on the noise addition towards $\mathbf{v}$. From Table 2 we observe our DP-SGDA achieves better AUC score than NSEG under the same privacy budget.

Different Hidden Units. In DP-SGDA under the MLP setting, the hidden unit is one of the most important factors affecting the model performance. Therefore, we compare the AUC performance with respect to the different hidden units in Figure 1(b). If we provide a small number of hidden units, the model will suffer from poor generalization capability. Using a large number of hidden units will make the model easier to fit the training set. For SGDA (non-private) training, it is often helpful to apply a large number of hidden units, as long as the model does not overfit. In agreement with this intuition, we find the model performance improves with increasing hidden units in Figure 1(b). However, for DP-SGDA training, more hidden units increase the sensitivity of the gradients, which leads to more noise added at each update. Therefore, in contrast to the non-private setting, we find the AUC performance decreases when the number of hidden units increases.

Different Mini-Batch Size. From Theorem 1 and Theorem 3, we find mini-batch size can influence the Gaussian noise variances $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{v}}^2$ as well as the convergence rate. Selecting the mini-batch size must balance two conflicting objectives. On one hand, a small mini-batch size may lead to sub-optimal performance. On the other hand, for large batch sizes, the added noise has a smaller relative effect. Therefore, we show the AUC score for DP-SGDA with different mini-batch sizes in Figure 1(c). The experimental results show that the mini-batch size has a relatively large impact on the AUC performance when the mini-batch size is small.

5 CONCLUSION

In this paper, we have used algorithmic stability to conduct utility analysis of the DP-SGDA algorithm for minimax problems under DP constraints. For the convex-concave setting, we proved that DP-SGDA can attain an optimal rate $\mathcal{O}(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})$ in terms of the weak primal-dual population risk while providing (ϵ, δ)-DP for both smooth and nonsmooth cases. For the nonconvex-strongly-concave case, assuming that the empirical risk satisfies the PL condition we proved the excess primal population risk of DP-SGDA can achieve a utility bound $\mathcal{O}\big(\frac{1}{n^{1/3}} + \frac{\sqrt{d\log(1/\delta)}}{n^{5/6}\epsilon}\big)$. Experiments on three benchmark datasets illustrate the effectiveness of DP-SGDA.

For future work, it would be interesting to improve the utility bound for the nonconvex-strongly-convex setting. It also remains unclear to us how to establish the utility bound for DP-SGDA when gradient clipping techniques are enforced at each iteration.

References

Martin Abadi et al. Deep learning with differential privacy. In *ACM CCS*, pages 308–318, 2016.

John M Abowd. The challenge of scientific reproducibility and privacy protection for statistical agencies. *Census Scientific Advisory Committee*, 2016.

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

Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Raef Bassily et al. Private stochastic convex optimization with optimal rates. In *NeurIPS*, pages 11279–11288, 2019.

Raef Bassily et al. Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*, 2020.

Digvijay Boob and Cristóbal Guzmán. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *arXiv preprint arXiv:2104.02988*, 2021.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2(Mar):499–526, 2002.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *TIST*, 2(3):27, 2011.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *ICML*, pages 745–754. PMLR, 2018.

Emily Diana et al. Minimax group fairness: Algorithms and experiments. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *arXiv preprint arXiv:1712.01524*, 2017.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork et al. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *ACM CCS*, pages 1054–1067, 2014.

Beaulieu-Jones et al. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.

Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *ICML*, pages 3174–3185. PMLR, 2021.

Vitaly Feldman et al. Private stochastic convex optimization: optimal rates in linear time. In *STOC*, pages 439–449, 2020.

Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass auc optimization. In *ICML*, pages 906–914. PMLR, 2013.

Ian Goodfellow et al. Generative adversarial nets. *NeurIPS*, 27, 2014.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, pages 1225–1234, 2016.

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

Hamed Karimi et al. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML PKDD*, pages 795–811, 2016.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *ICML*, pages 2815–2824. PMLR, 2018.

Yann LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *ICML*, pages 5809–5819. PMLR, 2020.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *ICLR*, 2021.

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.

Luo Luo et al. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *arXiv preprint arXiv:2001.03724*, 2020.

Natalia Martinez et al. Minimax pareto fairness: A multi objective perspective. In *ICML*, pages 6755–6764, 2020.

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

Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

Arkadi Nemirovski et al. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Maher Nouiehed et al. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17–32, 1964.

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

Hassan Rafique et al. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35, 2021.

