Stability and differential privacy of stochastic gradient methods

Zhenhuan Yang

University at Albany, State University of New York, zhenhuan.yang@hotmail.com

The University at Albany community has made this article openly available.

Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Applied Mathematics Commons, Computer Sciences Commons, and the Statistics and Probability Commons

Recommended Citation

https://scholarsarchive.library.albany.edu/legacy-etd/3059

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
To my parents
ABSTRACT

Recently there are a considerable amount of work devoted to the study of the algorithmic stability as well as differential privacy (DP) for stochastic gradient methods (SGM). However, most of the existing work focus on the empirical risk minimization (ERM) and the population risk minimization problems. In this paper, we study two types of optimization problems that enjoy wide applications in modern machine learning, namely the minimax problem and the pairwise learning problem.

For minimax problem, we establish a quantitative connection between stability and generalization for minimax learners in different forms including weak/strong primal-dual generalization, primal generalization and generalization with high probability. For the technical contributions, we introduce novel decompositions to handle the correlation between the primal model and dual model for connecting stability and generalization. We establish stability bounds of stochastic gradient descent ascent (SGDA) for convex-concave problems, from which we derive its optimal population risk bounds under an appropriate early-stopping strategy. We consider several measures of generalization and show that the optimal population risk bounds can be derived even in the nonsmooth case. To the best of our knowledge, our results are the first-ever known population risk bounds for minimax problems in the nonsmooth setting and the high-probability format. We further extend our analysis to the nonconvex-nonconcave setting and give the first generalization bounds for nonsmooth objective functions. Our analysis relaxes the range of step size for a controllable stability and implies meaningful primal population risk bounds under some regularity assumptions of objective functions, e.g., a decay of weak-convexity-weak-concavity parameter along the optimization process or a two-sided Polyak-Łojasiewicz (PL) condition.

We analyze the privacy and utility of differentially private stochastic gradient descent ascent (DP-SGDA) under the convex-concave setting in terms of the weak primal-dual population risk. Specifically, we show that it can guarantee $(\epsilon, \delta)$-DP and achieve the optimal rate for smooth and nonsmooth cases. To our best knowledge, this is the first-ever known result for DP-SGDA in the nonsmooth case. We further provide its utility analysis in the nonconvex-strongly-concave setting which is the first-ever-known result in terms of the pri-
mal population risk. The key techniques involve the convergence analysis and the stability analysis of the coupled minimax players. As far as we are aware of, these results are the first ones known for DP-SGDA in the nonconvex setting, which are of interest in their own rights.

For pairwise learning problem, we establish the first-ever-known stability bounds of stochastic gradient descent (SGD) for pairwise learning with non-smooth loss functions. Our results hold true for both bounded and unbounded parameter domains. The proof techniques are mainly motivated by the recent work where stability of SGD was established in the pointwise case. The main challenge here is that pairs of examples involved in pairwise learning are not statistically independent. To overcome this hurdle, we develop a novel approach for decoupling such pairwisely dependent random variables in the analysis. We also derive the first generalization bound in high probability for SGD in pairwise learning using the stability approach. We study the differential privacy guarantee and utility bounds of private SGD for pairwise learning by output perturbation method. Our idea is to use our stability results to derive its sensitivity with high probability w.r.t. the randomness of algorithm, and hence guarantee its differential privacy with smaller added noise. The resulting utility bound matches with the output perturbation method in recent works for private SGD in pairwise learning with smooth losses. We provide concrete examples of pairwise learning including AUC maximization and similarity metric learning to illustrate our results.

Then we revisit the pairwise learning problem in both the offline and online settings. We propose simple SGD and OGD algorithms for pairwise learning where the $t$-th update of the model parameter is based on the interacting of the current instance and the previous one which has a constant gradient complexity $O(1)$. We establish the stability results of the proposed SGD algorithms for pairwise learning and apply them to derive optimal excess generalization bounds for the proposed simple SGD algorithm for pairwise learning with both convex and nonconvex as well as both smooth and nonsmooth losses in the offline (finite-sum) setting where the training data of size $n$ is given. We introduce novel techniques to decouple the dependency of the current SGD iterate with the previous instance in both the generalization and optimization error analysis. We further develop a localization version of our SGD algorithms under $(\epsilon, \delta)$-DP constraints, and apply the obtained stability results
to derive an optimal utility (excess generalization) bound.
ACKNOWLEDGMENT

First and foremost, I would like to show my greatest gratitude to my advisor Prof. Yiming Ying for guiding me towards the cutting-edge of machine learning and data science field. His enthusiasm and work ethic have made some profound influence on my Ph.D. study as well as my future career. I am also grateful for his financial support, which provides me enough freedom to focus on my research interests. Overall, this thesis would not be done without his constant support and infinite patience.

I am grateful for all the support from my wonderful collaborators. In particular, I appreciate Dr. Yunwen Lei for teaching me useful optimization and statistics tools, which leads to the joint work [Lei et al., 2021, Yang et al., 2021b]. I also appreciate Dr. Shu Hu for showing me tricks in PyTorch, which leads to the joint work [Yang et al., 2022]. I would like to thank Dr. Baojian Zhou for teaching me coding in C, which leads to the joint work [Yang et al., 2020b]. I am also thankful for having the chance to collaborate with Dr. Puyu Wang, and Dr. Wei Shen.

I am grateful for the faculty and staff in University at Albany and other institutions. In particular, I would like to show my appreciation to Prof. Charles Michelli and Prof. Yuesheng Xu for leading me into this Ph.D. program; Dr. Penghang Yin for broadening my horizons on the job market; Dr. Li Shen and Dr. Xiaoting Zhao for mentoring me during my internship at Tencent and Etsy, respectively. I would also like to thank Joan Mainwaring for her patience and tolerance.

I feel blessed for having a bunch of helpful friends. In particular, I would like to thank Weichen Song, Dr. Wang Xu, Minghao Hu, Jiawen Gan, Dr. Weidong Ma, Dr. Hao Li, Dr. Feng Yu, Shiyun Zhou, Dr. Michael Natole, Jr., Dr. Wenbo Li, Dr. Bo Ye, Fu Li for encouraging me to pursue my career and life. Finally, I want to give my deepest gratitude to my girlfriend Thu Dang Lam for her generous love.
CONTENTS

ABSTRACT ........................................................................................................ iii

ACKNOWLEDGMENT .................................................................................. vi

LIST OF TABLES .......................................................................................... x

LIST OF FIGURES ....................................................................................... xi

1. Introduction ................................................................................................. 1

1.1 Part I: Stability and Differential Privacy of Minimax Problems .............. 4

1.2 Part II: Stability and Differential Privacy of Pairwise Learning Problems . 9

2. SGDA and its Variants for Minimax Problems ........................................ 16

2.1 Problem Set-up ....................................................................................... 16

2.2 Connecting Stability to Generalization ................................................. 18

2.3 SGDA: Convex-Concave Case ............................................................... 28

2.3.1 Stability Bounds .............................................................................. 28

2.3.2 Population Risks ............................................................................. 37

2.4 Nonconvex-Nonconcave Objectives ..................................................... 52

2.4.1 Stability and Generalization of SGDA ......................................... 53

2.4.2 Stability and Generalization of AGDA and Beyond ...................... 64

3. Differential Privacy and Utility for Minimax Problems .......................... 75

3.1 Problem Formulation ............................................................................. 75

3.2 Privacy and Utility Guarantees ............................................................. 76

3.2.1 Convex-Concave Case .................................................................. 79

3.2.2 Nonconvex-Strongly-Concave Case ............................................. 89

4. Stability and Differential Privacy of Non-Smooth Pairwise Learning Problems .............................................................................................................. 106
Nomenclature ................................................................. 187

Acronyms ................................................................. 190
# LIST OF TABLES

**A.1** Summary of Results in Chapter 2

- Page 171

**A.2** Summary of Results in Chapter 3

- Page 171

**B.1** Statistical information of each dataset for AUC optimization

- Page 174

**B.2** Training settings for each model and each dataset

- Page 175

**B.3** Comparison of AUC performance in NSEG and DP-SGDA (Linear and MLP settings) on three datasets with different $\epsilon$ and $\delta=1e-6$. The “Original” means no noise ($\epsilon = \infty$) is added in the algorithms

- Page 177

**B.4** Comparison of AUC performance in NSEG and DP-SGDA (Linear and MLP settings) on three datasets with different $\epsilon$ and different $\delta$. The “Original” means no noise ($\epsilon = \infty$) is added in the algorithms

- Page 180

**B.5** Data Statistics. $n$ is the number of samples and $d$ is the number of features

- Page 181

**B.6** Average AUC score $\pm$ standard deviation across multiple datasets. Our best results are highlighted in bold

- Page 182

**B.7** Average AUC $\pm$ standard deviation on diabetes. Non-Private result is $.813 \pm .016$

- Page 185

**B.8** Average AUC $\pm$ standard deviation on german. Non-Private result is $.763 \pm .016$

- Page 185

**B.9** Average AUC score and average CPU running time $\pm$ standard deviation

- Page 186
LIST OF FIGURES

B.1 △t versus the number of passes on svmguide3 (left) and w5a (right) .......................... 173

B.2 The parameter distance versus the number of passes. Left: generator, right: discriminator. ‘total’ is the mean normalized Euclidean distance across all layers and the shaded area is the standard deviation. ...................................................................................... 174

B.3 (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. ............................................................... 177

B.4 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. ......................................................................................................................... 181

B.5 CPU running time (log scale) versus the AUC score for the hinge loss .......................... 183

B.6 More CPU running time against AUC score for the hinge loss ........................................ 183

B.7 AUC score against CPU running time for the square loss ................................................ 184

B.8 CPU running time of different sampling schemes against the sample size n ..................... 185

B.9 Convergence of Algorithm 4 for the generalized linear model ....................................... 186
CHAPTER 1

Introduction

Numerical optimization algorithm is one of the foundations of machine learning, touching almost every aspect of the discipline. In particular, SGM plays a dominant role in solving large-scale machine learning problems \cite{Bottou2018}. As an iterative algorithm, SGM benefits from cheap per-iteration cost as well as utility guarantees. The prototypical stochastic optimization algorithm is SGD by \cite{Robbins1951} back in the fifties. The formal discussion of SGD and other related algorithms are given as follows in the setting of machine learning.

Let $S = \{z_1, \cdots, z_n\}$ be a set of training examples independently and identically drawn from a probability distribution $D$ defined over the sample space $Z = X \times Y$ where $X \in \mathbb{R}^d$ is an input space and $Y \in \mathbb{R}$ is an output space. The aim is to find a predictive model $w$ from a model parameter space $W$ based on the training dataset $S$. Such process is often fulfilled by solving a minimization problem. In this case, the performance of a prescribed model on a single example $z$ is measured by a nonnegative loss function $f : W \times Z \rightarrow \mathbb{R}$. The corresponding empirical risk and population risk are respectively given by

$$F_S(w) = \frac{1}{n} \sum_{i=1}^{n} f(w; z_i) \quad \text{and} \quad F(w) = \mathbb{E}_{z \sim D}[f(w; z)].$$

Here we use $\mathbb{E}_{z \sim D}[\cdot]$ to denote the expectation with respect to (w.r.t.) $z$. An (randomized) optimization algorithm $A : Z^n \rightarrow W$ utilizes $S$ to produce an output model $A(S)$. Our interest is to study the total error of $A$ defined as $F(A(S)) - \min_{w \in W} F(w)$. The expected total error can be decomposed as

$$\mathbb{E}_{S,A}[F(A(S)) - \min_{w \in W} F(w)] = \mathbb{E}[F(A(S)) - F_S(A(S))] + \mathbb{E}[F_S(A(S)) - \min_{w \in W} F(w)].$$

The first expected term is called the generalization error due to the discrepancy between the population risk and empirical risk of the output model. The second expected term is called the optimization error induced by running an optimization algorithm to minimize the
empirical risk. Given the above set-up, SGD updates the model sequentially upon receiving a training example at iteration \( t \), specifically, chosen \( \mathbf{w}_1 \in \mathcal{W} \), one has

\[
\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; \mathbf{z}_i)).
\]

Here \( \Pi_{\mathcal{W}} \) denotes the projection on \( \mathcal{W} \), \( \{\eta_t\}_t \) denotes a sequence of positive step sizes and \( \nabla f(\mathbf{w}_t; \mathbf{z}_i) \) denotes the gradient w.r.t. \( \mathbf{w}_t \) given the randomly chosen example \( \mathbf{z}_i \). There is a large amount of literature on studying the optimization error of SGD and its variants \cite{Karimi2016,Lacoste-Julien2012,Nemirovski2009,Shamir2013}. In contrast, there is far less work on studying the generalization error of SGD.

Stability is a classical approach to get bounds on the variance of the error of the k-Nearest Neighbors algorithm (k-NN) which dates back to the seventies \cite{Devroye1979,Rogers1978}. The idea originates from sensitivity analysis, which aims at determining how much the variation of the input can influence the output of a system. Then uniform stability was first introduced in the seminal work \cite{Bousquet2002} to derive tighter generalization error bounds for general learning algorithms. Uniform stability bounds the worst case change in loss of the model output by the algorithm on the worst case point when a single data point in the dataset is replaced. This stability was extended to study randomized algorithms \cite{Elisseeff2005}. It was shown that stability is closely related to the fundamental problem of learnability \cite{Rakhlin2005,Shalev-Shwartz2010}. In an influential work, \cite{Hardt2016} gave the first bounds on the generalization error of SGD by extending the uniform stability to randomized algorithm. Formally, the definition is given as follow.

**Definition 1.1** A randomized algorithm \( \mathcal{A} \) is \( \varepsilon \)-uniformly stable if for all training sets \( S, S' \in \mathcal{Z}^n \) such that \( S \) and \( S' \) differ in at most one example, one has

\[
\sup_{\mathbf{z} \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(S); \mathbf{z}) - f(\mathcal{A}(S'); \mathbf{z})] \leq \varepsilon.
\]

Following \cite{Hardt2016}, several upcoming works have used this approach to derive new generalization properties of SGD with different notions of stability measure under different mathematical assumptions, e.g., uniform stability \cite{Chen2018,Lin2016}. 

2
Madden et al. [2020], Mou et al. [2018], Richards et al. [2020], argument stability Bassily et al. [2020], Lei and Ying [2020], Liu et al. [2017], on-average stability Kuzborskij and Lampert [2018], Lei and Ying [2021a], hypothesis stability Charles and Papailiopoulos [2018], Foster et al. [2019], London [2017], Bayes stability Li et al. [2020], and locally elastic stability Deng et al. [2021].

On the other important front, differential privacy (DP) also stems from the sensitivity analysis. DP upper bounds the worst case change in the output distribution of an algorithm when a single data point in the dataset is replaced [Dwork et al., 2006].

**Definition 1.2** An (randomized) algorithm $A$ is called $(\epsilon, \delta)$-differentially private if, for all neighboring datasets $S, S'$ and for all events $O$ in the output space of $A$, the following holds:

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

As a definition of privacy tailored to the problem of privacy-preserving data analysis, DP promises to protect individuals from any addition harm by participating in the database (See Dwork et al. [2014] for a detailed discussion on the promise of DP). Its related technologies have been adopted by Google [Erlingsson et al., 2014], Apple [Ding et al., 2017], and the US Census Bureau [Abowd, 2016].

One of the challenges on DP research is to address the model’s utility for a fixed bound on privacy budget. In particular, the task of risk minimization under DP constraint, and the trade-off between the empirical/population risk versus privacy have been actively studied over a decade [Bassily et al. [2014, 2019], Chaudhuri and Monteleoni [2008], Chaudhuri et al. [2011]]. To design differentially private stochastic gradient method, Song et al. [2013] firstly proposed to perturb SGD at each iteration by Laplace noise and achieve $\epsilon$-DP; Bassily et al. [2014] proposed to use Gaussian noise at SGD update and achieve $(\epsilon, \delta)$-DP; Abadi et al. [2016] introduced the moment accountant technique and improved the privacy guarantee over Bassily et al. [2014]; Wu et al. [2017] perturbed the output of SGD by gaussian noise; Feldman et al. [2018] introduced the privacy amplification by iteration technique tailored for SGD with Gaussian noise. It is worth mentioning that the connection between uniform stability and differential privacy has been exploited in the design of several DP stochastic gradient methods as well as in the utility analysis [Bassily et al., 2019, 2020, Feldman et al., 2018].
Discussion hitherto is built upon the classical framework of empirical risk and population risk minimization. Despite its vast literature, there are notable machine learning problems that drift apart from this framework and classical results above may not apply directly. Therefore, a natural question is raised as follows

*How can one design stable and differentially private stochastic gradient methods beyond risk minimization framework?*

In the following chapters, we focus on two types of optimization problems that enjoy wide applications in modern machine learning but are less studied, namely the minimax problem and the pairwise learning problem. Novel analysis will be proposed to study the utility and privacy guarantee of SGMs therein.

## 1.1 Part I: Stability and Differential Privacy of Minimax Problems

In machine learning we often encounter minimax optimization problems, where the decision variables are partitioned into two groups: one for minimization and one for maximization. Given an objective function $f$, the population minimax problem can be formulated as follows

$$
\min_{w \in W} \max_{v \in V} \left\{ F(w, v) := \mathbb{E}_{z \sim D}[f(w, v; z)] \right\},
$$

(1.1)

where $W \subseteq \mathbb{R}^{d_1}$ and $V \subseteq \mathbb{R}^{d_2}$ are two nonempty closed and convex domains and $z$ is a random variable from some distribution $D$ taking values in $\mathcal{Z}$. Since the distribution $D$ is usually unknown and one has access only to an i.i.d. training dataset $S = \{z_1, \cdots, z_n\}$, one resorts to solving its empirical minimax problem

$$
\min_{w \in W} \max_{v \in V} \left\{ F_S(w, v) := \frac{1}{n} \sum_{i=1}^{n} f(w, v; z_i) \right\}.
$$

This framework covers many important problems as specific instantiations. Notable examples include generative adversarial networks (GANs) [Arjovsky et al., 2017, Goodfellow et al., 2014], AUC maximization [Gao et al., 2013, Liu et al., 2019, Ying et al., 2016].
Zhao et al. [2011], robust learning [Audibert and Catoni [2011], Xu et al. [2009], adversarial training [Sinha et al. [2017], algorithmic fairness [Diana et al. [2021], Li et al. [2019], Martinez et al. [2020], Mohri et al. [2019], Wang et al. [2020b], and Markov Decision Process (MDP) [Puterman [2014], Wang [2017]]. 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. Maximizing AUC with least square loss can be formulated as

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

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

$$
\min_{\theta, a, b} \max_v F(\theta, a, b, c) = \mathbb{E}_z[f(\theta, a, b, v; z)], \tag{1.2}
$$

where $f(\theta, a, b, v; z) = (1 - p)(h(\theta; x) - a)^2[|y = 1] + p(h(\theta; x) - b)^2[|y = -1] + 2(1 + v)(ph(\theta; x)[|y = -1] - (1 - p)h(\theta; x)[|y = 1]) - p(1 - p)v^2$ and $p = \mathbb{P}[y = 1]$. DP has been considered in the setting of AUC maximization [Wang et al. [2021]. The proposed privacy mechanisms there are objective perturbation and output perturbation [Chaudhuri et al. [2011].