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

Shai Shalev-Shwartz et al. Stochastic convex optimization. In *COLT*, 2009.

Aman Sinha et al. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Shuang Song et al. Stochastic gradient descent with differentially private updates. In *GlobalSIP*, pages 245–248. IEEE, 2013.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *ICML*, pages 6526–6535. PMLR, 2019a.

Di Wang et al. On differentially private stochastic convex optimization with heavy-tailed data. In *ICML*, pages 10081–10091. PMLR, 2020a.

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, Yunwen Lei, Yiming Ying, and Hai Zhang. Differentially private sgd with non-smooth loss. *Applied and Computational Harmonic Analysis (ACHA)*, 2021a.

Puyu Wang et al. Differentially private empirical risk minimization for auc maximization. *Neurocomputing*, 461: 419–437, 2021b.

Serena Wang et al. Robust optimization for fairness with noisy protected groups. In *NeurIPS*, volume 33, 2020b.

Yu-Xiang Wang et al. Subsampled Rényi differential privacy and analytical moments accountant. In *AISTATS*, pages 1226–1235. PMLR, 2019b.

Xi Wu et al. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *SIGMOD*, pages 1307–1322, 2017.

Han Xiao et al. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Liyang Xie et al. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Huan Xu et al. Robustness and regularization of support vector machines. *JMLR*, 10(7), 2009.

Yan Yan et al. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *arXiv preprint arXiv:2002.05309*, 2020.

Yiming. Ying et al. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems*, 2016.

Junyu Zhang et al. Generalization bounds for stochastic saddle point problems. In *AISTATS*, pages 568–576. PMLR, 2021a.

Qiuchen Zhang et al. Private stochastic non-convex optimization with improved utility rates. In *IJCAI*, pages 3370–3376, 2021b.

Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*, 2018.

Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbao Yang. Online auc maximization. In *ICML*, 2011.

Yingxue Zhou et al. Private stochastic non-convex optimization: Adaptive algorithms and tighter generalization bounds. *arXiv preprint arXiv:2006.13501*, 2020.

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 $p=\mathbb{P}[y=1]$. Differential privacy has been applied to learn private classifier by optimizing AUC Wang et al. [2021b]. 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

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

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, 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. [2021a] for a discussion on the measure of population risk.

Robust Optimization and Fairness. The aim is to minimize the worst population risks among multiple scenarios:

$$\min_{\mathbf{w}\in\mathcal{W}}L(\mathbf{w})=\max_{1\le i\le m}\big\{\mathbb{E}_{\mathbf{z}}[\ell_1(\mathbf{w};\mathbf{z})],\cdots,\mathbb{E}_{\mathbf{z}}[\ell_m(\mathbf{w};\mathbf{z})]\big\}$$

This problem can be reformulated as the following stochastic minimax problem

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

where $\Delta_m=\big\{v\in\mathbb{R}^n:v_i\ge0,\sum_{i=1}^m v_i=1\big\}$ 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 6. *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 7 (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 [3] 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 B.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*

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

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

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

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

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:

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

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*

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

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

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

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

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

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

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

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

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

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

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

We can take an expectation over both sides of (C.2) and get

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

where we used that the variance $\mathbb{E}[\|\xi_t\|_2^2] = d_1\sigma_{\mathbf{w}}^2$, the unbiasedness $\mathbb{E}[\langle\mathbf{w}_t,\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j})-\nabla_{\mathbf{w}}F_S(\mathbf{w}_t,\mathbf{v}_t)\rangle]=0$ and the independence $\mathbb{E}[\langle\frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;\mathbf{z}_{i_t^j}),\xi_t\rangle]=0$, $\mathbb{E}[\langle\mathbf{w}-\mathbf{w}_t,\xi_t\rangle]=0$. According to Jensen's inequality and $G_{\mathbf{w}}$-Lipschitz continuity we further derive

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

Plugging the above estimate into (C.3) we arrive

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

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

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

In a similar way, we can show that

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

Combining (C.4) and (C.5) with the fact that $D=\max\{D_{\mathbf{w}},D_{\mathbf{v}}\}$, $G=\max\{G_{\mathbf{w}},G_{\mathbf{v}}\}$ and $d=\max\{d_1,d_2\}$, we further derive

$$\mathbb{E}_A[\triangle_S^s(\bar{\mathbf{w}}_T,\bar{\mathbf{v}}_T)]=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_A[\sup_{\mathbf{v}\in\mathcal{V}}F_S(\bar{\mathbf{w}}_T,\mathbf{v})-\inf_{\mathbf{w}\in\mathcal{W}}F_S(\mathbf{w},\bar{\mathbf{v}}_T)]\leq\frac{\eta G^2}{2}+\frac{D^2}{\eta T}+\frac{DG}{\sqrt{mT}}+\eta d(\sigma_{\mathbf{w}}^2+\sigma_{\mathbf{v}}^2). \tag{C.6}$$

The proof is complete. □

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*

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

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

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

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,
\end{aligned} \tag{C.7}$$

where in the last inequality we used the elementary inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$ $(p > 0)$. 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

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

Applying this inequality recursively, we derive

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

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

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

By taking $p=n/t$ we get

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

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

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

According to Lemma 4 we know

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

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

$$\begin{aligned}\left\|\begin{pmatrix}\mathbf{w}_{t+1}-\mathbf{w}'_{t+1}\\ \mathbf{v}_{t+1}-\mathbf{v}'_{t+1}\end{pmatrix}\right\|_2^2 &\leq \left\|\begin{pmatrix}\mathbf{w}_t-\frac{\eta}{m}\sum_{j=1}^m\nabla_{\mathbf{w}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t^j})-\eta\xi_t-\mathbf{w}'_t+\frac{\eta}{m}\sum_{j=1}^m\nabla_{\mathbf{w}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t^j})+\eta\xi_t \\ \mathbf{v}_t+\frac{\eta}{m}\sum_{j=1}^m\nabla_{\mathbf{v}}f(\mathbf{w}_t,\mathbf{v}_t;z_{i_t^j})+\eta\zeta_t-\mathbf{v}'_t-\frac{\eta}{m}\sum_{j=1}^m\nabla_{\mathbf{v}}f(\mathbf{w}'_t,\mathbf{v}'_t;z_{i_t^j})-\eta\zeta_t\end{pmatrix}\right\|_2^2 \\ &\leq \left\|\begin{pmatrix}\mathbf{w}_t-\mathbf{w}'_t\\ \mathbf{v}_t-\mathbf{v}'_t\end{pmatrix}\right\|_2^2+4(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)\eta^2. \qquad \text{(C.8)}\end{aligned}$$

Combining the preceding inequality with (C.7) and using the probability of $n \notin I_t$, we derive

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

Applying this inequality recursively implies that

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

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

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

According to Lemma 4 we know

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

□

C.3 PROOF OF THEOREM 2

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

Theorem C.1 (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

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

and by Lemma 3 we know

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

Combining the above two quantities we have

$$\begin{aligned}\triangle^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) \leq & \frac{4\sqrt{e(T+T^2/n)}(G_{\mathbf{w}}+G_{\mathbf{v}})^2\eta\exp(L^2T\eta^2/2)}{\sqrt{n}} + \frac{\eta(G_{\mathbf{w}}^2+G_{\mathbf{v}}^2)}{2} + \frac{D_{\mathbf{w}}^2+D_{\mathbf{v}}^2}{2\eta T} \\ & + \frac{D_{\mathbf{w}}G_{\mathbf{w}}+D_{\mathbf{v}}G_{\mathbf{v}}}{\sqrt{mT}} + \eta d(\sigma_{\mathbf{w}}^2+\sigma_{\mathbf{v}}^2). \end{aligned} \tag{C.9}$$

Furthermore, by Theorem 1, we know

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

Plugging it back into (C.9) 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 $\eta \asymp 1/\Big(L\max\{\sqrt{n}, \sqrt{d\log(1/\delta)}/\epsilon\}\Big)$ 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

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

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 $\eta \asymp 1/\Big(n\max\{\sqrt{n}, \sqrt{d\log(1/\delta)}/\epsilon\}\Big)$ 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. □

C.4 CLARIFYING REMARK 3

The same optimal utility bound in terms of VI gap was claimed in Theorem 5.4 and Theorem 7.4 of Boob and Guzmán [2021], i.e.

$$\mathbb{E}[\sup_{w\in\mathcal{W}}\langle F(w), u-w\rangle] = \mathcal{O}\Big(\max\Big\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\Big\}\Big).$$