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

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

An heuristic differentially private version of RMSProp were employed to train GANs by Xie
Recently differential privacy has successfully applied to private synthetic data generation in the framework of GAN \cite{Beaulieu, Jordon}. \vspace{2pt}

**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 $r(a) \in [0,1]^n$ is the vector of expected state-transition rewards. In the infinite-horizon average-reward Markov decision problem, one aims to find a stationary policy $\pi$ to make an infinite sequence of actions and optimize the average-per-time-step reward $\bar{v}$. By the classical theory of dynamics programming \cite{Puterman}, finding an optimal policy is equivalent as solving the fixed-point Bellman equation

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

where $h \in \mathbb{R}^n$ is the difference-of-value vector. \cite{Wang} showed that this problem is equivalent to the following minimax problem

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

where $\mathcal{H}$ and $\mathcal{U}$ are the feasible regions chosen according to the mixing time and stationary distribution. We refer to \cite{Zhang} 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:

$$
\min_{w \in \mathcal{W}} L(w) = \max_{1 \leq i \leq m} \left\{ \mathbb{E}_z[\ell_1(w;z)], \ldots, \mathbb{E}_z[\ell_m(w;z)] \right\}.
$$

This problem can be reformulated as the following stochastic minimax problem

$$
\min \max \sum_{i=1}^m \mathbb{E}_z[v_i \ell_i(w;z)],
$$

where $\Delta_m = \{ v \in \mathbb{R}^m : v_i \geq 0, \sum_{i=1}^m v_i = 1 \}$ denotes the $m$-dimensional simplex. Such robust optimization formulation has been recently proposed to address fairness among subgroups \cite{Mohri} and federated learning on heterogeneous populations \cite{Li}.
One popular optimization algorithm for solving this problem is stochastic gradient descent ascent (SGDA). Specifically, at iteration $t$, upon receiving a random data point or mini-batch from $S$, it performs gradient descent over $w$ with the stepsize $\eta_{w,t}$ and gradient ascent over $v$ with the stepsize $\eta_{v,t}$.

It is a classical result that SGDA can achieve a convergence rate $O(1/\sqrt{T})$ in the convex and concave case [Nedić and Ozdaglar, 2009, Nemirovski et al., 2009], which can be further improved for SC-SC problems [Balamurugan and Bach, 2016, Hsieh et al., 2019]. For the nonconvex-(strongly)-concave case, the work of [Lin et al., 2020] shows the local convergence of SGDA if the stepsizes $\eta_{w,t}$ and $\eta_{v,t}$ 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], stochastic recursive SGDA [Luo et al., 2020] for nonconvex-strongly-concave problems, and alternating gradient descent ascent [Yang et al., 2020a] for two-sided PL condition, to mention but a few. However, there is relatively little work on studying the generalization, i.e., how the model trained based on the training examples would generalize to test examples. Indeed, a model with good performance on training data may not generalize well if the models are too complex. It is imperative to study the generalization error of the trained models to foresee their prediction behavior. This often entails the investigation of the tradeoff between optimization and estimation for an implicit regularization.

To our best knowledge, there is only two recent work on the generalization analysis for minimax optimization algorithms [Farnia and Ozdaglar, 2021, Zhang et al., 2021]. The argument stability for the specific empirical saddle point (ESP) was studied [Zhang et al., 2021], which implies weak generalization and strong generalization bounds. However, the discussion there ignored optimization errors and nonconvex-nonconcave cases, which can be restrictive in practice. For SC-SC, convex-concave, nonconvex-nonconcave objective functions, the uniform stability of several gradient-based minimax learners was developed in a smooth setting [Farnia and Ozdaglar, 2021], including gradient descent ascent (GDA), proximal point method (PPM) and GDmax. While they developed optimal generalization bounds for PPM, their discussions did not yield vanishing risk bounds for GDA in the gen-
eral convex-concave case since their generalization bounds grow exponentially in terms of the iteration number. Furthermore, the above mentioned papers only study generalization bounds in expectation, and there is a lack of high-probability analysis.

To this end, we leverage the lens of algorithmic stability to study the generalization behavior of minimax learners for both convex-concave and nonconvex-nonconcave problems. Our discussion shows how the optimization and generalization should be balanced for good prediction performance. Our main results are based on Lei et al. [2021]. In particular, our contributions can be summarized as follows.

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

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

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

Many studies analyze the privacy and utility of DP-SGD for the ERM problem that only involves the minimization over $w$ as we introduced earlier. In contrast, there is little work on analysing the utility of minimax optimization algorithms with DP constraints. 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.

Studying the computational and statistical behavior of DP-SGDA is fundamental towards understanding stochastic optimization algorithms for minimax problem under the differential privacy constraint. To this end, 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. Our main results are based on Yang et al. [2022]. 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. Specifically, we show that it can guarantee $(\epsilon, \delta)$-DP and achieve the optimal rate $O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d \log(1/\delta)}}{n}\epsilon\right)$ for smooth and nonsmooth cases. 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. In particular, under the Polyak-Łojasiewicz (PL) condition of $F_S$, we prove that the excess primal population risk enjoys a non-trivial rate $O\left(\frac{1}{n^{1/3}} + \frac{\sqrt{d \log(1/\delta)}}{n^{6/5} \epsilon}\right)$ while guaranteeing $(\epsilon, \delta)$-DP. The key techniques involve the convergence analysis of and the stability analysis for the primal variable 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.

### 1.2 Part II: Stability and Differential Privacy of Pairwise Learning Problems

Many important learning tasks involve pairwise loss functions which are often referred to as pairwise learning. As opposed to the learning tasks associated with pointwise loss functions in standard classification and regression, pairwise learning considers the quality of a model parameter $w$ by measuring it through a pairwise loss on a pair of examples.
Therefore, given a pairwise loss function $\ell : \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, we aim to minimize the following population risk

$$F(w) = \mathbb{E}_{z,z'}[\ell(w,z,z')] ,$$

where $z$ and $z'$ are drawn independently from the population distribution $\mathcal{D}$ on $\mathcal{Z}$. The population distribution is often unknown and we only have access to a set of i.i.d. training data $S = \{z_1,z_2,\ldots,z_n\} \in \mathcal{Z}^n$. The task then reduces to the following minimization problem of the empirical risk

$$\min_{w \in \mathcal{W}} F_S(w) := \frac{1}{n(n-1)} \sum_{i,j=1,i\neq j}^{n} \ell(w,z_i,z_j). \quad (1.3)$$


**AUC maximization.** Area under the ROC curve (AUC) of a prediction function $h_w$ is the probability that the function ranks a random positive example higher than a random negative example. The empirical risk of AUC maximization is given by

$$F_S(w) = \frac{1}{n(n-1)} \sum_{i,j \in [n], i \neq j} \ell(h_w(x_i) - h_w(x_j))\mathbb{I}_{[y_i=1]}\mathbb{I}_{[y_j=-1]} ,$$

where the loss $\ell(\cdot)$ can be the least square loss $\ell(t) = (1-t)^2$ or the hinge loss $\ell(t) = (1-t)^+$. 

**Minimum error entropy principle.** Minimum error entropy (MEE) is a principle of information theoretical learning [Hu et al. 2013, 2015], which aims to find a predictor $h : \mathcal{X} \mapsto \mathcal{Y}$ by minimizing the information entropy of the variable $E = Y - h(X)$. The Rényi’s entropy of order 2 for $E$ is defined as $H(E) = -\log \int p_E^2(e)de$. Here $p_E$ is probability density function and can be approximated by Parzen windowing $\hat{p}_E(e) = \frac{1}{n\gamma} \sum_{i=1}^{n} G\left(\frac{|e-e_i|}{2\gamma}\right)$, where $e_i = y_i - h(x_i)$, and $\gamma > 0$ is an MEE scaling parameter and $G : \mathbb{R} \mapsto \mathbb{R}^+$ is a
windowing function. The approximation of Rényi’s entropy is given by its empirical version
\[ \hat{H} = -\log \frac{1}{n^2} \sum_{i,j \in [n]} G\left( \frac{(e_i - e_j)^2}{2\gamma^2} \right). \]
The maximization of \( \hat{H} \) then leads to a pairwise learning problem since each loss function involves a pair of training examples [Hu et al., 2013].

**Metric learning.** The target of metric learning is to find a distance function \( h_w : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+ \) parameterized by \( w \) such that consistent with some supervised information, e.g., examples within the same class are close while examples from different classes are apart from each other under the learnt metric. If \( \mathcal{Y} = \{\pm 1\} \), the performance of \( h \) on an example pair \( z, z' \) can be quantified by a loss function of the form \( f(w; z, z') = \ell(yy' (1 - h_w(x, x'))) \), where \( \ell: \mathbb{R} \mapsto \mathbb{R}_+ \) is a decreasing function for which some typical choices include the hinge loss \( \ell(t) = (1 - t)_+ \) and the exponential function \( \ell(t) = \log(1 + \exp(-t)) \). Then one can minimize the training error \( F_S(w) = \frac{1}{n(n-1)} \sum_{i,j \in [n]: i \neq j} f(w; z_i, z_j) \) to learn a distance metric. Another closely related learning task to metric learning is contrastive learning [Chen et al., 2020, Oord et al., 2018, Sohn, 2016, Wu et al., 2018] which has become very popular recently for learning visual representations without supervision.

The ERM formulation for pairwise learning has been well studied theoretically using algorithmic stability for ranking problems [Agarwal and Niyogi, 2009], and for regularized metric learning with a strongly convex objective function [Jin et al., 2009]. U-statistics (e.g. De la Pena and Giné [2012]) tools for ranking [Clémençon et al., 2008, Rejchel, 2012] and metric learning [Cao et al., 2016, Verma and Branson, 2015] have also been applied to study the generalization error. At the same time, there are considerable interests on developing and studying online gradient descent (OGD) or stochastic gradient descent (SGD) algorithms for pairwise learning due to their scalability in practice [Kar et al., 2013, Lin et al., 2017, Wang et al., 2012, Ying et al., 2016].

In particular, a sequence of work [Kar et al., 2013, Wang et al., 2012, 2013, Zhao et al., 2011] assumed the online learning setting in which a stream of i.i.d. data \( \{z_1, z_2, \ldots, z_t, \ldots\} \) is continuously arriving. Upon receiving \( z_t \) at time \( t \), it is paired with all previous instances and then the model parameter is updated based on the local error \( F_t(w_{t-1}) = \frac{1}{t-1} \sum_{j=1}^{t-1} f(w_{t-1}; z_t, z_j) \). In particular, the work by [Wang et al., 2012, 2013] provided the first excess generalization bound for online learning methods by obtaining online-to-batch conversion bounds [Cesa-Bianchi et al., 2004] using covering numbers of function classes. [Kar et al., 2013] significantly improved the results using the so-called symmetrization of expec-
tations which reduce excess risk estimates to Rademacher complexities. To further reduce the expensive gradient complexity $O(t)$ at a large time $t$, the work [Kar et al., 2013; Wang et al., 2013; Zhao et al., 2011] proposed to use a buffering set $B_{t-1}$ with size $s$ instead of all previous instances. Generalization bounds of such OGD algorithms have been studied for convex and Lipschitz-continuous losses. The convergence (optimization error) of SGD type algorithms for pairwise learning was obtained in [Lin et al., 2017; Ying et al., 2016] where the algorithms there directly minimize the population risk. In this setting, there is no need to consider generalization (estimation error) i.e. the difference between the empirical risk and the true population risk. The work of [Shen et al., 2020] studied the stability and generalization of SGD in pairwise learning and derived lower bounds for their optimization error over a class of pairwise losses. This work used the uniform stability [Agarwal et al., 2010] in the pairwise learning setting, and this work was largely motivated by [Hardt et al., 2016] for studying SGD in the pointwise case. However, there are some fundamental limitations in the work by [Shen et al., 2020]: it requires the pairwise loss to be both Lipschitz continuous and strongly smooth, and the parameter domain $\mathcal{W}$ is assumed to be bounded. Such assumptions are very restrictive which are violated in many cases such as the least square loss for AUC maximization $(1 - w^T(x - x'))^2 \mathbb{I}_{y = 1 \land y' = -1}$ with $w \in \mathbb{R}^d$ ($\mathbb{I}$ is the indicator function) and the hinge loss for metric learning $(1 + \tau(y, y')(x - x')^T w(x - x'))_+$ where $w$ is a positive semi-definite matrix, and $\tau(y, y') = 1$ if $x, x'$ are from the same class and $-1$ otherwise.

While private SGD has been extensively studied in pointwise learning, there is little work on differentially private SGD for pairwise learning until recently. In particular, [Huale et al., 2020] considered both online and offline learning settings by following the update rule in [Kar et al., 2013; Wang et al., 2012; Zhao et al., 2011]. They derive the rate $\tilde{O}(\sqrt{d}/(\sqrt{n}\epsilon))$ for gradient perturbation and $\tilde{O}(\sqrt{d}/(\sqrt{n}\epsilon))$ for output perturbation. Note that the loss function there needs to be both Lipschitz continuous and strongly smooth.

To this end, we study the stability, generalization, and differential privacy of SGD for pairwise learning with non-smooth losses. Our main results are based on [Yang et al., 2021a]. Our contributions can be summarized as follows.

- We establish the first-ever-known stability bounds of SGD for pairwise learning with non-smooth loss functions. Our results hold true for both bounded and unbounded parameter domains. The proof techniques are mainly motivated by the recent work [Bassily et al.,...
where stability of SGD was established in the pointwise case. The main challenge here is that pairs of examples involved in pairwise learning are not statistically independent. To overcome this hurdle, we develop a novel approach for decoupling such pairwisely dependent random variables in the analysis. We also derive the first generalization bound in high probability for SGD in pairwise learning using the stability approach.

- We study the differential privacy guarantee and utility bounds of private SGD for pairwise learning by output perturbation method. Our idea is to use our stability results to derive its sensitivity with high probability w.r.t. the randomness of algorithm, and hence guarantee its differential privacy with smaller added noise. The resulting utility bound matches with the output perturbation method in [Huai et al., 2020] for private SGD in pairwise learning with smooth losses.

- We provide concrete examples of pairwise learning including AUC maximization and similarity metric learning to illustrate our stability and differential privacy results. In particular, we give an affirmative solution to the open question raised in [Cao et al., 2016] that whether similarity metric learning with nuclear-norm constraint can yield milder dependence on the dimensionality than the Frobenius-norm constraint.

The critical issue for designing such stochastic algorithms is to construct intersecting pairs of instances for updating the model parameter upon receiving individual instances. For the offline (finite-sum) setting where the prescribed training data $S = \{z_1, \ldots, z_n\}$ is available, one natural approach is to, at each time $t$, randomly select a pair of instances $(z_{i_t}, z_{j_t})$ from all $\binom{n}{2}$ pairs and update $w$ based on the gradient of the local error $f(w_{t-1}; z_{i_t}, z_{j_t})$. An excess risk bound $\tilde{O}(1/\sqrt{n})$ with high probability was derived for convex, Lipschitz and strongly smooth losses [Lei et al., 2020]. One popular approach [Kar et al., 2013; Wang et al., 2013; Ying and Zhou, 2006; Zhao et al., 2011] is to consider the online setting where the data is continuously arriving. This approach pairs the current datum $z_t = (x_t, y_t)$, which is received at time $t$, with all previous instances $S_{t-1} = \{z_1, \ldots, z_{t-1}\}$ and then performs the update based on the gradient of the local error $f(w_{t-1}; S_{t-1}) = \frac{1}{t-1} \sum_{z \in S_{t-1}} f(w_{t-1}; z_t, z)$. It requires a high gradient complexity $O(t)$ (i.e. the number of computing gradients) which is expensive when $t$ becomes large. To mitigate this potential limitation, [Kar et al., 2013],
Wang et al. [2012], Zhao et al. [2011] proposed to use a buffering set $B_{t-1} \subseteq S_{t-1}$ of size $s$ and the local error $f(w_{t-1}; B_{t-1}) = \frac{1}{s} \sum_{z \in B_{t-1}} f(w_{t-1}; z, z)$ which reduces the gradient complexity to $O(s)$. The excess generalization bound $O(\frac{1}{\sqrt{s}} + \frac{1}{\sqrt{n}})$ was established in [Kar et al., 2013] using the online-to-batch conversion method [Cesa-Bianchi et al., 2004] which is only meaningful for a very large $s$. In particular, this bound tends to zero only when $s = s(n)$ tends to infinity as $n$ tends to infinity. It was mentioned in [Kar et al., 2013] (see the discussion at the end of Section 7 there) as an open question on how to get a meaningful bound for a fixed constant $s$. It is worth mentioning that in the particular case of AUC maximization with the least square loss, Ying et al. [2016] considered the online learning setting and reformulated the problem as a stochastic saddle point (min-max) problem which decouples the pairwise structure. From this reformulation, efficient SGD-type algorithms [Liu et al., 2019, Ying and Li, 2012] have been developed.

We will show that optimal generalization bounds can be achieved for simple SGD and OGD algorithms for pairwise learning where, at time $t$, the current instance $z_t$ is only paired with the previous instance $z_{t-1}$. This is equivalent to the First-In-First-Out (FIFO) buffering strategy [Kar et al., 2013, Wang et al., 2012] while keeping the size $s$ of the buffering set $B_{t-1}$ to be $s = 1$, where, in this FIFO policy, the data $z_t$ arriving at time $t > 1$ is included into the buffer by removing $\{z_1, \ldots, z_{t-2}\}$ from the buffer. Our results are based on [Yang et al., 2021b]. In particular, our main contributions are summarized as follows.

- We propose simple SGD and OGD algorithms for pairwise learning where the $t$-th update of the model parameter is based on the interacting of the current instance and the previous one which has a constant gradient complexity $O(1)$.

- We establish the stability results of the proposed SGD algorithms for pairwise learning and apply them to derive optimal excess generalization bounds $O(1/\sqrt{n})$ for the proposed simple SGD algorithm for pairwise learning with both convex and nonconvex as well as both smooth and nonsmooth losses in the offline (finite-sum) setting where the training data of size $n$ is given. We introduce novel techniques to decouple the dependency of the current SGD iterate with the previous instance in both the generalization and optimization error analysis, which resolves the open question in [Kar et al., 2013] on how to develop meaningful generalization bounds when the buffering set of FIFO has a very small size.
We further develop a localization version of our SGD algorithms under \((\epsilon, \delta)\)-differential privacy (DP) constraints, and apply the obtained stability results to derive an optimal utility (excess generalization) bound \(\tilde{O}(1/\sqrt{n} + \sqrt{d/n}\epsilon)\). In contrast to the existing work [Huai et al., 2020] which requires the loss function to be smooth and an at least quadratic gradient complexity, our proposed DP algorithms only need a linear gradient complexity \(\tilde{O}(n)\) for smooth convex losses to achieve the optimal utility bound and can also be applied to non-smooth convex losses.
CHAPTER 2
SGDA and its Variants for Minimax Problems

2.1 Problem Set-up

Our main results in this chapter are based on [Lei et al. 2021]. Unlike the standard statistical learning theory (SLT) setting where there is only a minimization of \( w \), we have different measures of population risk due to the minimax structure [Zhang et al., 2021].

Definition 2.1 (Weak Primal-Dual Risk) The weak Primal-Dual (PD) population risk of a (randomized) model \((w, v)\) is defined as [Zhang et al., 2021]

\[
\Delta^w(w, v) = \sup_{v' \in V} \mathbb{E}[F(w, v') - \inf_{w' \in W} \mathbb{E}[F(w', v)]].
\]

The weak PD empirical risk of \((w, v)\) is defined as

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

We refer to \(\Delta^w(w, v) - \Delta^w_S(w, v)\) as the weak PD generalization error of the model \((w, v)\).

Definition 2.2 (Strong Primal-Dual Risk) The strong PD population risk of a model \((w, v)\) is defined as

\[
\Delta^s(w, v) = \sup_{v' \in V} F(w, v') - \inf_{w' \in W} F(w', v).
\]

The strong PD empirical risk of \((w, v)\) is defined as

\[
\Delta^s_S(w, v) = \sup_{v' \in V} F_S(w, v') - \inf_{w' \in W} F_S(w', v).
\]

We refer to \(\Delta^s(w, v) - \Delta^s_S(w, v)\) as the strong PD generalization error of the model \((w, v)\).

Definition 2.3 (Primal Risk) The primal population risk of a model \(w\) is defined as

\[
R(w) = \sup_{v \in V} F(w, v).
\]
The primal empirical risk of \( w \) is defined as

\[
R_S(w) = \sup_{v \in V} F_S(w, v).
\]

We refer to \( R(w) - R_S(w) \) as the primal generalization error of the model \( w \), and \( R(w) - \inf_{w'} R(w') \) as the excess primal population risk.