By the monotonicity assumption, their results imply the same bound on the primal-dual risk $\triangle(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) = \mathbb{E}[\max_{\mathbf{v}} F(A_{\mathbf{w}}(S), \mathbf{v}) - \min_{\mathbf{w}} F(\mathbf{w}, A_{\mathbf{v}}(S))]$ of the minimax problem (1). Such measure of generalization seems to be stronger than ours since $\triangle^{w}(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)) \leq \triangle(A_{\mathbf{w}}(S), A_{\mathbf{v}}(S))$. However, the discussion between stability and generalization there is not rigorous in Theorem 4.1 (i.e. stability implies generalization for SVI). Indeed, borrowing their notations, equation (4.3) from their paper, i.e.,

$$\mathbb{E}_{\beta_i}\langle F(u;\beta_i), \mathcal{A}(S) - u\rangle \leq \mathbb{E}_{\beta_j}\langle F(u;\beta_j), \mathcal{A}(S) - u\rangle + M\mathbb{E}_{\beta_j'}[\|\mathcal{A}(S_j^i) - \mathcal{A}(S^j)\|],$$

only holds when u is independent of β_j. But applying (4.3) with $u = \hat{u}_i = \arg\max_{u\in\mathcal{W}} \mathbb{E}_{\beta_i}\langle F(u;\beta_i), \mathcal{A}(S) - u\rangle$ to get the second inequality in (4.5) of their paper is problematic because now $u = \hat{i}_i$ involves sample S and after taking expectation of β_i, it depends on β_j.

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

$$\|\nabla R_S(\mathbf{w})\|_2^2 \geq 2\mu\big(F_S(\mathbf{w}, \hat{\mathbf{v}}_S(\mathbf{w})) - \min_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}', \hat{\mathbf{v}}_S(\mathbf{w}))\big). \tag{D.1}$$

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

$$\min_{\mathbf{w}'\in\mathcal{W}} F_S(\mathbf{w}', \hat{\mathbf{v}}_S(\mathbf{w})) \leq \min_{\mathbf{w}'\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{V}} F_S(\mathbf{w}', \mathbf{v}) = \min_{\mathbf{w}'\in\mathcal{W}} R_S(\mathbf{w}') \tag{D.2}$$

Combining equation (D.1) and (D.2), 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*

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

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

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

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

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

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*

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

Proof. By Young's inequality, we have

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

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

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

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

$$\begin{aligned}
\mathbb{E}_t[\|\hat{\mathbf{v}}_S(\mathbf{w}_{t+1}) - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] \leq& 2L^4/\rho^2\eta_{\mathbf{w},t}^2\|\hat{\mathbf{v}}_S(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + 4L^2/\rho^2(L + L^2/\rho)\eta_{\mathbf{w},t}^2(R_S(\mathbf{w}_t) - R_S(\mathbf{w}^*)) \\
&+ L^2/\rho^2\eta_{\mathbf{w},t}^2(\frac{B_{\mathbf{w}}^2}{m} + d\sigma_{\mathbf{w}}^2). \qquad \text{(D.3)}
\end{aligned}$$

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

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

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

$$\begin{aligned}
\mathbb{E}_t[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2] \leq& (1 - \rho\eta_{\mathbf{v},t})\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 - \frac{\eta_{\mathbf{v},t}}{L}\|\nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \eta_{\mathbf{v},t}^2\|\nabla_{\mathbf{v}} F_S(\mathbf{w}_t, \mathbf{v}_t)\|_2^2 + \eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2) \\
\leq& (1 - \rho\eta_{\mathbf{v},t})\|\mathbf{v}_t - \hat{\mathbf{v}}_S(\mathbf{w}_t)\|_2^2 + \eta_{\mathbf{v},t}^2(\frac{B_{\mathbf{v}}^2}{m} + d\sigma_{\mathbf{v}}^2). \qquad \text{(D.4)}
\end{aligned}$$

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

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

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*

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

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

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

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 D.1 (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*

$$\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] = \mathcal{O}(\min\left\{\frac{1}{L}, \frac{1}{\mu}\right\}(\frac{B_{\mathbf{w}}^2/m + d\sigma_{\mathbf{w}}^2}{T^{2/3}}) + \max\left\{1, \sqrt{\frac{L\kappa}{\mu}}\right\}\frac{L\kappa^3}{\mu^2}(\frac{B_{\mathbf{v}}^2/m + d\sigma_{\mathbf{v}}^2}{T^{2/3}})). \tag{D.5}$$

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

$$\begin{aligned}
&\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] \\
=&\mathcal{O}(\min\left\{\frac{1}{L}, \frac{1}{\mu}\right\}(\frac{B_{\mathbf{w}}^2}{mT^{2/3}} + \frac{G_{\mathbf{w}}^2 dT^{1/3}\log(1/\delta)}{n^2\epsilon^2}) + \max\left\{1, \sqrt{\frac{L\kappa}{\mu}}\right\}\frac{L\kappa^3}{\mu^2}(\frac{B_{\mathbf{v}}^2}{mT^{2/3}} + \frac{G_{\mathbf{v}}^2 dT^{1/3}\log(1/\delta)}{n^2\epsilon^2})).
\end{aligned} \tag{D.6}$$

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

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

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

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

where we used $\mu \leq 2L^2$. 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

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

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

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

Consequently,

$$\begin{aligned}
\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] \leq &a_{T+1} + \lambda_{T+1}b_{T+1} \\
\leq &\frac{2(L + L^2/\rho)(B_{\mathbf{w}}^2/m + d\sigma_{\mathbf{w}}^2)}{\mu^2}\frac{\log(T)}{T} + \frac{32(B_{\mathbf{w}}^2/m + d\sigma_{\mathbf{w}}^2)L^4/\rho^3\min\{\frac{\sqrt{\mu}}{8\kappa^2\sqrt{L+L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}^2}{\mu\rho}\frac{1}{T^{2/3}} \\
&+ \frac{16(B_{\mathbf{v}}^2/m + d\sigma_{\mathbf{v}}^2)L^2\max\{8\kappa^2\sqrt{(L+L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}}{2\mu^2\rho}\frac{1}{T^{2/3}}. \qquad \text{(D.7)}
\end{aligned}$$

Therefore, the estimation (D.5) follows from the fact that $\kappa = L/\rho$.

The result in Theorem 3 follows by observing $\max\left\{1, \sqrt{\frac{L\kappa}{\mu}}\right\}\frac{L\kappa^3}{\mu^2} \geq \min\left\{\frac{1}{L}, \frac{1}{\mu}\right\}$. 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 (D.5) yields the desired estimation (D.6). □

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*

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

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

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

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

$$F_S(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S) \geq F_S(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}'_S) + \frac{\rho}{2}\|\hat{\mathbf{v}}_S - \hat{\mathbf{v}}'_S\|_2^2 \geq F_S(\hat{\mathbf{w}}'_S, \hat{\mathbf{v}}'_S) + \frac{\rho}{2}\|\hat{\mathbf{v}}_S - \hat{\mathbf{v}}'_S\|_2^2, \tag{D.8}$$

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

$$F_S(\hat{\mathbf{w}}'_S, \hat{\mathbf{v}}'_S) \geq F_S(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S) + \frac{\rho}{2}\|\hat{\mathbf{v}}_S - \hat{\mathbf{v}}'_S\|_2^2. \tag{D.9}$$

Adding (D.8) and (D.9) 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

$$\begin{aligned}4\mu\|\pi_S(A_{\mathbf{w}}(S)) - \pi_{S'}(A_{\mathbf{w}}(S'))\|_2^2 \leq& F_S(\pi_{S'}(A_{\mathbf{w}}(S')), \hat{\mathbf{v}}_S) - F_S(\pi_S(A_{\mathbf{w}}(S)), \hat{\mathbf{v}}_S) \\ &+ F_{S'}(\pi_S(A_{\mathbf{w}}(S)), \hat{\mathbf{v}}_{S'}) - F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')), \hat{\mathbf{v}}_{S'}).\end{aligned} \tag{D.10}$$

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

$$\begin{aligned}\rho\|\hat{\mathbf{v}}_S - \hat{\mathbf{v}}_{S'}\|_2^2 \leq & F_S(\pi_S(A_{\mathbf{w}}(S)), \hat{\mathbf{v}}_S) - F_S(\pi_S(A_{\mathbf{w}}(S)), \hat{\mathbf{v}}_{S'}) \\ & + F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')), \hat{\mathbf{v}}_{S'}) - F_{S'}(\pi_{S'}(A_{\mathbf{w}}(S')), \hat{\mathbf{v}}_S). \end{aligned} \tag{D.11}$$

Summing up (D.10) and (D.11) 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 D.2. *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 D.1, 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 D.3. *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*