According to the above definitions\(^1\) we know \( \Delta^w(w, v) \leq \mathbb{E}[\Delta^s(w, v)] \) by applying Jensen’s inequality and \( R(w) - R_S(w) \) is closely related to \( \Delta^s(w, v) - \Delta^s_S(w, v) \). However, when \( F \) is strongly-convex-strongly-concave, the point distance from the model \((\mathcal{A}_w(S), \mathcal{A}_v(S))\) to the true saddle point \((w^*, v^*) \in \arg\min_{w \in W} \max_{v \in V} F(w, v)\) can be bounded by the weak PD population risk, i.e. \( \mathbb{E}[\|\mathcal{A}_w(S) - w^*\|^2_2 + \|\mathcal{A}_v(S) - v^*\|^2_2] \leq O(\Delta^w(\mathcal{A}_w(S), \mathcal{A}_v(S))) \). The key difference between the weak PD risk and the strong PD risk is that the expectation is inside of the supremum/infimum for weak PD risk, while outside of the supremum/infimum for strong PD risk. In this way, one does not need to consider the coupling between primal and dual models for studying weak PD risks, and has to consider this coupling for strong PD risks. For certain problems, it is suffices to bound the weak PD risk, such as the learning problem for Markov decision process \[\text{[Zhang et al., 2021]}\]. 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.

Furthermore, we refer to \( F(w, v) - F_S(w, v) \) as the plain generalization error as it is standard in SLT. An approach to handle a population risk is to decompose it into a generalization error (estimation error) and an empirical risk (optimization error) \[\text{[Bousquet and Bottou, 2008]}\]. For example, the weak PD population risk can be decomposed as

\[
\Delta^w(w, v) = (\Delta^w(w, v) - \Delta^s_S(w, v)) + \Delta^s_S(w, v). \quad (2.1)
\]

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

\(^1\)Primal generalization bounds were presented in \[\text{[Farnia and Ozdaglar, 2021]}\]. However, the stability analysis there actually only implies bounds on weak PD risk.
We now introduce necessary definitions and assumptions. A function $g : \mathcal{W} \mapsto \mathbb{R}$ is said to be $\rho$-strongly-convex ($\rho \geq 0$) iff for all $w, w' \in \mathcal{W}$ there holds $g(w) \geq g(w') + \langle w - w', \nabla g(w') \rangle + \frac{\rho}{2} \| w - w' \|^2$, where $\nabla$ is the gradient operator. We say $g$ is convex if $g$ is 0-strongly-convex. We say $g$ is $\rho$-strongly concave if $-g$ is $\rho$-strongly convex and concave if $-g$ is convex.

Definition 2.4 Let $\rho \geq 0$ and $g : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$. We say

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

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

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

The following two assumptions are standard [Farnia and Ozdaglar, 2021] Zhang et al. 2021. Assumption 2.1 amounts to saying $f$ is Lipschitz continuous with respect to (w.r.t.) both $w$ and $v$. Let $\nabla_w f$ denote the gradient w.r.t. $w$.

Assumption 2.1 The function $f$ is said to be Lipschitz continuous if there exist $G_w, G_v > 0$ such that, for any $w, w' \in \mathcal{W}, v, v' \in \mathcal{V}$ and $z \in \mathcal{Z}$, $\| f(w, v; z) - f(w', v; z) \|_2 \leq G_w \| w - w' \|_2$, and $\| f(w, v; z) - f(w, v'; z) \|_2 \leq G_v \| v - v' \|_2$. And denote $G = \max \{ G_w, G_v \}$.

Assumption 2.2 Let $L > 0$. For any $z$, the function $(w, v) \mapsto f(w, v; z)$ is said to be $L$-smooth, if the following inequality holds for all $w \in \mathcal{W}, v \in \mathcal{V}$ and $z \in \mathcal{Z}$

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

2.2 Connecting Stability to Generalization

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

**Definition 2.5 (Algorithmic Stability)** Let \( \mathcal{A} \) be a randomized algorithm, \( \epsilon > 0 \) and \( \delta \in (0,1) \). Then we say

(a) \( \mathcal{A} \) is \( \epsilon \)-weakly-stable if for all neighboring \( S \) and \( S' \), there holds

\[
\sup_{z} \left( \sup_{\mathbf{v}' \in \mathcal{V}} \mathbb{E}_{\mathcal{A}} \left[ f(\mathcal{A}_w(S), \mathbf{v'}; z) - f(\mathcal{A}_w(S'), \mathbf{v'}; z) \right] 
+ \sup_{\mathbf{w}' \in \mathcal{W}} \mathbb{E}_{\mathcal{A}} \left[ f(\mathbf{w'}, \mathcal{A}_v(S); z) - f(\mathbf{w'}, \mathcal{A}_v(S'; z)) \right] \right) \leq \epsilon.
\]

(b) \( \mathcal{A} \) is \( \epsilon \)-argument-stable in expectation if for all neighboring \( S \) and \( S' \), there holds

\[
\mathbb{E}_{\mathcal{A}} \left[ \left\| \left( \mathcal{A}_w(S) - \mathcal{A}_w(S') \right) - \left( \mathcal{A}_v(S) - \mathcal{A}_v(S') \right) \right\|_2 \right] \leq \epsilon.
\]

\( \mathcal{A} \) is \( \epsilon \)-argument-stable with probability at least \( 1 - \delta \) if with probability at least \( 1 - \delta \)

\[
\left\| \left( \mathcal{A}_w(S) - \mathcal{A}_w(S') \right) - \left( \mathcal{A}_v(S) - \mathcal{A}_v(S') \right) \right\|_2 \leq \epsilon.
\]

(c) \( \mathcal{A} \) is \( \epsilon \)-uniformly-stable with probability at least \( 1 - \delta \) if with probability at least \( 1 - \delta \)

\[
\sup_{z} \left[ f(\mathcal{A}_w(S), \mathcal{A}_v(S); z) - f(\mathcal{A}_w(S'), \mathcal{A}_v(S'; z)) \right] \leq \epsilon.
\]

Under Assumption 2.1, argument stability implies weak and uniform stability. As we will see, argument stability plays an important role in deriving primal population risk bounds.

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

**Theorem 2.1** Let $A$ be a randomized algorithm and $\epsilon > 0$.

(a) If $A$ is $\epsilon$-weakly-stable, then the weak PD generalization error of $(A_w(S), A_v(S))$ satisfies

$$\Delta^w(A_w(S), A_v(S)) - \Delta^w_S(A_w(S), A_v(S)) \leq \epsilon.$$  

(b) If $A$ is $\epsilon$-argument-stable in expectation, the function $v \mapsto F(w, v)$ is $\rho$-strongly-concave and Assumptions 2.1, 2.2 hold, then the primal generalization error satisfies

$$E_{S,A} \left[ R(A_w(S)) - R_S(A_w(S)) \right] \leq (1 + L/\rho)G\epsilon.$$  

(c) If $A$ is $\epsilon$-argument-stable in expectation, $v \mapsto F(w, v)$ is $\rho$-SC-SC and Assumptions 2.1, 2.2 hold, then the strong PD generalization error satisfies

$$E_{S,A} \left[ \Delta^s(A_w(S), A_v(S)) - \Delta^s_S(A_w(S), A_v(S)) \right] \leq (1 + L/\rho)G\sqrt{2}\epsilon.$$  

(d) Assume $|f(w, v; z)| \leq R$ for some $R > 0$ and $w \in W, v \in V, z \in Z$. Assume for all $w$, the function $v \mapsto F(w, v)$ is $\rho$-strongly-concave and Assumptions 2.1, 2.2 hold. Let $\delta \in (0, 1)$. If $A$ is $\epsilon$-uniformly stable almost surely (a.s.), then with probability at least $1 - \delta$

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

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

$$| F(A_w(S), A_v(S)) - F_S(A_w(S), A_v(S)) | = O(\epsilon \log n \log(1/\delta) + Rn^{-\frac{1}{2}} \sqrt{\log(1/\delta)}) .$$

**Proof:** For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$. For any function $g, \tilde{g}$, we have the basic inequalities

$$\sup_w g(w) - \sup_w \tilde{g}(w) \leq \sup_w g(w) - \tilde{g}(w)$$

$$\inf_w g(w) - \inf_w \tilde{g}(w) \leq \sup_w g(w) - \tilde{g}(w).$$ \hspace{1cm} (2.2)

We first prove Part (a). It follows from (2.2) that

$$\triangle^w(A_w(S), A_v(S)) - \triangle^w_S(A_w(S), A_v(S)) \leq \sup_{v' \in V} \mathbb{E}[F(A_w(S), v') - F_S(A_w(S), v')]$$

$$+ \sup_{w' \in W} \mathbb{E}[F_S(w', A_v(S)) - F(w', A_v(S))] .$$

According to the symmetry between $z_i$ and $z'_i$ we know

$$\mathbb{E}[F(A_w(S), v') - F_S(A_w(S), v')] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[F(A_w(S^{(i)}), v') - F_S(A_w(S), v')]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(A_w(S^{(i)}), v'; z_i) - f(A_w(S), v'; z_i)],$$

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

$$\mathbb{E}[F_S(w', A_v(S)) - F(w', A_v(S))] = \frac{1}{n} \sum_{i=1}^{n} [f(w', A_v(S^{(i)}); z_i) - f(w', A_v(S); z_i)].$$
As a combination of the above three inequalities we get

\[
\triangle \! w(A_w(S), A_v(S)) - \triangle \! w(S(A_w(S), A_v(S))) \\
\leq \sup_{v' \in V} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ f(A_w(S^{(i)}), v'; z_i) - f(A_w(S), v'; z_i) \right] \right] \\
+ \sup_{w' \in W} \left[ \frac{1}{n} \sum_{i=1}^{n} \left[ f(w', A_v(S^{(i)}); z_i) - f(w', A_v(S); z_i) \right] \right].
\]

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

The following lemma quantifies the sensitivity of the optimal \(v\) w.r.t. the perturbation of \(w\).

**Lemma 2.1 (Lin et al. 2020)** Let \(\phi : W \times V \mapsto \mathbb{R}\). Assume that for any \(w\), the function \(v \mapsto \phi(w, v)\) is \(\rho\)-strongly-concave. Suppose for any \((w, v), (w', v')\) we have

\[
\| \nabla_v \phi(w, v) - \nabla_v \phi(w', v) \|_2 \leq L \| w - w' \|_2.
\]

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

\[
\| v^*(w) - v^*(w') \|_2 \leq \frac{L}{\rho} \| w - w' \|_2.
\]

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

\[
\mathbb{E} \left[ \sup_{v' \in V} F(A_w(S), v') \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{v' \in V} F(A_w(S^{(i)}), v') \right] \\
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ F(A_w(S^{(i)}), v^*_S) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ f(A_w(S^{(i)}), v^*_{S^{(i)}}; z_i) \right],
\]

where the last identity holds since \(z_i\) is independent of \(A_w(S^{(i)})\) and \(v^*_{S^{(i)}}\).
According to Assumption 2.1, we know
\[
 f(A_w(S(i)), v_s^*(i); z_i) - f(A_w(S), v_s^*(i); z_i)
 = f(A_w(S(i)), v_s^*(i); z_i) - f(A_w(S(i)), v_s^*(i); z_i) + f(A_w(S(i)), v_s^*(i); z_i) - f(A_w(S), v_s^*(i); z_i)
 \leq G \|A_w(S(i)) - A_w(S)\|_2 + G \|v_s^*(i) - v_s^*(i)\|_2 \leq (1 + L/\rho) G \|A_w(S(i)) - A_w(S)\|_2, \quad (2.3)
\]
where in the last inequality we have used Lemma 2.1 due to the strong concavity of \( v \mapsto F(w, v) \) for any \( w \). As a combination of the above two inequalities, we get
\[
\mathbb{E}\left[ \sup_{v' \in V} F(A_w(S), v') \right]
\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[ f(A_w(S), v_s^*(i); z_i) \right] + \frac{(1 + L/\rho) G}{n} \sum_{i=1}^{n} \mathbb{E}\left[ \|A_w(S(i)) - A_w(S)\|_2 \right]
= \mathbb{E}\left[ F_S(A_w(S), v_s^*(i); z_i) \right] + \frac{(1 + L/\rho) G}{n} \sum_{i=1}^{n} \mathbb{E}\left[ \|A_w(S(i)) - A_w(S)\|_2 \right]
\leq \mathbb{E}\left[ \sup_{v' \in V} F_S(A_w(S), v') \right] + \frac{(1 + L/\rho) G}{n} \sum_{i=1}^{n} \mathbb{E}\left[ \|A_w(S(i)) - A_w(S)\|_2 \right]. \quad (2.4)
\]
The stated bound in Part (b) then follows. In a similar way, one can show that
\[
\mathbb{E}\left[ \inf_{w' \in W} F_S(w', A_v(S)) \right] - \mathbb{E}\left[ \inf_{w' \in W} F(w', A_v(S)) \right] \leq \frac{(1 + L/\rho) G}{n} \sum_{i=1}^{n} \mathbb{E}\left[ \|A_v(S(i)) - A_v(S)\|_2 \right].
\]
The above inequality together with (2.4) then implies
\[
\mathbb{E}\left[ \triangle^s(A_w(S), A_v(S)) \right] - \mathbb{E}\left[ \triangle^s(A_w(S), A_v(S)) \right]
= \mathbb{E}\left[ \sup_{v' \in V} F(A_w(S), v') \right] - \mathbb{E}\left[ \sup_{v' \in V} F_S(A_w(S), v') \right] + \mathbb{E}\left[ \inf_{w' \in W} F_S(w', A_v(S)) \right]
- \mathbb{E}\left[ \inf_{w' \in W} F(w', A_v(S)) \right]
\leq (1 + L/\rho) G \mathbb{E}\left[ \|A_w(S(i)) - A_w(S)\|_2 \right] + (1 + L/\rho) G \mathbb{E}\left[ \|A_v(S(i)) - A_v(S)\|_2 \right]
\leq (1 + L/\rho) G \sqrt{2} \mathbb{E}\left[ \left\| \begin{pmatrix} A_w(S(i)) - A_w(S) \\ A_v(S(i)) - A_v(S) \end{pmatrix} \right\|_2 \right],
\]
where we have used the elementary inequality \( a + b \leq \sqrt{2(a^2 + b^2)} \). This proves the stated bound in Part (c).
To prove Part (d) on high-probability bounds, we need to introduce some lemmas.

**Lemma 2.2 (Bousquet et al. 2020)** Let $S = \{z_1, \ldots, z_n\}$ be a set of independent random variables each taking values in $\mathbb{Z}$ and $M > 0$. Let $g_1, \ldots, g_n$ be some functions $g_i : \mathbb{Z}^n \mapsto \mathbb{R}$ such that the following holds for any $i \in [n]$

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

$$|g_i(S) - g_i(z_1, \ldots, z_{j-1}, z''_j, z_{j+1}, \ldots, z_n)| \leq \beta. \quad (2.5)$$

Then, for any $p \geq 2$

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

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

**Lemma 2.3 (Bousquet et al. 2020, Vershynin 2018)** Let $a, b \in \mathbb{R}_+$ and $\delta \in (0, 1/e)$. Let $Z$ be a random variable with $\|Z\|_p \leq \sqrt{pa} + pb$ for any $p \geq 2$. Then with probability at least $1 - \delta$

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

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

$$v_S^* = \arg \max_{v \in \mathcal{V}} F(A_w(S), v). \quad (2.6)$$
We have the following error decomposition

\[ n F(\mathbb{A}_w(S), v^*_S) - n \sup_{v' \in \mathcal{V}} F_S(\mathbb{A}_w(S), v') \]

\[ \leq \sum_{i=1}^{n} \mathbb{E}_{Z}[f(\mathbb{A}_w(S), v^*_S; Z) - \mathbb{E}_{Z'}[f(\mathbb{A}_w(S^{(i)}), v^*_S; Z)]] \]

\[ + \sum_{i=1}^{n} \mathbb{E}_{Z'}[\mathbb{E}_{Z}[f(\mathbb{A}_w(S^{(i)}), v^*_S; Z)] - f(\mathbb{A}_w(S^{(i)}), v^*_S; z_i)] \]

\[ + \sum_{i=1}^{n} \mathbb{E}_{Z'}[f(\mathbb{A}_w(S^{(i)}), v^*_S; z_i)] - n \sup_{v' \in \mathcal{V}} F_S(\mathbb{A}_w(S), v'). \]

By the definition of \( v^*_S(i) \) we know \( \mathbb{E}_Z[f(\mathbb{A}_w(S^{(i)}), v^*_S; Z)] \geq \mathbb{E}_Z[f(\mathbb{A}_w(S^{(i)}), v^*_S; Z)] \). It then follows that

\[ n F(\mathbb{A}_w(S), v^*_S) - n \sup_{v' \in \mathcal{V}} F_S(\mathbb{A}_w(S), v') \]

\[ \leq \sum_{i=1}^{n} \mathbb{E}_{Z}[f(\mathbb{A}_w(S), v^*_S; Z) - \mathbb{E}_{Z'}[f(\mathbb{A}_w(S^{(i)}), v^*_S; Z)]] \]

\[ + \sum_{i=1}^{n} \mathbb{E}_{Z'}[\mathbb{E}_{Z}[f(\mathbb{A}_w(S^{(i)}), v^*_S; Z)] - f(\mathbb{A}_w(S^{(i)}), v^*_S; z_i)] \]

\[ + \sum_{i=1}^{n} \mathbb{E}_{Z'}[f(\mathbb{A}_w(S^{(i)}), v^*_S; z_i)] - n \sup_{v' \in \mathcal{V}} F_S(\mathbb{A}_w(S), v'). \]

According to [2,3], we know

\[ \sum_{i=1}^{n} \mathbb{E}_{Z'}[f(\mathbb{A}_w(S^{(i)}), v^*_S; z_i)] \]

\[ \leq (1 + L/\rho) G \sum_{i=1}^{n} \mathbb{E}_{Z'}[\| A_w(S^{(i)}) - A_w(S) \|_2] + \sum_{i=1}^{n} f(\mathbb{A}_w(S), v^*_S; z_i) \]

\[ = (1 + L/\rho) G \sum_{i=1}^{n} \mathbb{E}_{Z'}[\| A_w(S^{(i)}) - A_w(S) \|_2] + \sum_{i=1}^{n} f(\mathbb{A}_w(S), v^*_S; z_i) \]

\[ \leq (1 + L/\rho) G \sum_{i=1}^{n} \mathbb{E}_{Z'}[\| A_w(S^{(i)}) - A_w(S) \|_2] + n \sup_{v' \in \mathcal{V}} F_S(\mathbb{A}_w(S), v'). \]
As a combination of the above two inequalities, we derive

\[n F(A_w(S), v_S^*) - n \sup_{v' \in V} F_S(A_w(S), v') \leq (2 + L/\rho) n G\epsilon + \sum_{i=1}^{n} g_i(S), \quad (2.7)\]

where we introduce

\[g_i(S) = \mathbb{E}_{z_i} \left[ \mathbb{E}_Z [f(A_w(S^{(i)}), v^*_{S^{(i)}}; Z)] - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z_i) \right] \]

and use the inequality

\[f(A_w(S), v_S^*; Z) - f(A_w(S^{(i)}), v^*_{S^{(i)}}; Z) \leq G \| A_w(S) - A_w(S^{(i)}) \|_2 \leq G\epsilon.\]

Due to the symmetry between \(z_i\) and \(Z\), we know \(\mathbb{E}_{z_i}[g_i(S)] = 0\). Furthermore, the inequality \(|\mathbb{E}_{S \setminus \{z_i}\}[g_i(S)]| \leq 2R\) is also clear. For any \(j \neq i\) and any \(z''_j \in Z\), we know

\[\left| \mathbb{E}_{z_i} \left[ \mathbb{E}_Z [f(A_w(S^{(i)}), v^*_{S^{(i)}}; Z)] - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z_i) \right] \right.\]

\[- \mathbb{E}_{z'_i} \left[ \mathbb{E}_Z [f(A_w(S^{(i)}), v^*_{S^{(i)}}; Z)] - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z_i) \right]\]

\[\leq \left| \mathbb{E}_{z_i} \left[ \mathbb{E}_Z [f(A_w(S^{(i)}), v^*_{S^{(i)}}; Z)] - \mathbb{E}_Z [f(A_w(S^{(i)}), v^*_{S^{(i)}}; Z)] \right] \right|\]

\[+ \mathbb{E}_{z'_i} \left[ f(A_w(S^{(i)}), v^*_{S^{(i)}}; z_i) - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z_i) \right],\]

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

\[\left| f(A_w(S^{(i)}), v^*_{S^{(i)}}; z) - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z) \right|\]

\[\leq \left| f(A_w(S^{(i)}), v^*_{S^{(i)}}; z) - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z) \right| + \left| f(A_w(S^{(i)}), v^*_{S^{(i)}}; z) - f(A_w(S^{(i)}), v^*_{S^{(i)}}; z) \right|\]

\[\leq G \|v^*_{S^{(i)}} - v^*_{S^{(i)}}\|_2 + G \|A_w(S^{(i)}) - A_w(S^{(i)})\|_2 \leq (L/\rho + 1) G\epsilon.\]

Therefore \(g_i(S)\) satisfies the condition (2.5) with \(\beta = (L/\rho + 1) G\epsilon\). Therefore, all the conditions of Lemma 2.2 hold and we can apply Lemma 2.2 to derive the following inequality.
for any $p \geq 2$

$$
\left\| \sum_{i=1}^{n} g_i(S) \right\|_p \leq 12\sqrt{6pn}(L/\rho + 1)G\epsilon[\log_2 n] + 6\sqrt{2R\sqrt{pn}}.
$$

This together with Lemma 2.3 implies the following inequality with probability $1 - \delta$

$$
\left| \sum_{i=1}^{n} g_i(S) \right| \leq e\left( 6R\sqrt{2n\log(e/\delta)} + 12\sqrt{6n}(L/\rho + 1)G\epsilon \log(e/\delta)[\log_2 n] \right).
$$

We can plug the above inequality back into (2.7) and derive the following inequality with probability at least $1 - \delta$

$$
F(\mathcal{A}_w(S), v^*_S) - \sup_{v' \in \mathcal{V}} F_S(\mathcal{A}_w(S), v')
\leq (2 + L/\rho)G\epsilon + e\left( 6R\sqrt{2n\log(e/\delta)} + 12\sqrt{6n}(L/\rho + 1)G\epsilon \log(e/\delta)[\log_2 n] \right).
$$

This proves the stated bound in Part (d). Part (e) is standard in the literature [Bousquet et al., 2020].

**Remark 2.1** We compare Theorem 2.1 with related work. Weak and strong PD generalization error bounds were established for (R)-ESP [Zhang et al., 2021]. However, the discussion there does not consider the connection between stability and generalization. Primal generalization bounds were studied for stable algorithms [Farnia and Ozdaglar, 2021]. However, the discussion there is not rigorous since they used an identity $nR_S(\mathcal{A}_w(S)) = \sum_{i=1}^{n} \max_v f(\mathcal{A}_w(S), v; z_i)$, which does not hold. To our best knowledge, Theorem 2.1 gives the first systematic connection between stability and generalization for minimax problems.

**Remark 2.2** We provide some intuitive understanding of Theorem 2.1 here. Part (a) shows that weak-stability is sufficient for weak PD generalization. This is as expected since both the supremum over $w'$ and $v'$ are outside of the expectation operator in the definition of weak stability/generalization. We do not need to consider the correlation between $\mathcal{A}_w(S)$ and $v'$. As a comparison, the primal generalization needs the much stronger argument-stability. The reason is that the supremum over $w'$ is inside the expectation and $v^{(i)} := \arg\sup_{v'} F(\mathcal{A}_w(S^{(i)}), v')$ is different for different $i$ ($v^{(i)}$ correlates to $\mathcal{A}_w(S^{(i)})$ and $S^{(i)}$ is
derived by replacing the \( i \)-th example in \( S \) with \( z'_i \). We need to estimate how \( v^{(i)} \) differs from each other due to the difference among \( A_w(S^{(i)}) \). This explains why we need argument-stability and a strong-concavity in Parts (b), (d) for primal generalization. Similarly, the strong PD generalization assumes SC-SC functions.

2.3 SGDA: Convex-Concave Case

In this section, we are interested in SGDA for solving minimax optimization problems in the convex-concave case. Let \( w_1 = 0 \in W \) and \( v_1 = 0 \in V \) be the initial point. Let \( \Pi_W(\cdot) \) and \( \Pi_V(\cdot) \) denote the projections onto \( W \) and \( V \), respectively. Let \( \{\eta_t\}_t \) be a sequence of positive stepsizes. At each iteration, we randomly draw \( i_t \) from the uniform distribution over \( [n] := \{1, 2, \ldots, n\} \) and do the update

\[
\begin{align*}
w_{t+1} &= \Pi_W(w_t - \eta_t \nabla_w f(w_t, v_t, z_{i_t})), \\
v_{t+1} &= \Pi_V(v_t + \eta_t \nabla_v f(w_t, v_t, z_{i_t})).
\end{align*}
\]

(2.8)

2.3.1 Stability Bounds

The stability analysis of SGDA replies on the properties of the gradient map

\[
G_{f,\eta} : \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{w} - \eta \nabla_w f(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta \nabla_v f(\mathbf{w}, \mathbf{v}) \end{pmatrix}.
\]

The following lemma shows that \( G_{f,\eta} \) is approximately nonexpansive and is nonexpansive if \( f \) is SC-SC and the step size is small. Part (b) can be found in Farnia and Ozdaglar 2021

**Lemma 2.4** Let \( f \) be \( \rho \)-SC-SC with \( \rho \geq 0 \).

(a) If Assumption 2.1 holds, then

\[
\left\| \begin{pmatrix} \mathbf{w} - \eta \nabla_w f(\mathbf{w}, \mathbf{v}) \\ \mathbf{v} + \eta \nabla_v f(\mathbf{w}, \mathbf{v}) \end{pmatrix} - \begin{pmatrix} \mathbf{w}' - \eta \nabla_w f(\mathbf{w}', \mathbf{v}') \\ \mathbf{v}' + \eta \nabla_v f(\mathbf{w}', \mathbf{v}') \end{pmatrix} \right\|_2^2 \leq (1 - 2\rho \eta) \left\| \begin{pmatrix} \mathbf{w} - \mathbf{w}' \\ \mathbf{v} - \mathbf{v}' \end{pmatrix} \right\|_2^2 + 8\rho^2 \eta^2.
\]
(b) If Assumption 2.2 holds, then
\[ \left\| \left( \begin{array}{c} w - \eta \nabla_w f(w, v) \\
 v + \eta \nabla_v f(w, v) \end{array} \right) - \left( \begin{array}{c} w' - \eta \nabla_w f(w', v') \\
 v' + \eta \nabla_v f(w', v') \end{array} \right) \right\|_2^2 \leq \left( 1 - 2\rho\eta + L^2\eta^2 \right) \left\| \left( \begin{array}{c} w - w' \\
 v - v' \end{array} \right) \right\|_2^2. \]

To prove Lemma 2.4 we require the following standard lemma [Rockafellar, 1976].

**Lemma 2.5** Let \( f \) be a \( \rho \)-SC-SC function, \( \rho \geq 0 \). Then
\[ \left\langle \left( \begin{array}{c} w - w' \\
 v - v' \end{array} \right), \left( \begin{array}{c} \nabla_w f(w, v) - \nabla_w f(w', v') \\
 \nabla_v f(w', v') - \nabla_v f(w, v) \end{array} \right) \right\rangle \geq \rho \left\| \left( \begin{array}{c} w - w' \\
 v - v' \end{array} \right) \right\|_2^2. \] (2.9)

**Proof:** [Proof of Lemma 2.4] It is clear that
\[
A := \left\| \left( \begin{array}{c} w - \eta \nabla_w f(w, v) \\
 v + \eta \nabla_v f(w, v) \end{array} \right) - \left( \begin{array}{c} w' - \eta \nabla_w f(w', v') \\
 v' + \eta \nabla_v f(w', v') \end{array} \right) \right\|_2^2
= \left\| \left( \begin{array}{c} w - w' \\
 v - v' \end{array} \right) \right\|_2^2 + \eta^2 \left\| \left( \begin{array}{c} \nabla_w f(w', v') - \nabla_w f(w, v) \\
 \nabla_v f(w', v') - \nabla_v f(w, v) \end{array} \right) \right\|_2^2
- 2\eta \left\langle \left( \begin{array}{c} w - w' \\
 v - v' \end{array} \right), \left( \begin{array}{c} \nabla_w f(w, v) - \nabla_w f(w', v') \\
 \nabla_v f(w', v') - \nabla_v f(w, v) \end{array} \right) \right\rangle.
\]

Plugging (2.9) to the above inequality, we derive
\[ A \leq \left( 1 - 2\rho\eta \right) \left\| \left( \begin{array}{c} w - w' \\
 v - v' \end{array} \right) \right\|_2^2 + \eta^2 \left\| \left( \begin{array}{c} \nabla_w f(w', v') - \nabla_w f(w, v) \\
 \nabla_v f(w', v') - \nabla_v f(w, v) \end{array} \right) \right\|_2^2. \]

We can combine the above inequality with the Lipschitz continuity to derive Part (a). We refer the interested readers to [Farnia and Ozdaglar, 2021] for the proof of Part (b).

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

**Lemma 2.6 (Chernoff’s Bound)** Let \( X_1, \ldots, X_t \) be independent random variables taking values in \( \{0, 1\} \). Let \( X = \sum_{j=1}^t X_j \) and \( \mu = \mathbb{E}[X] \). Then for any \( \delta > 0 \) with probability at
least $1 - \exp(-\mu \tilde{\delta}^2/(2 + \tilde{\delta}))$ we have $X \leq (1 + \tilde{\delta})\mu$. Furthermore, for any $\delta \in (0,1)$ with probability at least $1 - \delta$ we have

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

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

**Theorem 2.2** Assume for all $z$, the function $(w, v) \mapsto f(w, v; z)$ is convex-concave. Let the algorithm $A$ be SGDA (2.8) with $t$ iterations. Let $\delta \in (0, 1)$.

(a) Assume $\eta_t = \eta$. If Assumption 2.1 holds, then $A$ is $4\eta G(\sqrt{t} + t/n)$-argument-stable in expectation.

(b) If Assumptions 2.1, 2.2 hold, then $A$ is $\epsilon$-argument-stable in expectation, where

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

(c) Let $\eta_t = \eta$. If Assumption 2.1 holds, then $A$ is $\epsilon$-argument-stable with probability at least $1 - \delta$, where

$$\epsilon \leq \sqrt{8eG\eta} \left(\sqrt{t} + t/n + \log(1/\delta) + \sqrt{2tn^{-1}\log(1/\delta)}\right).$$

(d) Let $\eta_t = \eta$. If Assumptions 2.1, 2.2 hold, then $A$ is $\epsilon$-argument-stable with probability at least $1 - \delta$, where

$$\epsilon \leq \sqrt{8eG\eta} \exp\left(2^{-1}L^2 t\eta_t^2\right) \times \left(1 + t/n + \log(1/\delta) + \sqrt{2tn^{-1}\log(1/\delta)}\right).$$

(e) If $(w, v) \mapsto f(w, v; z)$ is $\rho$-SC-SC, Assumption 2.1 holds and $\eta_t = 1/(\rho t)$, then $A$ is
\[ \epsilon \text{-argument-stable in expectation, where} \]
\[ \epsilon \leq \frac{2\sqrt{2}G}{\rho} \left( \frac{\log(et)}{t} + \frac{1}{n(n-2)} \right)^{\frac{1}{2}}. \]

**Proof:** Let \( \{w_t, v_t\} \) and \( \{w'_t, v'_t\} \) be the sequence produced by (2.8) w.r.t. \( S \) and \( S' \), respectively.

We first prove Part (a). Note that the projection step is nonexpansive. We consider two cases at the \( t \)-th iteration. If \( i_t \neq n \), then it follows from Part (a) of Lemma 2.4 that
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} w_t - \eta_t \nabla_w f(w_t, v_t; z_i_t) - w'_t + \eta_t \nabla_w f(w'_t, v'_t; z_i_t) \\ v_t + \eta_t \nabla_v f(w_t, v_t; z_i_t) - v'_t - \eta_t \nabla_v f(w'_t, v'_t; z_i_t) \end{pmatrix} \right\|_2^2
\]
\[
\leq \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + 8G^2 \eta_t^2. \tag{2.10}
\]

If \( i_t = n \), then it follows from the elementary inequality \((a + b)^2 \leq (1 + p)a^2 + (1 + 1/p)b^2\) \((p > 0)\) that
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} w_t - \eta_t \nabla_w f(w_t, v_t; z_n) - w'_t + \eta_t \nabla_w f(w'_t, v'_t; z'_n) \\ v_t + \eta_t \nabla_v f(w_t, v_t; z_n) - v'_t - \eta_t \nabla_v f(w'_t, v'_t; z'_n) \end{pmatrix} \right\|_2^2
\]
\[
\leq (1 + p) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + (1 + 1/p) \eta_t^2 \left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_n) - \nabla_w f(w'_t, v'_t; z'_n) \\ \nabla_v f(w_t, v_t; z_n) - \nabla_v f(w'_t, v'_t; z'_n) \end{pmatrix} \right\|_2^2. \tag{2.11}
\]

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

$$\mathbb{E}_{it} \left[ \left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|^2 \right] \leq \frac{n-1}{n} \left( \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2 + 8G^2 \eta_t^2 \right)$$

$$+ \frac{1+p}{n} \left( \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2 + \frac{8(1+1/p)}{n} \eta_t^2 G^2 \right)$$

$$= (1+p/n) \left( \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2 + 8\eta_t^2 G^2 (1+1/(np)) \right).$$

Applying this inequality recursively implies that

$$\mathbb{E}_{A} \left[ \left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|^2 \right] \leq 8\eta_t^2 G^2 \sum_{k=1}^{t} \left( 1 + \frac{p}{n} \right)^{-k}$$

$$= 8\eta_t^2 G^2 \left( 1 + \frac{1}{np} \right) \frac{n}{p} \left( \left( 1 + \frac{p}{n} \right)^t - 1 \right)$$

$$= 8\eta_t^2 G^2 \left( \frac{n}{p} + \frac{1}{p^2} \right) \left( \left( 1 + \frac{p}{n} \right)^t - 1 \right).$$

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

$$\mathbb{E}_{A} \left[ \left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|^2 \right] \leq 16\eta_t^2 G^2 \left( t + \frac{t^2}{n^2} \right).$$

The stated bound then follows by Jensen’s inequality.

We now prove Part (b). Analogous to (2.10), we can use Part (b) of Lemma 2.4 to derive

$$\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|^2 \leq (1 + L^2 \eta_t^2) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2$$
in the case $i_t \neq n$. We can combine the above inequality and (2.11) to derive

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

Applying this inequality recursively, we derive

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

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

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

By taking $p = n/t$ we get

$$
\mathbb{E}_{A} \left[ \left\| \left( \begin{array}{c} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{array} \right) \right\|_2^2 \right] \leq \frac{8G^2(1 + t/n)}{n} \exp \left(L^2 \sum_{j=1}^{t} \eta_j^2 \right) \sum_{k=1}^{t} \eta_k^2.
$$

The stated result then follows from the Jensen’s inequality.
We now prove Part (c). According to the analysis in Part (a), we know
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left( \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + 8G^2\eta_t^2 \right) \mathbb{I}_{[i \neq n]} \\
+ (1 + p) \left\| \begin{pmatrix} w_t - w'_{t+1} \\ v_t - v'_{t+1} \end{pmatrix} \right\|_2^2 + 8(1 + 1/p)\eta_t^2 G^2 \right) \mathbb{I}_{[i = n]}.
\]

It then follows that
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq 1 + p \mathbb{I}_{[i = n]} \left( \left\| \begin{pmatrix} w_t - w'_{t+1} \\ v_t - v'_{t+1} \end{pmatrix} \right\|_2^2 + 8G^2\eta_t^2 \right) \mathbb{I}_{[i = n]}.
\]

(2.12)

Applying this inequality recursively gives
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8G^2\eta_t^2 \sum_{k=1}^t (1 + \mathbb{I}_{[i_k = n]}/p) \prod_{j=k+1}^{t} (1 + p\mathbb{I}_{[i_j = n]}) \\
= 8G^2\eta_t^2 \sum_{k=1}^t (1 + \mathbb{I}_{[i_k = n]}/p) \prod_{j=k+1}^{t} (1 + p\mathbb{I}_{[i_j = n]}) \\
\leq 8G^2\eta_t^2 (1 + p)\sum_{j=1}^t \mathbb{I}_{[i_j = n]}(t + \sum_{k=1}^t \mathbb{I}_{[i_k = n]}/p).
\]

Applying Lemma 2.6 with \(X_j = \mathbb{I}_{[i_j = n]}\) and \(\mu = t/n\) (note \(\mathbb{E}[X_j] = 1/n\)), with probability \(1 - \delta\) there holds
\[
\sum_{j=1}^t \mathbb{I}_{[i_j = n]} \leq t/n + \log(1/\delta) + \sqrt{2tn^{-1}\log(1/\delta)}.
\]

(2.13)

The following inequality then holds with probability at least \(1 - \delta\)
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8G^2\eta_t^2 (1 + p)^{t/n+\log(1/\delta)+\sqrt{2tn^{-1}\log(1/\delta)}} \\
\times \left( t + t/(pn) + p^{-1}\log(1/\delta) + p^{-1}\sqrt{2tn^{-1}\log(1/\delta)} \right).
\]

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

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

This finishes the proof of Part (c).

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

$$\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left( 1 + L^2\eta^2_t \right) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 \mathbb{I}_{[i_t \neq n]}$$

$$+ \left( 1 + p \right) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + 8(1 + 1/p)\eta^2_t G^2 \mathbb{I}_{[i_t = n]}.$$ 

It then follows that

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

We can apply the above inequality recursively and derive

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

$$\leq 8(1 + 1/p)G^2\eta^2 \sum_{k=1}^t \mathbb{I}_{[i_k = n]} \prod_{j=k+1}^t \left( 1 + L^2\eta^2_j \right) \prod_{j=k+1}^t \left( 1 + p\mathbb{I}_{[i_j = n]} \right)$$

$$= 8(1 + 1/p)G^2\eta^2 \sum_{k=1}^t \mathbb{I}_{[i_k = n]} \prod_{j=k+1}^t \left( 1 + L^2\eta^2_j \right) \prod_{j=k+1}^t \left( 1 + p\mathbb{I}_{[i_j = n]} \right)$$

$$\leq 8(1 + 1/p)G^2\eta^2 \prod_{j=1}^t \left( 1 + L^2\eta^2_j \right) \prod_{j=1}^t \left( 1 + p\mathbb{I}_{[i_j = n]} \right) \sum_{k=1}^t \mathbb{I}_{[i_k = n]}.$$ 

35
It then follows from the elementary inequality $1 + x \leq e^x$ that
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8(1 + 1/p)G^2 \eta^2 \exp \left( L^2 \sum_{j=1}^{t} \eta_j^2 \right) \left( 1 + p \right) \sum_{k=1}^{t} \sum_{[i_k = n]} \mathbb{I}
\]
According to (2.13), we get the following inequality with probability at least $1 - \delta$
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8(1 + 1/p)G^2 \eta^2 \exp \left( L^2 t \eta^2 \right) \left( 1 + p \right)^{t/n + \log(1/\delta) + \sqrt{2tn^{-1} \log(1/\delta)}}
\]
\[
\times \left( t/n + \log(1/\delta) + \sqrt{2tn^{-1} \log(1/\delta)} \right).
\]

We can choose $p = \frac{1}{t/n + \log(1/\delta) + \sqrt{2tn^{-1} \log(1/\delta)}}$ and derive the following inequality with probability at least $1 - \delta$
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq 8G^2 \eta^2 \exp \left( L^2 t \eta^2 \right) \left( 1 + t/n + \log(1/\delta) + \sqrt{2tn^{-1} \log(1/\delta)} \right)^2.
\]

The stated bound then follows.