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

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 $\mathbb{E}\big[F(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F(\mathbf{w}^*, \mathbf{v}^*)\big] \leq 0$ 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

$$\begin{aligned}
F_S(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F_S(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*) =& \frac{1}{n}\sum_{j\neq i}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_j) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_j)\Big) + \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i)\Big) \\
=& \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i') - f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i')\Big) + \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i)\Big) \\
&+ F_{S^{(i)}}(\mathbf{w}^*, \hat{\mathbf{v}}_S^*) - F_{S^{(i)}}(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*) \\
\leq& \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i') - f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i')\Big) + \frac{1}{n}\Big(f(\mathbf{w}^*, \hat{\mathbf{v}}_S^*; z_i) - f(\mathbf{w}^*, \hat{\mathbf{v}}_{S^{(i)}}^*; z_i)\Big) \\
\leq& \frac{2G_{\mathbf{v}}}{n}\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}_{S^{(i)}}^*\big\|_2,
\end{aligned} \tag{D.12}$$

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

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

Combining it with (D.12) we get $\big\|\hat{\mathbf{v}}_S^* - \hat{\mathbf{v}}_{S^{(i)}}^*\big\|_2 \leq 4G_{\mathbf{v}}/(\rho n)$. By Lipschitz continuity, the following inequality holds for any z

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

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

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

The proof is complete. □

Theorem D.4 (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 $\mathbb{E}[R_S(\mathbf{w}_{T+1}) - R_S^*] \leq \varepsilon_T$, then*

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

and

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

Proof. It follows directly from Theorem D.2 and D.3. □

D.3 PROOF OF THEOREM 5

Theorem D.5 (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

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

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 D.2, 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 D.3, 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 F.1. 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 E.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 E.2.

Methods	Datasets	Batch Size	Learning Rate				Epochs		Projection Size	
			Ori		DP		Ori	DP	Ori	DP
			$\mathbf{w}$	$\mathbf{v}$	$\mathbf{w}$	$\mathbf{v}$				
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 E.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

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

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

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

The pseudo-code can be found in Algorithm 2.

Algorithm 2 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. (E.1)

  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

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

$$\mathbf{v}_{t+1} = \Pi\Big\{ \mathbf{v}_t + \lambda_t \big(\frac{1}{m} \sum_{j \in I_t} \nabla_{\mathbf{v}} f(\theta_t, a_t, b_t, \mathbf{v}_t; \mathbf{z}_j) + \zeta_t \big) \Big\}$$

end for

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

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 F.1. From Table F.1, 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 F.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.

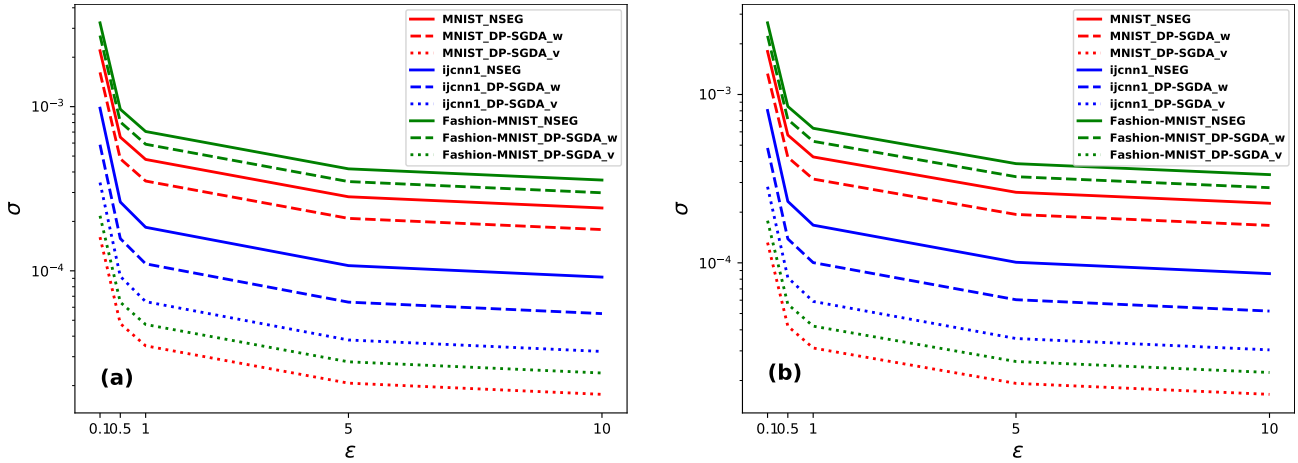

Figure F.1: *Comparison of σ in NSEG and DP-SGDA (with Linear setting) on three datasets with different ϵ and (a) δ=1e-5 and (b) δ=1e-4.*

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 F.1: *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.*