If $i_t \neq n$, we can analyze analogously to (2.10) excepting using the strong convexity, and show
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq (1 - 2\rho \eta_t) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + 8G^2 \eta_t^2.
\]
If $i_t = n$, then (2.11) holds. We can combine the above two cases and derive
\[
\mathbb{E}_{i_t} \left[ \left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \right]
\leq \frac{n - 1}{n} \left( 1 - 2\rho \eta_t \right) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + \frac{1 + p}{n} \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + \frac{8(1 + 1/p)}{n} \eta_t^2 G^2
\]
\[
= (1 - 2\rho \eta_t + (2\rho \eta_t + p)/n) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + 8\eta_t^2 G^2(1 + 1/(np)).
\]
We can choose \( p = \rho \eta (n - 2) \) to derive

\[
\mathbb{E}_{\tau_t} \left[ \left\| \left( \begin{array}{c}
w_{t+1} - w'_{t+1} \\
v_{t+1} - v'_{t+1}
\end{array} \right) \right\|_2^2 \right] \leq (1 - \rho \eta_t) \left( \left\| \left( \begin{array}{c}w_t - w'_t \\
v_t - v'_t
\end{array} \right) \right\|_2^2 + 8 \eta_t^2 G^2 \left( 1 + \frac{1}{n(n - 2) \rho \eta_t} \right). \]

It then follows that

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

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

\[
\mathbb{E}_A \left[ \left\| \left( \begin{array}{c}
w_{t+1} - w'_{t+1} \\
v_{t+1} - v'_{t+1}
\end{array} \right) \right\|_2^2 \right] \leq \frac{8 G^2}{t \rho} \sum_{j=1}^{t} \left( (\rho t)^{-1} + \frac{1}{n(n - 2) \rho} \right) \leq \frac{8 G^2}{\rho^2} \left( \frac{\log(\epsilon t)}{t} + \frac{1}{n(n - 2)} \right).
\]

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

\[ \Box \]

**Remark 2.3** We consider stability bounds under various assumptions on loss functions. We now sketch the technical difference in our analysis. Let \( \delta_t := \| w_t - w'_t \|^2_2 + \| v_t - v'_t \|^2_2 \), where \((w_t, v_t), (w'_t, v'_t)\) are SGDA iterates for \( S \) and \( S' \) differing only by the last element. For convex-concave and nonsmooth problems, we show \( \delta_{t+1} = \delta_t + O(\eta_t^2) \) if \( i_t \neq n \). For convex-concave and smooth problems, we show \( \delta_{t+1} = (1 + O(\eta_t^2)) \delta_t \) if \( i_t \neq n \). For \( \rho \)-SC-SC and nonsmooth problems, we show \( \delta_{t+1} = (1 - 2 \rho \eta_t) \delta_t + O(\eta_t^2) \) if \( i_t \neq n \). For the above cases, we first control \( \delta_{t+1} \) and then take expectation w.r.t. \( i_t \). A key point to tackle nonsmooth problems is to consider the evolution \( \delta_t \) instead of \( \| w_t - w'_t \|_2 + \| v_t - v'_t \|_2 \), which is able to yield nontrivial bounds by making \( \sum_t \eta_t^2 = o(1) \) with sufficiently small \( \eta_t \).

### 2.3.2 Population Risks

In this section, we present optimization error bounds for SGDA, which are standard in the literature [Nedić and Ozdaglar 2009; Nemirovski et al. 2009]. We give both bounds in expectation and bounds with high probability. The high-probability analysis requires to use concentration inequalities for martingales. Lemma 2.7 is an Azuma-Hoeffding inequality for real-valued martingale difference sequence [Hoeffding 1963], while Lemma 2.8 is a Bernstein-
type inequality for martingale difference sequences in a Hilbert space [Tarres and Yao, 2014].

Lemma 2.7 Let \( \{\xi_k : k \in \mathbb{N}\} \) be a martingale difference sequence taking values in \( \mathbb{R} \), i.e., \( \mathbb{E}[\xi_k | \xi_1, \ldots, \xi_{k-1}] = 0 \). Assume that \( |\xi_k - \mathbb{E}_k[\xi_k]| \leq b_k \) for each \( k \). For \( \delta \in (0, 1) \), with probability at least \( 1 - \delta \) we have

\[
\sum_{k=1}^{n} \xi_k \leq \left(2 \sum_{k=1}^{n} b_k^2 \log \frac{1}{\delta}\right)^{\frac{1}{2}}.
\]  (2.14)

Lemma 2.8 Let \( \{\xi_k : k \in \mathbb{N}\} \) be a martingale difference sequence in a Hilbert space with the norm \( \| \cdot \|_2 \). Suppose that almost surely \( \|\xi_k\| \leq B \) and \( \sum_{k=1}^{t} \mathbb{E}[\|\xi_k\|^2 | \xi_1, \ldots, \xi_{k-1}] \leq \sigma_t^2 \) for \( \sigma_t \geq 0 \). Then, for any \( 0 < \delta < 1 \), the following inequality holds with probability at least \( 1 - \delta \)

\[
\max_{1 \leq j \leq t} \left\| \sum_{k=1}^{j} \xi_k \right\| \leq 2 \left( \frac{B}{3} + \sigma_t \right) \log \frac{2}{\delta}.
\]

Lemma 2.9 Let \( \{\mathbf{w}_t, \mathbf{v}_t\} \) be the sequence produced by (2.8) with \( \eta_t = \eta \). Let Assumption 2.1 hold and \( F_S \) be convex-concave. Assume \( \sup_{\mathbf{w} \in \mathcal{W}} \|\mathbf{w}\|_2 \leq D_w \) and \( \sup_{\mathbf{v} \in \mathcal{V}} \|\mathbf{v}\|_2 \leq D_v \). Then the following inequality holds

\[
\mathbb{E}_A \left[ \sup_{\mathbf{v} \in \mathcal{V}} F_S(\mathbf{w}_T, \mathbf{v}) - \inf_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w}, \mathbf{v}_T) \right] \leq \eta G^2 + \frac{D_w^2 + D_v^2}{2T\eta} + \frac{G(D_w + D_v)}{\sqrt{T}},
\]  (2.15)

where \( (\mathbf{w}_T, \mathbf{v}_T) \) is defined in (2.26). Let \( \delta \in (0, 1) \). Then with probability at least \( 1 - \delta \) we have

\[
\sup_{\mathbf{v} \in \mathcal{V}} F_S(\mathbf{w}_T, \mathbf{v}) - \inf_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w}, \mathbf{v}_T) \leq \eta G^2 + \frac{D_w^2 + D_v^2}{2T\eta} + \frac{G(D_w + D_v)(9\log(6/\delta) + 2)}{\sqrt{T}}.
\]  (2.16)

Proof: According to the non-expansiveness of projection and (2.8), we know

\[
\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \leq \|\mathbf{w}_t - \eta_t \nabla_w f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_t) - \mathbf{w}\|_2^2
\]

\[
= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\nabla_w f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_t)\|_2^2 + 2\eta_t \langle \mathbf{w} - \mathbf{w}_t, \nabla_w f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_t) \rangle
\]

\[
\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 G^2 + 2\eta_t \langle \mathbf{w} - \mathbf{w}_t, \nabla_w F_S(\mathbf{w}_t; \mathbf{v}_t) \rangle
\]

\[
+ 2\eta_t \langle \mathbf{w} - \mathbf{w}_t, \nabla_w f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_t) - \nabla_w F_S(\mathbf{w}_t, \mathbf{v}_t) \rangle,
\]

38
where we have used Assumption 2.1. According to the convexity of \( F_S(\cdot, v_t) \), we know

\[
2\eta_t \langle F_S(w_t, v_t) - F_S(w, v_t) \rangle \leq \|w_t - w\|^2_2 - \|w_{t+1} - w\|^2_2 + \eta_t^2 G^2 + 2\eta_t \langle w - w_t, \nabla_w f(w_t, v_t; z_i) - \nabla_w F_S(w_t, v_t) \rangle. \tag{2.17}
\]

Taking a summation of the above inequality from \( t = 1 \) to \( t = T \) (\( w_1 = 0 \)), we derive

\[
2\eta \sum_{t=1}^{T} (F_S(w_t, v_t) - F_S(w, v_t)) \leq \|w\|^2_2 + T\eta^2 G^2
\]

\[
+ 2\eta \sum_{t=1}^{T} \langle w_t, \nabla_w F_S(w_t, v_t) - \nabla_w f(w_t, v_t; z_i) \rangle + 2\eta \sum_{t=1}^{T} (w, \nabla_w f(w_t, v_t; z_i) - \nabla_w F_S(w_t, v_t)).
\]

It then follows from the concavity of \( F_S(w, \cdot) \) and Schwartz’s inequality that

\[
2\eta \sum_{t=1}^{T} (F_S(w_t, v_t) - F_S(w, \bar{v}_T)) \leq 2\eta \sum_{t=1}^{T} (w_t, \nabla_w F_S(w_t, v_t) - \nabla_w f(w_t, v_t; z_i))
\]

\[
+ 2\eta D_w \left\| \sum_{t=1}^{T} (\nabla_w f(w_t, v_t; z_i) - \nabla_w F_S(w_t, v_t)) \right\|_2^2 + D_w^2 + T\eta^2 G^2. \tag{2.18}
\]

Since the above inequality holds for all \( w \), we further get

\[
2\eta \sum_{t=1}^{T} (F_S(w_t, v_t) - \inf_w F_S(w, \bar{v}_T)) \leq 2\eta \sum_{t=1}^{T} (w_t, \nabla_w F_S(w_t, v_t) - \nabla_w f(w_t, v_t; z_i))
\]

\[
+ 2\eta D_w \left\| \sum_{t=1}^{T} (\nabla_w f(w_t, v_t; z_i) - \nabla_w F_S(w_t, v_t)) \right\|_2^2 + D_w^2 + T\eta^2 G^2. \tag{2.18}
\]

Note

\[
\mathbb{E}_{\xi_i} \left[ (w_t, \nabla_w F_S(w_t, v_t) - \nabla_w f(w_t, v_t; z_i)) \right] = 0. \tag{2.19}
\]

We can take an expectation over both sides of (2.18) and get

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_A [F_S(w_t, v_t)] - \mathbb{E}_A [\inf_w F_S(w, \bar{v}_T)]
\]

\[
\leq \frac{D_w^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{D_w}{T} \mathbb{E}_A \left[ \left\| \sum_{t=1}^{T} (\nabla_w f(w_t, v_t; z_i) - \nabla_w F_S(w_t, v_t)) \right\|_2^2 \right].
\]
According to Jensen’s inequality and (2.19), we know

\[
\mathbb{E}_A \left[ \left\| \sum_{t=1}^{T} (\nabla_w f(w_t, v_t; z_{t_t}) - \nabla_w F_S(w_t, v_t)) \right\|^2 \right]^2 \leq \mathbb{E}_A \left[ \left\| \sum_{t=1}^{T} (\nabla_w f(w_t, v_t; z_{t_t}) - \nabla_w F_S(w_t, v_t)) \right\|^2 \right] = \sum_{t=1}^{T} \mathbb{E}_A \left[ \left\| \nabla_w f(w_t, v_t; z_{t_t}) - \nabla_w F_S(w_t, v_t) \right\|^2 \right] \leq TG^2.
\]

It then follows that

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_A \left[ F_S(w_t, v_t) \right] - \mathbb{E}_A \left[ \inf_w F_S(w, \bar{v}_T) \right] \leq \frac{D_w^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{D_w G}{\sqrt{T}}. \tag{2.20}
\]

In a similar way, we can show that

\[
\mathbb{E}_A \left[ \sup_v F_S(\bar{w}_T, v) \right] - \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_A \left[ F_S(w_t, v_t) \right] \leq \frac{D_v^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{D_v G}{\sqrt{T}}. \tag{2.21}
\]

The stated bound (2.15) then follows from (2.20) and (2.21).

We now turn to (2.16). It is clear that \(|\langle w_t, \nabla F_S(w_t, v_t) - \nabla f(w_t, v_t; z_{t_t}) \rangle| \leq 2GD_w\), and therefore we can apply Lemma 2.7 to derive the following inequality with probability at least \(1 - \delta/6\) that

\[
\sum_{t=1}^{T} \langle w_t, \nabla_w F_S(w_t, v_t) - \nabla_w f(w_t, v_t; z_{t_t}) \rangle \leq 2GD_w \left( 2T \log(6/\delta) \right)^{1/2}. \tag{2.22}
\]

For any \(t \in \mathbb{N}\), define \(\xi_t = \nabla_w f(w_t, v_t; z_{t_t}) - \nabla_w F_S(w_t, v_t)\). Then it is clear that \(\|\xi_t\|_2 \leq 2G\) and

\[
\sum_{t=1}^{T} \mathbb{E}[\|\xi_t\|^2] \leq 4TG^2.
\]

Therefore, we can apply Lemma 2.8 to derive the following inequality with probability at
least $1 - \delta/3$
\[
\left\| \sum_{t=1}^{T} \xi_t \right\|_2 \leq 2 \left( \frac{2G}{3} + 2G \sqrt{T} \right) \log(6/\delta).
\]

Then, the following inequality holds with probability at least $1 - \delta/3$
\[
\left\| \sum_{t=1}^{T} \left( \nabla_w f(w_t, v_t; z_{it}) - \nabla_w F_S(w_t, v_t) \right) \right\|_2 \leq 4G \left( 1 + \sqrt{T} \right) \log(6/\delta).
\]

We can plug the above inequality and (2.22) back into (2.18), and derive the following inequality with probability at least $1 - \delta/2$
\[
\frac{1}{T} \sum_{t=1}^{T} F_S(w_t, v_t) - \inf_w F_S(w, v_T) \leq \frac{D_w^2}{2T\eta} + \frac{\eta G^2}{2} + \frac{2GD_w \sqrt{2\log(6/\delta)}}{\sqrt{T}} + \frac{8D_w G \log(6/\delta)}{\sqrt{T}}.
\]

In a similar way, we can get the following inequality with probability at least $1 - \delta/2$
\[
\sup_{v \in V} F_S(w_T, v) - \frac{1}{T} \sum_{t=1}^{T} F_S(w_t, v_t) \leq \frac{D_v^2}{2T\eta} + \frac{\eta G^2}{2} + \frac{9D_v G \log(6/\delta)}{2 \sqrt{T}} + 2D_v G.
\]

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

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

**Lemma 2.10** Let Assumption 2.1 hold, $t_0 \geq 0$ and $F_S(\cdot, \cdot)$ be $\rho$-SC-SC with $\rho > 0$. Let $\{w_t, v_t\}$ be the sequence produced by (2.8) with $\eta_t = 1/(\rho(t + t_0))$. If $t_0 = 0$, then for $(\bar{w}_T, \bar{v}_T)$ defined in (2.26) we have
\[
\mathbb{E}_A \left[ \sup_{v \in V} F_S(\bar{w}_T, v) - \inf_{w \in W} F_S(w, \bar{v}_T) \right] \leq \frac{G^2 \log(eT)}{\rho T} - \frac{(D_w + D_v)G}{\sqrt{T}}.
\]  
(2.23)

If $\sup_{w \in W} \|w\|_2 \leq D_w$ and $\sup_{v \in V} \|v\|_2 \leq D_v$, then
\[
\Delta_S^w(\bar{w}_T, \bar{v}_T) \leq \frac{2\rho t_0 (D_w^2 + D_v^2)}{T} + \frac{G^2 \log(eT)}{\rho T}.
\]  
(2.24)

*Proof:* Analyzing analogously to (2.17) but using the strong convexity of $w \mapsto F_S(w, v)$,
we derive

\[2\eta_t [F_S(w_t, v_t) - F_S(w, v_t)] \leq (1 - \eta_t \rho)\|w_t - w\|_2^2 - \|w_{t+1} - w\|_2^2 + \eta_t^2 G^2 + \xi_t(w),\]

where \(\xi_t(w) = 2\eta_t \langle w - w_t, \nabla_w f(w_t, v_t; z_i) - \nabla_w F_S(w_t, v_t)\rangle\). Since \(\eta_t = 1/(\rho(t + t_0))\), we further get

\[
\frac{2}{\rho(t + t_0)} [F_S(w_t, v_t) - F_S(w, v_t)] \leq (1 - 1/(t + t_0))\|w_t - w\|_2^2 - \|w_{t+1} - w\|_2^2 + \frac{G^2}{\rho^2(t + t_0)^2} + \xi_t(w).
\]

Multiplying both sides by \(t + t_0\) gives

\[
\frac{2}{\rho} [F_S(w_t, v_t) - F_S(w, v_t)] \leq (t + t_0 - 1)\|w_t - w\|_2^2 - (t + t_0)\|w_{t+1} - w\|_2^2 + (t + t_0)\xi_t(w) + \frac{G^2}{\rho^2(t + t_0)}.
\]

Taking a summation of the above inequality further gives

\[
\sum_{t=1}^{T} [F_S(w_t, v_t) - F_S(w, v_t)] \leq 2\rho t_0 D_w^2 + \frac{G^2 \log(eT)}{2\rho} + \frac{\rho}{2} \sum_{t=1}^{T} (t + t_0)\xi_t(w),
\]

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

\[
\sum_{t=1}^{T} [F_S(w_t, v_t) - F_S(w, \bar{v}_T)] \leq 2\rho t_0 D_w^2 + \frac{G^2 \log(eT)}{2\rho} + \frac{\rho}{2} \sum_{t=1}^{T} (t + t_0)\xi_t(w). \quad (2.25)
\]

Since the above inequality holds for any \(w\), we know

\[
\sum_{t=1}^{T} [F_S(w_t, v_t) - \inf_{w \in W} F_S(w, \bar{v}_T)] \leq 2\rho t_0 D_w^2 + \frac{G^2 \log(eT)}{2\rho} + \frac{\rho}{2} \sup_{w \in W} \sum_{t=1}^{T} (t + t_0)\xi_t(w).
\]
Since \( \mathbb{E}_A[(\mathbf{w}_t, \nabla_w f(\mathbf{w}_t, v_t; z_t)) - \nabla_w F_S(\mathbf{w}_t, v_t)] = 0 \) we know

\[
\mathbb{E}_A\left[ \sup_{\mathbf{w} \in W} \sum_{t=1}^T (t + t_0) \xi_t(\mathbf{w}) \right] = 2 \mathbb{E}_A\left[ \sup_{\mathbf{w} \in W} \sum_{t=1}^T (t + t_0) \eta_t(\mathbf{w}, \nabla_w f(\mathbf{w}_t, v_t; z_t) - \nabla_w F_S(\mathbf{w}_t, v_t)) \right]
\leq 2 \sup_{\mathbf{w} \in W} \left\| \mathbf{w} \right\|_2 \mathbb{E}_A\left[ \sum_{t=1}^T (t + t_0) \eta_t(\nabla_w f(\mathbf{w}_t, v_t; z_t) - \nabla_w F_S(\mathbf{w}_t, v_t)) \right]
\leq 2D_w \left( \mathbb{E}_A\left[ \sum_{t=1}^T (t + t_0) \eta_t \nabla_w f(\mathbf{w}_t, v_t; z_t) - \nabla_w F_S(\mathbf{w}_t, v_t) \right] \right)^{1/2}
\leq 2D_w \left( \sum_{t=1}^T (t + t_0)^2 \eta_t \mathbb{E}_A\left[ \nabla_w f(\mathbf{w}_t, v_t; z_t) \right] \right)^{1/2}
\leq 2D_w \left( \sum_{t=1}^T (t + t_0)^2 \eta_t \mathbb{E}_A\left[ \nabla_w f(\mathbf{w}_t, v_t; z_t) \right] \right)^{1/2} \leq 2D_w G \rho^{-1} \sqrt{T}.
\]

We can combine the above two inequalities together and derive

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

In a similar way one can show

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

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

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

This proves (2.23) with \( t_0 = 0 \).

We now turn to (2.24). Since \( \mathbb{E}_A[\xi_t(\mathbf{w})] = 0 \), it follows from (2.25) that

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

In a similar way, one can show

\[
\sum_{t=1}^T \mathbb{E}_A[F_S(\bar{w}_T, v) - F_S(\mathbf{w}_t, v_t)] \leq 2\rho t_0 D_v^2 + \frac{G^2 \log(eT)}{2\rho}.
\]
We can combine the above two inequalities together and derive

$$E[F_S(\bar{w}_T, v) - F_S(w, \bar{v}_T)] \leq \frac{2\rho t_0 (D^2_w + D^2_v)}{T} + \frac{G^2 \log(eT)}{\rho T}.$$  

The stated bound (2.24) then follows by taking the supremum over $w$ and $v$. 

We now use stability bounds in Theorem 2.2 to develop error bounds of SGDA which outputs an average of iterates

$$\bar{w}_T = \frac{\sum_{t=1}^{T} \eta_t w_t}{\sum_{t=1}^{T} \eta_t} \quad \text{and} \quad \bar{v}_T = \frac{\sum_{t=1}^{T} \eta_t v_t}{\sum_{t=1}^{T} \eta_t}. \quad (2.26)$$

The underlying reason to introduce the average operator is to simplify the optimization error analysis [Nemirovski et al., 2009]. Indeed, our stability and generalization analysis applies to any individual iterates. As a comparison, the optimization error analysis for the last iterate is much more difficult than that for the averaged iterate. We use the notation $B \asymp \tilde{B}$ if there exist constants $c_1, c_2 > 0$ such that $c_1 e^B < B \leq c_2 e^B$.

**Theorem 2.3 (Weak PD risk)** Let $\{w_t, v_t\}$ be produced by (2.8). Assume for all $z$, the function $(w, v) \mapsto f(w, v; z)$ is convex-concave. Let $A$ be defined by $A_w(S) = \bar{w}_T$ and $A_v(S) = \bar{v}_T$ for $(\bar{w}_T, \bar{v}_T)$ in (2.26). Assume $\sup_{w \in W} \|w\|_2 \leq D_w$ and $\sup_{v \in V} \|v\|_2 \leq D_v$.

(a) If $\eta_t = \eta$ and Assumption 2.1 holds, then

$$\Delta^w(\bar{w}_T, \bar{v}_T) \leq 4\sqrt{\eta} G^2 \left( \sqrt{T} + \frac{T}{n} \right) + \eta G^2 + \frac{D^2_w + D^2_v}{2\eta T} + \frac{G(D_w + D_v)}{\sqrt{T}}. \quad (2.27)$$

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

(b) If $\eta_t = \eta$ and Assumptions 2.1, 2.2 hold, then

$$\Delta^w(\bar{w}_T, \bar{v}_T) \leq \frac{4\sqrt{e(T + T^2/n)} G^2 \eta \exp(LT\eta^2/2)}{\sqrt{n}} + \eta G^2 + \frac{D^2_w + D^2_v}{2\eta T} + \frac{G(D_w + D_v)}{\sqrt{T}}. \quad (2.28)$$

We can choose $T \asymp n$ and $\eta \asymp T^{-1/2}$ to derive $\Delta^w(\bar{w}_T, \bar{v}_T) = O(n^{-1/2})$. 
(c) If $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is $\rho$-SC-SC ($\rho > 0$), Assumption 2.1 holds, $\eta_t = 1/(\rho T)$ and $T \asymp n^2$, then
\[ \Delta^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = \mathcal{O}(\sqrt{\log n/(n \rho)}). \]

(d) If $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is $\rho$-SC-SC ($\rho > 0$), Assumptions 2.1, 2.2 hold, $\eta_t = 1/(\rho (t + t_0))$ with $t_0 \geq L^2/\rho^2$ and $T \asymp n$, then $\Delta^w(\mathbf{w}_T, \mathbf{v}_T) = \mathcal{O}(\log(n)/(n \rho))$.

Proof: We first prove Part (a). We have the decomposition
\[ \Delta^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = \Delta^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) - \Delta^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) + \Delta^w_S(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T). \quad (2.29) \]

According to Part (a) of Theorem 2.2 we know the following inequality for all $t$
\[ \mathbb{E}_A \left[ \left\| \left( \begin{array}{c} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{array} \right) \right\|_2 \right] \leq 4\eta G \left( \sqrt{t} + \frac{t}{n} \right). \]

It then follows from the convexity of a norm that
\[ \mathbb{E}_A \left[ \left\| \left( \begin{array}{c} \bar{w}_T - \bar{w}'_T \\ \bar{v}_T - \bar{v}'_T \end{array} \right) \right\|_2 \right] \leq 4\eta G \left( \sqrt{T} + \frac{T}{n} \right) \]

and therefore
\[ \sup_z \left( \sup_{\mathbf{v}' \in \mathcal{V}} \mathbb{E}_A[f(\mathbf{w}_T, \mathbf{v}', z) - f(\mathbf{w}'_T, \mathbf{v}', z)] + \sup_{\mathbf{w}' \in \mathcal{W}} \mathbb{E}_A[f(\mathbf{w}', \mathbf{v}_T, z) - f(\mathbf{w}', \mathbf{v}'_T; z)] \right) \]
\[ \leq G \left( \mathbb{E}_A \left[ \left\| \mathbf{w}_T - \mathbf{w}'_T \right\|_2 \right] + \mathbb{E}_A \left[ \left\| \mathbf{v}_T - \mathbf{v}'_T \right\|_2 \right] \right) \leq 4\sqrt{2} \eta G^2 \left( \sqrt{T} + \frac{T}{n} \right). \]

According to Part (a) of Theorem 2.1, we know
\[ \Delta^w(\mathbf{w}_T, \mathbf{v}_T) - \Delta^w_S(\mathbf{w}_T, \mathbf{v}_T) \leq 4\sqrt{2} \eta G^2 \left( \sqrt{T} + \frac{T}{n} \right). \]

According to Eq. (2.15), we know
\[ \Delta^w_S(\mathbf{w}_T, \mathbf{v}_T) \leq \eta G^2 + \frac{D_w^2 + D_v^2}{2 \eta T} + \frac{G(D_w + D_v)}{\sqrt{T}}. \]
The bound (2.27) then follows directly from (2.29).

Eq. (2.28) in Part (b) can be proved in a similar way (e.g., by combining the stability bounds in Part (b) of Theorem 2.2 and optimization error bounds in Eq. (2.15) together). We omit the proof for brevity.

We now turn to Part (c). According to Part (e) of Theorem 2.2 and the convexity of norm, we know
\[
\mathbb{E}_A \left[ \left\| \begin{pmatrix} \bar{w}_T - \bar{w}'_T \\ \bar{v}_T - \bar{v}'_T \end{pmatrix} \right\|_2^2 \right] \leq \frac{2\sqrt{2}G}{\rho} \left( \frac{\log^2(eT)}{\sqrt{T}} + \frac{1}{\sqrt{n(n-2)}} \right).
\]

Analyzing analogous to Part (a), we further know
\[
\Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^w_S(\bar{w}_T, \bar{v}_T) \leq \frac{4G^2}{\rho} \left( \frac{\log^2(eT)}{\sqrt{T}} + \frac{1}{\sqrt{n(n-2)}} \right).
\]

This together with the optimization error bounds in Lemma 2.10 and (2.29) gives
\[
\Delta^w(\bar{w}_T, \bar{v}_T) \leq \frac{4G^2}{\rho} \left( \frac{\log^2(eT)}{\sqrt{T}} + \frac{1}{\sqrt{n(n-2)}} \right) + \frac{G^2\log(eT)}{\rho T} + \frac{(D_w + D_v)G}{\sqrt{T}}.
\]

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

Finally, we consider Part (d). Since $t_0 \geq L^2/\rho^2$ we know $\eta_t = 1/(\rho(t + t_0)) \leq \rho/L^2$. The stability analysis in Farnia and Ozdaglar [2021] then shows that $A$ is $\epsilon$-argument stable with $\epsilon = \mathcal{O}(1/(\rho n))$. This together with Part (a) of Theorem 2.1 then shows that
\[
\Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^w_S(\bar{w}_T, \bar{v}_T) = \mathcal{O}(1/(\rho n)).
\]

We can combine the above generalization bound and the optimization error bound in (2.24) together, and get
\[
\Delta^w(\bar{w}_T, \bar{v}_T) = \mathcal{O}(1/(\rho n)) + \mathcal{O}\left( \frac{\rho}{T} + \frac{\log(eT)}{\rho T} \right).
\]

The stated bound then follows from $T \asymp n$.

---

Farnia and Ozdaglar [2021] considered the constant step size $\eta_t = \eta \leq \rho/L^2$. It is direct to extend the analysis there to any step size $\eta_t \leq \rho/L^2$ since an algorithm would be more stable if the step size decreases.
Remark 2.4 We first compare our bounds with the related work in a convex-concave setting. Weak PD population risk bounds were established for PPM under Assumptions 2.1, 2.2 [Farinia and Ozdaglar, 2021], which updates $(w_{t+1}^{PPM}, v_{t+1}^{PPM})$ as the saddle point of the following minimax problem

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

In particular, they developed population risk bounds $O(1/\sqrt{n})$ by taking $T \approx \sqrt{n}$ for PPM. However, the implementation of PPM requires to find the exact saddle point at each iteration, which is often computationally expensive. As a comparison, Part (b) shows the minimax optimal population risk bounds $O(1/\sqrt{n})$ for SGDA with $O(n)$ iterations. Weak PD population risk bounds $O(1/\sqrt{n})$ were established for R-ESP [Zhang et al., 2021] without a smoothness assumption, which, however, ignore the interplay between generalization and optimization. In this setting, we show SGDA achieves the same population risk bounds $O(1/\sqrt{n})$ by taking $\eta \approx T^{-3/4}$ and $T \approx n^2$ in Part (a). We now consider the SC-SC setting. Weak PD risk bounds $O(1/(n\rho))$ were established for ESP [Zhang et al., 2021]. Since Farinia and Ozdaglar [2021] did not present an explicit risk bound, we use their stability analysis to give an explicit risk bound $O(\log(n)/(n\rho))$ in the smooth case (Part (d)). As a comparison, we establish the same population risk bounds for SGDA within a logarithmic factor by taking $\eta_t = 1/(\rho_t)$ and $T \approx n^2$ without the smoothness assumption (Part (c)).

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

**Theorem 2.4 (Excess primal risk)** Let $\{w_t, v_t\}$ be produced by (2.8) with $\eta_t = \eta$. Assume for all $z$, the function $(w, v) \mapsto f(w, v; z)$ is convex-concave and the function $v \mapsto F(w, v)$ is $\rho$-strongly-concave. Assume $\sup_{w \in W} \|w\|_2 \leq D_w$ and $\sup_{v \in V} \|v\|_2 \leq D_v$. Let the algorithm $A$ be defined by $A_w(S) = \bar{w}_T$ and $A_v(S) = \bar{v}_T$ for $(\bar{w}_T, \bar{v}_T)$ in (2.26). If
Assumptions 2.1, 2.2 hold, then

\[ \mathbb{E}[R(\tilde{w}_T)] - \inf_{w \in \mathcal{W}} R(w) \leq \eta G^2 + \frac{D^2_w + D^2_v}{2\eta T} + \frac{G(D_w + D_v)}{\sqrt{T}} \]

\[ + \frac{(1 + L/\rho)\sqrt{32e(T + T^2/n)G^2\eta \exp(L^2T\eta^2/2)}}{\sqrt{n}}. \]

In particular, if we choose \( T \approx n \) and \( \eta \approx T^{-1/2} \) then

\[ \mathbb{E}[R(\tilde{w}_T)] - \inf_{w \in \mathcal{W}} R(w) = \mathcal{O}(\sqrt{n}). \] (2.30)

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

\[ R(\tilde{w}_T) - R(w^*) = \mathcal{O}\left((L/\rho)n^{-1/2} \log n \log^2(1/\delta)\right). \] (2.31)

**Proof:** We have the decomposition

\[ R(\tilde{w}_T) - R(w^*) = (R(\tilde{w}_T) - R_S(\tilde{w}_T)) + (R_S(\tilde{w}_T) - F_S(w^*, \tilde{v}_T)) \]

\[ + (F_S(w^*, \tilde{v}_T) - F(w^*, \tilde{v}_T)) + (F(w^*, \tilde{v}_T) - F(w^*, v^*)). \]

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

\[ R(\tilde{w}_T) - R(w^*) \leq (R(\tilde{w}_T) - R_S(\tilde{w}_T)) + (R_S(\tilde{w}_T) - F_S(w^*, \tilde{v}_T)) + (F_S(w^*, \tilde{v}_T) - F(w^*, \tilde{v}_T)). \] (2.32)

Taking an expectation on both sides gives

\[ \mathbb{E}[R(\tilde{w}_T) - R(w^*)] \leq \mathbb{E}[R(\tilde{w}_T) - R_S(\tilde{w}_T)] + \mathbb{E}[R_S(\tilde{w}_T) - F_S(w^*, \tilde{v}_T)] \]

\[ + \mathbb{E}[F_S(w^*, \tilde{v}_T) - F(w^*, \tilde{v}_T)]. \] (2.33)

Note that the first and the third term on the right-hand side is related to generalization, while the second term \( R_S(\tilde{w}_T) - F_S(w^*, \tilde{v}_T) \) is related to optimization. According to Part (b) of Theorem 2.2 we know the following inequality for all \( t \)

\[ \mathbb{E}_{\mathcal{A}} \left[ \left\| \left( \begin{array}{c} \mathbf{w}_{t+1} - \mathbf{w}'_{t+1} \\ \mathbf{v}_{t+1} - \mathbf{v}'_{t+1} \end{array} \right) \right\|_2 \right] \leq \frac{G \sqrt{8e(t + t^2/n)}}{\sqrt{n}} \exp(L^2t\eta^2/2). \]
It then follows from the convexity of a norm that

$$\mathbb{E}_A \left[ \left\| \begin{pmatrix} \bar{w}_T & \bar{v}_T \\ \bar{w}'_T & \bar{v}'_T \end{pmatrix} \right\|_2 \right] \leq G \frac{\sqrt{8c(T + T^2/n)}}{\sqrt{n}} \exp \left( \frac{L^2 T \eta^2}{2} \right) \eta. \quad (2.34)$$

This together with Part (b) of Theorem 2.1 implies that

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

Similarly, the stability bound (2.34) also implies the following bound on the gap between the population and empirical risk

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

According to Lemma 2.9, we know

$$\mathbb{E}_A \left[ R_S(\bar{w}_T) - F_S(\bar{w}_T, \bar{v}_T) \right] \leq \mathbb{E}_A \left[ \sup_{v \in V} F_S(\bar{w}_T, v) - \inf_{w \in W} F_S(w, \bar{v}_T) \right] \leq \eta G^2 + \frac{D_w^2 + D_v^2}{2\eta T} + \frac{G(D_w + D_v)}{\sqrt{T}}.$$

We can plug the above three inequalities back into (2.33), and derive the stated bound on the excess primal population risk in expectation.

We now turn to the high-probability bounds. According to Assumption 2.1 and Part (d) of Theorem 2.2, we know that with probability at least $1 - \delta/4$ that SGDA is $\epsilon$-uniformly stable, where $\epsilon$ satisfies

$$\epsilon = O \left( \eta \exp \left( \frac{L^2 T \eta^2}{2} \right) \left( T n^{-1} + \log(1/\delta) \right) \right). \quad (2.35)$$

This together with Part (d) of Theorem 2.1 implies the following inequality with probability at least $1 - \delta/2$

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

where $\epsilon$ satisfies (2.35). In a similar way, one can use Part (d) of Theorem 2.1 and stability
bounds in Part (d) of Theorem 2.2 to show the following inequality with probability at least $1 - \delta/4$

$$F_S(w^*, v_T) - F(w^*, v_T) = \mathcal{O}\left(\log n \log(1/\delta)\epsilon\right) + \mathcal{O}\left(n^{-\frac{1}{2}} \log^2(1/\delta)\right). \tag{2.36}$$

According to (2.16), we derive the following inequality with probability at least $1 - \delta/4$

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

We can plug the above three inequalities back into (2.32) and derive the following inequality with probability at least $1 - \delta$

$$R(\bar{w}_T) - R(w^*) = \mathcal{O}\left(L^{-1} \eta \exp(L^2 T \eta^2/2) \log n \log(1/\delta)(Tn^{-1} + \log(1/\delta))\right)$$

$$+ \mathcal{O}\left(n^{-\frac{1}{2}} \sqrt{\log(1/\delta)}\right) + \mathcal{O}\left(\eta + (T\eta)^{-1} + T^{-\frac{1}{2}} \log(1/\delta)\right). \tag{2.37}$$

The high-probability bound (2.31) then follows from the choice of $T$ and $\eta$.

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

**Theorem 2.5 (High-probability bounds)** Let $\{w_t, v_t\}$ be the sequence produced by (2.8) with $\eta_t = \eta$. Assume for all $z$, the function $(w, v) \mapsto f(w, v; z)$ is convex-concave. Let $\mathcal{A}$ be defined by $A_w(S) = \bar{w}_T$ and $A_v(S) = \bar{v}_T$ for $(\bar{w}_T, \bar{v}_T)$ in (2.26). Let $\sup_{w \in \mathcal{W}} \|w\|_2 \leq D_w, \sup_{v \in \mathcal{V}} \|v\|_2 \leq D_v$ and $\delta \in (0, 1)$. Let $\tilde{\Delta}_T = |F(\bar{w}_T, \bar{v}_T) - F(w^*, v^*)|$.

(a) If Assumption 2.1 holds, then with probability at least $1 - \delta$

$$\tilde{\Delta}_T = \mathcal{O}\left(\eta \log n \log(1/\delta)(\sqrt{T} + Tn^{-1} + \log(1/\delta))\right)$$

$$+ \mathcal{O}\left(n^{-\frac{1}{2}} \log^2(1/\delta)\right) + \mathcal{O}\left((T\eta)^{-1} + T^{-\frac{1}{2}} \log(1/\delta)\right).$$

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

$$\tilde{\Delta}_T = \mathcal{O}(n^{-1/2} \log n \log^2(1/\delta)). \tag{2.38}$$

(b) If Assumptions 2.1, 2.2 hold, then the following inequality holds with probability at least
\[ \Delta_T = \mathcal{O} \left( \eta \log n \log(1/\delta) \exp \left( L^2 T \eta^2 / 2 \right) \left( T n^{-1} + \log(1/\delta) \right) \right) + n^{-\frac{1}{2}} \log^{\frac{3}{2}}(1/\delta) + (T \eta)^{-1} + T^{-\frac{1}{2}} \log(1/\delta). \]

In particular, we can choose \( T \approx n \) and \( \eta \approx T^{-1/2} \) to derive (2.38) with probability at least \( 1 - \delta \).

**Proof**: We use the error decomposition

\[
F(\bar{w}_T, \bar{v}_T) - F(w^*, v^*) = F(\bar{w}_T, \bar{v}_T) - F_S(\bar{w}_T, \bar{v}_T) + F_S(\bar{w}_T, \bar{v}_T) - F_S(w^*, \bar{v}_T) + F_S(w^*, \bar{v}_T) - F_S(w^*, v_T) + F_S(w^*, v_T) - F(w^*, v^*). \tag{2.39}
\]

We first prove Part (a). According to Assumption 2.1 and Part (c) of Theorem 2.2, we know that SGDA is \( \epsilon \)-uniformly stable with probability at least \( 1 - \delta/4 \), where

\[
\epsilon = \mathcal{O} \left( \eta \left( \sqrt{T} + T n^{-1} + \log(1/\delta) \right) \right).
\]

This together with Part (e) of Theorem 2.1 implies the following inequality with probability at least \( 1 - \delta/2 \)

\[
F(\bar{w}_T, \bar{v}_T) - F_S(\bar{w}_T, \bar{v}_T) = \mathcal{O} \left( \eta \log n \log(1/\delta) \left( \sqrt{T} + T n^{-1} + \log(1/\delta) \right) \right) + \mathcal{O}(n^{-\frac{1}{2}} \log^{\frac{1}{2}}(1/\delta)). \tag{2.40}
\]

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

\[
F_S(w^*, \bar{v}_T) - F(w^*, \bar{v}_T) = \mathcal{O} \left( \eta \log n \log(1/\delta) \left( \sqrt{T} + T n^{-1} + \log(1/\delta) \right) \right) + \mathcal{O}(n^{-\frac{1}{2}} \log^{\frac{1}{2}}(1/\delta)). \tag{2.41}
\]
According to Lemma 2.9, the following inequality holds with probability at least $1 - \delta/4$

$$F_S(\bar{w}_T, \bar{v}_T) - F_S[w^*, \bar{v}_T] \leq \sup_v F_S(\bar{w}_T, v) - \inf_w F_S(w, \bar{v}_T)$$

$$= \mathcal{O} \left( \eta + (T\eta)^{-1} + T^{-\frac{1}{2}} \log(1/\delta) \right). \quad (2.42)$$

According to the definition of $(w^*, v^*)$, we know $F(w^*, \bar{v}_T) \leq F(w^*, v^*)$. We can plug this inequality and (2.40), (2.41), (2.42) back into (2.39), and derive the following inequality with probability at least $1 - \delta/2$

$$F(\bar{w}_T, \bar{v}_T) - F(w^*, v^*) = \mathcal{O} \left( \eta \log n \log(1/\delta)(\sqrt{T} + Tn^{-1} + \log(1/\delta)) \right)$$

$$+ \mathcal{O}(n^{-\frac{1}{2}} \log^\frac{3}{2}(1/\delta)) + \mathcal{O} \left( (T\eta)^{-1} + T^{-\frac{1}{2}} \log(1/\delta) \right).$$

Analyzing in a similar way but using the error decomposition

$$F(w^*, v^*) - F(\bar{w}_T, \bar{v}_T) = F(w^*, v^*) - F(\bar{w}_T, v^*) + F(\bar{w}_T, v^*) - F_S(\bar{w}_T, v^*)$$

$$+ F_S(\bar{w}_T, v^*) - F_S(\bar{w}_T, \bar{v}_T) + F_S(\bar{w}_T, \bar{v}_T) - F(\bar{w}_T, \bar{v}_T),$$

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

$$F(w^*, v^*) - F(\bar{w}_T, \bar{v}_T) = \mathcal{O} \left( \eta \log n \log(1/\delta)(\sqrt{T} + Tn^{-1} + \log(1/\delta)) \right)$$

$$+ \mathcal{O}(n^{-\frac{1}{2}} \log^\frac{3}{2}(1/\delta)) + \mathcal{O} \left( (T\eta)^{-1} + T^{-\frac{1}{2}} \log(1/\delta) \right).$$

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

Part (b) can be derived similarly excepting using the stability bounds in Part (d) of Theorem 2.2. We omit the proof for brevity.

2.4 Nonconvex-Nonconcave Objectives

In this section, we extend our analysis to nonconvex-nonconcave minimax learning problems.
2.4.1 Stability and Generalization of SGDA

Lemma 2.11 Assume $|f(\cdot, \cdot, z)| \leq 1$ for any $z$ and let Assumption 2.1 hold. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z_1, \ldots, z_{n-1}, z'_n\}$. Let $\{w_t, v_t\}$ and $\{w'_t, v'_t\}$ be the sequence produced by (2.8) w.r.t. $S$ and $S'$, respectively. Denote

$$\Delta_t = \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2. \quad (2.43)$$

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

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

Proof: According to Assumption 2.1 we know

$$f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z) \leq G\sqrt{2}\Delta_T. \quad (2.44)$$

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

$$\mathbb{E}[f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z)] = \mathbb{P}[\mathcal{E}]\mathbb{E}[f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z) | \mathcal{E}]$$

$$+ \mathbb{P}[\mathcal{E}^c]\mathbb{E}[f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z) | \mathcal{E}^c]$$

$$\leq \sqrt{2}G\mathbb{E}[\Delta_T | \mathcal{E}] + 4\mathbb{P}[\mathcal{E}^c],$$

where in the last step we have used (2.44) and the condition $|f(\cdot, \cdot, z)| \leq 1$. Using the union bound on the outcome $i_t = n$ we obtain that

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

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

Lemma 2.12 shows the monotonity of the gradient for weakly-convex-weakly-concave functions. Its proof is well known in the literature [Liu et al., 2021, Rockafellar, 1976].
Lemma 2.12 Let $f$ be a $\rho$-weakly-convex-weakly-concave function. Then

$$\langle (\mathbf{w} - \mathbf{w}'), (\nabla_w f(\mathbf{w}, \mathbf{v}) - \nabla_w f(\mathbf{w}', \mathbf{v}')) \rangle \geq -\rho \| (\mathbf{w} - \mathbf{w}') \|^2.$$ \hfill (2.45)

We first study the generalization bounds of SGDA for WC-WC problems.

Theorem 2.6 (Weak generalization bound) Let $\{\mathbf{w}_t, \mathbf{v}_t\}$ be produced by (2.8) with $T$ iterations. Assume for all $z$, the function $(\mathbf{w}, \mathbf{v}) \mapsto f(\mathbf{w}, \mathbf{v}; z)$ is $\rho$-WC-WC and $|f(\cdot, \cdot; z)| \leq 1$. If Assumption 2.7 holds and $\eta_t = c/t$, then the weak PD generalization error of SGDA is bounded by

$$\mathcal{O}\left( (1 + \sqrt{T/n}) T^{\rho} \right)^{\frac{2}{2\rho+3}} \left( \frac{1}{n} \right)^{\frac{2\rho+1}{2\rho+3}}.$$ \hfill (2.46)

Proof: Note that the projection step is nonexpansive. We consider two cases at the $t$-th iteration. If $i_t \neq n$, then it follows from Lemma 2.12 and the Lipschitz continuity of $f$ that

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

$$= \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|^2_2 + \eta_t^2 \left\| \begin{pmatrix} \nabla_w f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t}) - \nabla_w f(\mathbf{w}'_t, \mathbf{v}'_t; \mathbf{z}_{i_t}) \\ \nabla_v f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t}) - \nabla_v f(\mathbf{w}'_t, \mathbf{v}'_t; \mathbf{z}_{i_t}) \end{pmatrix} \right\|^2_2$$

$$- 2\eta_t \left\langle \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix}, \begin{pmatrix} \nabla_w f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t}) - \nabla_w f(\mathbf{w}'_t, \mathbf{v}'_t; \mathbf{z}_{i_t}) \\ \nabla_v f(\mathbf{w}_t, \mathbf{v}_t; \mathbf{z}_{i_t}) - \nabla_v f(\mathbf{w}'_t, \mathbf{v}'_t; \mathbf{z}_{i_t}) \end{pmatrix} \right\rangle$$

$$\leq (1 + 2\eta_t \rho) \left\| \begin{pmatrix} \mathbf{w}_t - \mathbf{w}'_t \\ \mathbf{v}_t - \mathbf{v}'_t \end{pmatrix} \right\|^2_2 + 8G^2 \eta_t^2 \eta_t^2.$$ \hfill (2.46)

If $i_t = n$, then it follows from the elementary inequality $(a + b)^2 \leq (1 + p)a^2 + (1 + 1/p)b^2$
that
\[
\left\| \begin{pmatrix} w_{t+1} - w_{t+1}' \\ v_{t+1} - v_{t+1}' \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} w_t - \eta_t \nabla_w f(w_t, v_t; z_n) - w_t' + \eta_t \nabla_w f(w_t', v_t'; z_n') \\ v_t + \eta_t \nabla_v f(w_t, v_t; z_n) - v_t' - \eta_t \nabla_v f(w_t', v_t'; z_n') \end{pmatrix} \right\|_2^2 \\
\leq (1 + p) \left\| \begin{pmatrix} w_t - w_t' \\ v_t - v_t' \end{pmatrix} \right\|_2^2 + (1 + 1/p) \eta_t^2 \left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_n) - \nabla_w f(w_t', v_t'; z_n') \\ \nabla_v f(w_t, v_t; z_n) - \nabla_v f(w_t', v_t'; z_n') \end{pmatrix} \right\|_2^2. \tag{2.47}
\]

Note that the event \( i_t \neq n \) happens with probability \( 1 - 1/n \) and the event \( i_t = n \) happens with probability \( 1/n \). Therefore, we know
\[
E_{i_t} \left[ \left\| \begin{pmatrix} w_{t+1} - w_{t+1}' \\ v_{t+1} - v_{t+1}' \end{pmatrix} \right\|_2^2 \right] \leq \frac{n - 1}{n} \left( 1 + 2 \eta_t \rho \right) \left\| \begin{pmatrix} w_t - w_t' \\ v_t - v_t' \end{pmatrix} \right\|_2^2 + 8 \eta_t^2 \left( 1 + 2 \eta_t \rho \right) \left( 1 + 2 \eta_t \rho + p/n \right)
\]
\[
\leq (1 + 2 \eta_t \rho + p/n) \left\| \begin{pmatrix} w_t - w_t' \\ v_t - v_t' \end{pmatrix} \right\|_2^2 + 8 \eta_t^2 G^2 (1 + 1/(np)).
\]

Let \( t_0 \in \mathbb{N} \) and \( E \) be defined as in the proof of Lemma 2.11. We apply the above equation recursively from \( t = t_0 + 1 \) to \( T \), then
\[
E_A \left[ \left\| \begin{pmatrix} w_T - w_T' \\ v_T - v_T' \end{pmatrix} \right\|_2^2 \right] \leq 8 G^2 (1 + 1/(np)) \sum_{t=t_0+1}^T \eta_t^2 \prod_{k=t+1}^T \left( 1 + 2 \eta_k \rho + p/n \right)
\]

By the elementary inequality \( 1 + a \leq \exp(a) \) and \( \eta_t = \frac{c}{t} \), we further derive
\[
E_A \left[ \left\| \begin{pmatrix} w_{t+1} - w_{t+1}' \\ v_{t+1} - v_{t+1}' \end{pmatrix} \right\|_2^2 \right] \leq 8 G^2 (1 + 1/(np)) \sum_{t=t_0+1}^T \frac{c^2}{t^2} \prod_{k=t+1}^T \exp \left( \frac{2c \rho}{k} + \frac{p}{n} \right)
\]
\[
\leq 8 G^2 (1 + 1/(np)) \sum_{t=t_0+1}^T \frac{c^2}{t^2} \exp \left( \sum_{k=t+1}^T \frac{2c \rho}{k} + \frac{pT}{n} \right).
\]

55
By taking $p = n/T$ in the above inequality, we further derive

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

Combining the above inequality and Lemma 2.11 together, we obtain

$$
\mathbb{E}_A \left[ f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z) \right] \\
\leq \frac{4t_0}{n} + \frac{4\sqrt{e}G^2}{\sqrt{2c\rho + 1}} \left(1 + \frac{\sqrt{T}}{n}\right) \left( \frac{T}{t_0} \right)^{c\rho} \frac{1}{\sqrt{t_0}}. \quad (2.48)
$$

The right hand side is approximately minimized when

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

Plugging it into the Eq. (2.48) we have (for simplicity we assume the above $t_0$ is an integer)

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

Since the above bound holds for all $z, S, S'$ and $w', v'$, we immediately get the same upper bound on the weak stability. Finally the theorem holds by calling Theorem 2.1, Part (a). □

**Remark 2.5** When $T = O(n^2)$, our weak PD generalization error bound is of the order $O\left(n^{-\frac{2c\rho + 1}{2c\rho + 3}} T^{-\frac{2c\rho}{2c\rho + 3}} \right)$. This is the first generalization bound of SGDA for nonsmooth and nonconvex-nonconcave objectives. Farnia and Ozdaglar [2021] also studied generalization under nonconvex-nonconcave setting but required the objectives to be smooth, which is relaxed to a milder WC-WC assumption here. Our analysis readily applies to stochastic gradient
descent (SGD) with nonsmooth weakly-convex functions, which has not been studied in the literature.

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

Lemma 2.13 Let $c_t \in (0, 1]$ and let $X_1, \ldots, X_T$ be independent Bernoulli random variables with the success rate of $X_t$ being $p_t \in [0, 1]$. Denote $s = \sum_{t=1}^{T} c_t p_t$. Then, for all $a > 0$,

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

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

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

Theorem 2.7 Let $\{w_t, v_t\}$ be the sequence produced by (2.8) with $\eta_t \leq \frac{c}{t}$ for some $c > 0$. Assume Assumptions 2.1, 2.2 hold and $|f(\cdot, \cdot; z)| \leq 1$. For any $\delta \in (0, 1)$, if $c \leq \frac{1}{(n \log(2/\delta) - 1)L}$, then with probability at least $1 - \delta$ we have

$$|F(w_T, v_T) - F_S(w_T, v_T)| = O\left(T^{cL} \log(n) \log^{3/2}(1/\delta)n^{-1/2} + n^{-1/2}\log^{1/2}(1/\delta)\right).$$

Proof: Let $S' = \{z_1, \ldots, z_{n-1}, z'_n\}$ and $\{w'_t, v'_t\}$ be the sequence produced by (2.8) w.r.t. $S'$. If $i_t \neq n$, it follows from the $L$-smoothness of $f$ that

$$\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2 \leq \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2 + \eta_t \left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w f(w'_t, v'_t; z_{i_t}) \\ \nabla_v f(w_t, v_t; z_{i_t}) - \nabla_v f(w'_t, v'_t; z_{i_t}) \end{pmatrix} \right\|_2 \leq (1 + L\eta_t) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2.$$
If $i_t = n$, we have
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2 \leq \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2 + 4\eta_t G.
\]

We can combine the above two inequalities together and get
\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2 \leq (1 + L\eta_t) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2 + 4G\eta_t \mathbb{I}_{[i_t = n]}.
\]

We apply the above inequality recursively from $t = 1$ to $T$ and get
\[
\left\| \begin{pmatrix} w_T - w'_T \\ v_T - v'_T \end{pmatrix} \right\|_2 \leq 4G \sum_{t=1}^T \eta_t \mathbb{I}_{[i_t = n]} \prod_{k=t+1}^T \left( 1 + L\eta_k \right).
\]

By the elementary inequality $1 + a \leq \exp(a)$ and $\eta_t \leq \frac{\xi}{t}$, we further derive
\[
\left\| \begin{pmatrix} w_T - w'_T \\ v_T - v'_T \end{pmatrix} \right\|_2 \leq 4cG \sum_{t=1}^T \frac{\mathbb{I}_{[i_t = n]}}{t} \prod_{k=t+1}^T \exp \left( \frac{cL}{k} \right) = 4cG \sum_{t=1}^T \frac{\mathbb{I}_{[i_t = n]}}{t} \exp \left( \sum_{k=t+1}^T \frac{cL}{k} \right) \leq 4cG \sum_{t=1}^T \frac{\mathbb{I}_{[i_t = n]}}{t} \exp \left( cL \log \left( \frac{T}{t} \right) \right) \leq 4cGT cL \sum_{t=1}^T \frac{\mathbb{I}_{[i_t = n]}}{t^{cL+1}}.
\]

By Lemma 2.13 for any $\delta > 0$ such that $\log(2/\delta) < \sum_{t=1}^T \frac{1}{t^{cL+1}}$, with probability at least $1 - \delta/2$ we have
\[
\left\| \begin{pmatrix} w_T - w'_T \\ v_T - v'_T \end{pmatrix} \right\|_2 \leq 4cG T cL \left( \sum_{t=1}^T \frac{1}{t^{cL+1}} + (e - 1) \sqrt{\log(1/\delta) \sum_{t=1}^T \frac{1}{t^{cL+1}}} \right). \tag{2.49}
\]

Note that
\[
\sum_{t=1}^T \frac{1}{t^{cL+1}} \leq 1 + \int_{t=1}^T \frac{dt}{t^{cL+1}} \leq 1 + \frac{1}{cL}.
\]
Plugging the above bound into Equation (2.49), we know with probability at least $1 - \delta/2$
\[
\left\| \begin{pmatrix} w_T - w'_T \\ v_T - v'_T \end{pmatrix} \right\|_2 \leq 4cGT^{cL} \left( \frac{cL + 1}{cLn} + (e - 1)\sqrt{\frac{(cL + 1) \log(1/\delta)}{cLn}} \right).
\]
By the Lipschitz continuity of $f$, the above equation implies SGDA is $\epsilon$-uniformly stable with probability at least $1 - \delta/2$ and
\[
\epsilon = \mathcal{O}\left(T^{cL} \sqrt{\log(1/\delta)n^{-1/2}}\right).
\]
This together with Part (e) of Theorem 2.1 implies the following inequality with probability at least $1 - \delta$
\[
|F(w, v) - F_S(w_T, v_T)| = \mathcal{O}\left(T^{cL} \log(n) \log^{3/2}(1/\delta)n^{-1/2} + n^{-1/2} \log^{1/2}(1/\delta)\right).
\]

We further consider a variant of weak-convexity-weak-concavity.

**Theorem 2.8 (Weak generalization bound)** Let $\{w_t, v_t\}$ be produced by (2.8) with $T$ iterations. Let Assumptions 2.1, 2.2 hold. Assume there are non-negative numbers $\{\rho_t\}_{t \in \mathbb{N}}$ such that the following inequality holds a.s.
\[
\left\langle \begin{pmatrix} w_t - w \\ v_t - v \end{pmatrix}, \begin{pmatrix} \nabla_w F_S(w_t, v_t) - \nabla_w F_S(w, v) \\ \nabla_v F_S(w, v) - \nabla_v F_S(w_t, v_t) \end{pmatrix} \right\rangle \geq -\rho_t \left\| \begin{pmatrix} w_t - w \\ v_t - v \end{pmatrix} \right\|_2^2, \ \forall w \in \mathcal{W}, v \in \mathcal{V}.
\]
(2.50)

Then the weak PD generalization error of SGDA with $T$ iterations can be bounded by
\[
\mathcal{O}\left(n^{-1} \sum_{t=1}^{T} \left( \eta_t^2 + \frac{1}{n} \right) \exp \left( \sum_{k=t+1}^{T} \left( 2\rho_k \eta_k + (L^2 + 1)\eta_k^2 \right) \right) \right)^{1/2}.
\]

**Proof:** Without loss of generality, we assume $z_i = z'_i$ for $i \in [n - 1]$. If $i_t \neq n$, then it
follows from Assumption 2.2 that

\[
\left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_{it}) - \nabla_w f(w'_t, v'_t; z'_{it}) \\ \nabla_v f(w_t, v_t; z_{it}) - \nabla_v f(w'_t, v'_t; z'_{it}) \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_{it}) - \nabla_w f(w'_t, v'_t; z'_{it}) \\ \nabla_v f(w_t, v_t; z_{it}) - \nabla_v f(w'_t, v'_t; z'_{it}) \end{pmatrix} \right\|_2^2 \\
\leq L^2 \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2
\]

If \( i_t = n \), then it follows from Assumption 2.1 that

\[
\left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_{it}) - \nabla_w f(w'_t, v'_t; z'_{it}) \\ \nabla_v f(w_t, v_t; z_{it}) - \nabla_v f(w'_t, v'_t; z'_{it}) \end{pmatrix} \right\|_2^2 \leq 8G^2.
\]

Therefore, we have

\[
\mathbb{E}_{i_t} \left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_{it}) - \nabla_w f(w'_t, v'_t; z'_{it}) \\ \nabla_v f(w_t, v_t; z_{it}) - \nabla_v f(w'_t, v'_t; z'_{it}) \end{pmatrix} \right\|_2^2 \leq \frac{(n-1)L^2}{n} \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + \frac{8G^2}{n}. \tag{2.51}
\]

According to (2.8), we know

\[
\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \leq \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + \eta_t^2 \left\| \begin{pmatrix} \nabla_w f(w_t, v_t; z_{it}) - \nabla_w f(w'_t, v'_t; z'_{it}) \\ \nabla_v f(w_t, v_t; z_{it}) - \nabla_v f(w'_t, v'_t; z'_{it}) \end{pmatrix} \right\|_2^2 \\
- 2\eta_t \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w f(w_t, v_t; z_{it}) - \nabla_w f(w'_t, v'_t; z'_{it}) \\ \nabla_v f(w_t, v_t; z_{it}) - \nabla_v f(w'_t, v'_t; z'_{it}) \end{pmatrix} \right\rangle.
\]
Taking a conditional expectation w.r.t. $i_t$ gives

\[
\mathbb{E}_{i_t} \left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|^2_2 \\
\leq \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2_2 + L^2 \eta_t^2 \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2_2 + \frac{8G^2 \eta_t^2}{n} \\
- 2\eta_t \mathbb{E}_{i_t} \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w f(w_t, v_t, z_{i_t}) - \nabla_w f(w'_t, v'_t, z'_{i_t}) \\ \nabla_v f(w'_t, v'_t, z'_{i_t}) - \nabla_v f(w_t, v'_t, z_{i_t}) \end{pmatrix} \right\rangle \\
= \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2_2 + L^2 \eta_t^2 \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2_2 + \frac{8G^2 \eta_t^2}{n} \\
- 2\eta_t \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w F_S(w_t, v_t) - \nabla_w F'_S(w'_t, v'_t) \\ \nabla_v F'_S(w'_t, v'_t) - \nabla_v F_S(w_t, v_t) \end{pmatrix} \right\rangle,
\]

where we have used (2.51) in the first step and used the fact

\[
\mathbb{E}_{i_t} \nabla f(w, v, z_{i_t}) = \nabla F_S(w, v), \quad \mathbb{E}_{i_t} \nabla f(w, v, z'_{i_t}) = \nabla F'_S(w, v)
\]
in the second step. According to (2.50), we know

\[
\left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w F_S(w_t, v_t) - \nabla_w F'_S(w'_t, v'_t) \\ \nabla_v F'_S(w'_t, v'_t) - \nabla_v F_S(w_t, v_t) \end{pmatrix} \right\rangle \\
= \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w F_S(w_t, v_t) - \nabla_w F_S(w'_t, v'_t) \\ \nabla_v F'_S(w'_t, v'_t) - \nabla_v F_S(w_t, v_t) \end{pmatrix} \right\rangle \\
+ \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w F'_S(w'_t, v'_t) - \nabla_w F'_S(w'_t, v'_t) \\ \nabla_v F'_S(w'_t, v'_t) - \nabla_v F'_S(w'_t, v'_t) \end{pmatrix} \right\rangle \\
\geq -\rho_t \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|^2_2 + \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w F'_S(w'_t, v'_t) - \nabla_w F'_S(w'_t, v'_t) \\ \nabla_v F'_S(w'_t, v'_t) - \nabla_v F'_S(w'_t, v'_t) \end{pmatrix} \right\rangle.
\]
It follows from Assumption 2.1 that

\[
\begin{align*}
&\left\langle (w_t - w'_t), \left( \nabla_w F_S(w'_t, v'_t) - \nabla_w F_S(w'^t, v'^t) \right) \right\rangle \\
&\quad = \frac{1}{n} \left\langle \left( w_t - w'_t \right), \left( \nabla_w f(w'_t, v'_t; z_n) - \nabla_w f(w'_t, v'_t; z'_n) \right) \right\rangle \\
&\quad \geq - \frac{1}{n} \left\| \left( w_t - w'_t \right) \right\|_2 \left\| \left( \nabla_w f(w'_t, v'_t; z_n) - \nabla_w f(w'_t, v'_t; z'_n) \right) \right\|_2 \\
&\quad \geq - \frac{2\sqrt{2}G}{n} \left\| \left( w_t - w'_t \right) \right\|_2.
\end{align*}
\]

We can combine the above three inequalities together and derive

\[
\mathbb{E}_{\eta_t} \left\| \left( w_{t+1} - w'_{t+1} \right) \right\|_2^2 \leq \left( 1 + 2\rho_t \eta_t + L^2 \eta_t^2 \right) \left\| \left( w_t - w'_t \right) \right\|_2^2 + \frac{8\eta_t^2 G^2}{n} \\
+ \frac{4\sqrt{2}G \eta_t}{n} \left\| \left( w_t - w'_t \right) \right\|_2 \\
\leq \left( 1 + 2\rho_t \eta_t + L^2 \eta_t^2 \right) \left\| \left( w_t - w'_t \right) \right\|_2^2 + \frac{8\eta_t^2 G^2}{n} \\
+ \eta_t^2 \left\| \left( w_t - w'_t \right) \right\|_2^2 + \frac{8G^2}{n^2}.
\]

Applying the above inequality recursively, we get

\[
\mathbb{E}_A \left\| \left( w_{t+1} - w'_{t+1} \right) \right\|_2^2 \leq \frac{8G^2}{n} \sum_{j=1}^t \left( \eta_t^2 + \frac{1}{n} \right) \prod_{k=j+1}^t \left( 1 + 2\rho_k \eta_k + L^2 \eta_k^2 + \eta_k^2 \right).
\]

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

\[
\mathbb{E}_A \left\| \left( w_{t+1} - w'_{t+1} \right) \right\|_2^2 \leq \frac{8G^2}{n} \sum_{j=1}^t \left( \eta_t^2 + \frac{1}{n} \right) \exp \left( \sum_{k=j+1}^t \left( 2\rho_k \eta_k + (L^2 + 1) \eta_k^2 \right) \right).
\]

62
It then follows from the Jensen’s inequality that

$$E_A \left\| \left( \begin{array}{c} w_{t+1} - w_{t+1}' \\ v_{t+1} - v_{t+1}' \end{array} \right) \right\|_2^2 \leq \frac{2\sqrt{2}G}{\sqrt{n}} \left( \sum_{j=1}^{t} \left( \eta_j^2 + \frac{1}{n} \right) \exp \left( \sum_{k=j+1}^{t} \left( 2\rho_k \eta_k + (L^2 + 1) \eta_k^2 \right) \right) \right)^{\frac{1}{2}}.$$

The stated bound then follows from Part (a) of Theorem 2.1 and Assumption 2.1.

Eq. (2.50) allows the empirical objective $F_S$ to have varying weak-convexity-weak-concavity at different iterates encountered by the algorithm. This is motivated by the observation that the nonconvex-nonconcave function can have approximate convexity-concavity around a saddle point. For these problems, we can expect the weak-convexity-weak-concavity parameter $\rho_t$ to decrease along the optimization process [Sagun et al., 2017, Yuan et al., 2019].

Remark 2.6 If $F_S$ is convex-concave, then $\rho_t = 0$ and we can take $\eta_t \asymp 1/\sqrt{T}$ to show that SGDA with $T$ iterations enjoys the generalization bound $O(1/\sqrt{n} + \sqrt{T}/n)$. This extends Theorem 2.3 since we only require the convexity-concavity of $F_S$ here instead of $f(\cdot, \cdot; z)$ for all $z$ in Theorem 2.3. If $\rho_t = O(t^{-\alpha})$ ($\alpha \in (0, 1)$), then we can take $\eta_t \asymp t^{\min(\alpha - 1, -\frac{1}{2})}/\log T$ (note $\sum_{t=1}^{T} \eta_t^2 = O(1), \sum_{t=1}^{T} \eta_t \rho_t = O(1)$) to show that SGDA with $T$ iterations enjoys the weak PD generalization bound $O(1/\sqrt{n} + \sqrt{T}/n)$. As compared to Theorem 2.6, the assumption (2.50) allows us to use much larger step sizes ($O(t^{-\beta}), \beta \in (0, 1)$ vs $O(t^{-1})$). This larger step size allows for a better trade-off between generalization and optimization. We note that a recent work [Richards and Rabbat, 2021] considered gradient descent under an assumption similar to (2.50), and developed interesting generalization bounds for $\eta_t = O(t^{-\beta})$ ($\beta \in (0, 1)$). However, their discussions do not apply to the important SGD and require an additional assumption on the Lipschitz continuity of Hessian matrix which may be restrictive. It is direct to extend Theorem 2.8 to SGD for learning with weakly-convex functions for relaxing the step size under Eq. (2.50). Therefore, our stability analysis even gives novel results in the standard nonconvex learning setting. We introduce a novel technique in achieving this improvement. Specifically, let $\delta_i := \|w_t - w_i'\|_2^2 + \|v_t - v_i'\|_2^2$, where $(w_t, v_t), (w_i', v_i')$ are SGDA iterates for neighboring datasets $S$ and $S'$. For the stability bounds in Section 2.3.1, we first handle $\delta_{i+1}$ according to different realizations of $i_t$ and then consider the expectation w.r.t. $i_t$. While for $\rho$-WC-WC problems, we first take expectation w.r.t. $i_t$ and then show how $E_{i_t}[\delta_{i+1}]$ would change along the iterations.
2.4.2 Stability and Generalization of AGDA and Beyond

In this section, we give the proof on the stability and generalization bounds of AGDA for nonconvex-nonconcave functions. The next lemma is similar to Lemma 2.11, which shows AGDA typically runs several iterations before encountering the different example between \( S \) and \( S' \).

**Lemma 2.14** Assume \(|f(\cdot, \cdot, z)| \leq 1\) for any \( z \) and let Assumption 2.1 hold. Let \( S = \{z_1, \ldots, z_n\} \) and \( S' = \{z_1, \ldots, z_{n-1}, z'_n\} \). Let \( \{w_t, v_t\} \) and \( \{w'_t, v'_t\} \) be the sequence produced by (2.54) w.r.t. \( S \) and \( S' \), respectively. Denote

\[
\Delta_t = \|w_t - w'_t\|_2 + \|v_t - v'_t\|_2.
\]  

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

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

**Proof:** Let \( \mathcal{E} \) denote the event that \( \Delta_{t_0} = 0 \). Then we have

\[
\mathbb{E}[f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z)] = \mathbb{P}[\mathcal{E}]\mathbb{E}[f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z)|\mathcal{E}]
\]

\[
+ \mathbb{P}[\mathcal{E}^c]\mathbb{E}[f(w_T, v'; z) - f(w'_T, v'; z) + f(w', v_T; z) - f(w', v'_T; z)|\mathcal{E}^c]
\]

\[
\leq G\mathbb{E}[\Delta_T|\mathcal{E}] + 4\mathbb{P}[\mathcal{E}^c],
\]

where we have used (2.44) and the assumption \(|f(\cdot, \cdot, z)| \leq 1\). Using the union bound on the outcome \( i_t = n \) and \( j_t = n \) we obtain that

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

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

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

\[
\begin{align*}
\mathbf{w}_{t+1} &= \Pi_{W}(\mathbf{w}_t - \eta_{w,t} \nabla_w f(\mathbf{w}_t, \mathbf{v}_t; z_{i_t})) , \\
\mathbf{v}_{t+1} &= \Pi_{V}(\mathbf{v}_t + \eta_{v,t} \nabla_v f(\mathbf{w}_{t+1}, \mathbf{v}_t; z_{j_t})) .
\end{align*}
\] (2.54)

This algorithm differs from SGDA in two aspects. First, it randomly selects two examples to update \(\mathbf{w}\) and \(\mathbf{v}\) per iteration. Second, it uses the updated \(\mathbf{w}_{t+1}\) when updating \(\mathbf{v}_{t+1}\).

Theorem 2.9 provides generalization bounds for AGDA.

**Theorem 2.9 (Weak generalization bounds)** Let \(\{\mathbf{w}_t, \mathbf{v}_t\}\) be the sequence produced by (2.54). If Assumptions 2.1, 2.2 hold and \(\eta_{w,t} + \eta_{v,t} \leq \frac{c}{t}\) for some \(c > 0\), then the weak PD generalization error can be upper bounded by \(O(n^{-1}T^{\frac{d}{d+1}})\).

**Proof:** Since \(z_{i_t}\) and \(z_{j_t}\) are i.i.d, we can analyze the update of \(\mathbf{w}\) and \(\mathbf{v}\) separately. Note that the projection step is nonexpansive. We consider two cases at the \(t\)-th iteration.

If \(i_t \neq n\), then it follows from Assumption 2.2 that

\[
\begin{align*}
\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \\
\leq \|\mathbf{w}_t - \eta_{w,t} \nabla_w f(\mathbf{w}_t, \mathbf{v}_t, z_{i_t}) - \mathbf{w}'_t + \eta_{w,t} \nabla_w f(\mathbf{w}'_t, \mathbf{v}_t, z_{i_t})\|_2 \\
\leq \|\mathbf{w}_t - \eta_{w,t} \nabla_w f(\mathbf{w}_t, \mathbf{v}_t, z_{i_t}) - \mathbf{w}'_t + \eta_{w,t} \nabla_w f(\mathbf{w}'_t, \mathbf{v}_t, z_{i_t})\|_2 \\
+ \|\eta_{w,t} \nabla_w f(\mathbf{w}'_t, \mathbf{v}_t, z_{i_t}) - \eta_{w,t} \nabla_w f(\mathbf{w}'_t, \mathbf{v}_t, z_{i_t})\|_2 \\
\leq (1 + \eta_{w,t})\|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \eta_{w,t}\|\mathbf{v}_t - \mathbf{v}'_t\|_2 .
\end{align*}
\]

If \(i_t = n\), then it follows from Assumption 2.1 that

\[
\begin{align*}
\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \\ \leq \|\mathbf{w}_t - \eta_{w,t} \nabla_w f(\mathbf{w}_t, \mathbf{v}_t, z_{i_t}) - \mathbf{w}'_t + \eta_{w,t} \nabla_w f(\mathbf{w}'_t, \mathbf{v}_t, z_{i_t})\|_2 \\
\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2G\eta_{w,t} .
\end{align*}
\]
According to the distribution of \( i_t \), we have
\[
\mathbb{E}_A[\|w_{t+1} - w'_{t+1}\|_2] \leq \frac{n-1}{n} \mathbb{E}_A\left[(1 + \eta_{w,t}L)\|w_t - w'_t\|_2 + L\eta_{w,t}\|v_t - v'_t\|_2\right]
\]
\[
+ \frac{1}{n}(\|w_t - w'_t\|_2 + 2\eta_{w,t}G)
\]
\[
\leq (1 + \eta_{w,t}L)\mathbb{E}_A[\|w_t - w'_t\|_2] + L\eta_{w,t}\mathbb{E}_A[\|v_t - v'_t\|_2] + \frac{2\eta_{w,t}G}{n}. \quad (2.55)
\]

Similarly, for \( v \) we also have
\[
\mathbb{E}_A[\|v_{t+1} - v'_{t+1}\|_2] \leq (1 + \eta_{v,t}L)\mathbb{E}_A[\|v_t - v'_t\|_2] + L\eta_{v,t}\mathbb{E}_A[\|w_t - w'_t\|_2] + \frac{2\eta_{v,t}G}{n}. \quad (2.56)
\]

Combining (2.55) and (2.56) we have
\[
\mathbb{E}_A[\|w_{t+1} - w'_{t+1}\|_2 + \|v_{t+1} - v'_{t+1}\|_2] \leq (1 + (\eta_{w,t} + \eta_{v,t})L)\mathbb{E}_A[\|w_t - w'_t\|_2 + \|v_t - v'_t\|_2]
\]
\[
+ \frac{2(\eta_{w,t} + \eta_{v,t})G}{n}.
\]

Recalling the event \( E \) that \( \Delta_{t_0} = 0 \), we apply the above equation recursively from \( t = t_0 + 1 \) to \( T \), then
\[
\mathbb{E}_A[\|w_{t+1} - w'_{t+1}\|_2 + \|v_{t+1} - v'_{t+1}\|_2 | \Delta_{t_0} = 0]
\]
\[
\leq \frac{2G}{n} \sum_{t=t_0+1}^T (\eta_{w,t} + \eta_{v,t}) \prod_{k=t_0}^T (1 + (\eta_{w,k} + \eta_{v,k})L).
\]

By the elementary inequality \( 1 + x \leq \exp(x) \) and \( \eta_{w,t} + \eta_{v,t} \leq \frac{\xi}{t} \), we have
\[
\mathbb{E}_A[\|w_{t+1} - w'_{t+1}\|_2 + \|v_{t+1} - v'_{t+1}\|_2 | \Delta_{t_0} = 0]
\]
\[
\leq \frac{2cG}{n} \sum_{t=t_0+1}^T \frac{1}{t} \prod_{k=t_0+1}^T \exp\left(\frac{cL}{k}\right) = \frac{2cG}{n} \sum_{t=t_0+1}^T \frac{1}{t} \exp\left(\sum_{k=t_0+1}^T \frac{cL}{k}\right)
\]
\[
\leq \frac{2cG}{n} \sum_{t=t_0+1}^T \frac{1}{t} \exp\left(cL \log\left(\frac{T}{t}\right)\right) \leq \frac{2cGT^cL}{n} \sum_{t=t_0+1}^T \frac{1}{t} \leq \frac{2G}{Ln} \left(\frac{T}{t_0}\right)^{cL}.
\]
By Lemma 2.14 we have

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

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

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

Plugging it into Eq. (2.57) we have (for simplicity we assume the above $t_0$ is an integer)

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

Since the above bound holds for all $z, S, S'$ and $w', v'$, we immediately get the same upper bound on the weak stability. Finally the theorem holds by calling Theorem 2.1, Part (a). □

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

**Assumption 2.3** Assume $F_S$ satisfies the two-sided PL condition, i.e., there exist constants $\beta_1(S), \beta_2(S) > 0$ such that the following inequalities hold for all $w \in W, v \in V$

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

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

As a combination of our generalization bounds and optimization error bounds in [Yang et al., 2020a], we can derive the following informal corollary on primal population risks by early stopping the algorithm to balance the optimization and generalization. It gives the first primal risk bounds for learning with nonconvex-strongly-concave functions.

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

**Assumption 2.4 (Existence of Saddle Point)** Assume for any $S$, $F_S$ has at least one
saddle point. Assume for any \( \mathbf{v} \), \( \min_{\mathbf{w}} F_S(\mathbf{w}, \mathbf{v}) \) has a nonempty solution set and a finite optimal value. Assume for any \( \mathbf{w} \), \( \max_{\mathbf{v}} F_S(\mathbf{w}, \mathbf{v}) \) has a nonempty solution set and a finite optimal value.

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

**Lemma 2.15** Assume that for any \( \mathbf{w} \) and \( S \), the function \( \mathbf{v} \mapsto F_S(\mathbf{w}, \mathbf{v}) \) is \( \rho \)-strongly-concave. Suppose for any \( \mathbf{w}, \mathbf{v}, \mathbf{v}' \) and for any \( z \) we have

\[
|f(\mathbf{w}, \mathbf{v}; z) - f(\mathbf{w}, \mathbf{v}'; z)| \leq G \|\mathbf{v} - \mathbf{v}'\|_2. \tag{2.58}
\]

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

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

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

\[
F_S(\mathbf{w}, \hat{\mathbf{v}}^*_S) - F_S(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}})
= \frac{1}{n} \sum_{j \neq i} \left( f(\mathbf{w}, \hat{\mathbf{v}}^*_S; z_j) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_j) \right) + \frac{1}{n} \left( f(\mathbf{w}, \hat{\mathbf{v}}^*_S; z_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i) \right)
= \frac{1}{n} \left( f(\mathbf{w}, \hat{\mathbf{v}}^*_S; z'_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z'_i) \right) + \frac{1}{n} \left( f(\mathbf{w}, \hat{\mathbf{v}}^*_S; z_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i) \right)
+ F_{S^{(i)}}(\mathbf{w}, \hat{\mathbf{v}}^*_S) - F_{S^{(i)}}(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}})
\leq \frac{1}{n} \left( f(\mathbf{w}, \hat{\mathbf{v}}^*_S; z'_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z'_i) \right) + \frac{1}{n} \left( f(\mathbf{w}, \hat{\mathbf{v}}^*_S; z_i) - f(\mathbf{w}, \hat{\mathbf{v}}^*_{S^{(i)}}; z_i) \right)
\leq \frac{2G}{n} \|\hat{\mathbf{v}}^*_S - \hat{\mathbf{v}}^*_{S^{(i)}}\|_2, \tag{2.59}
\]

where the first inequality follows from the fact that \( \hat{\mathbf{v}}^*_{S^{(i)}} \) is the maximizer of \( F_{S^{(i)}}(\mathbf{w}, \cdot) \) and the second inequality follows from (2.58). Since \( F_S \) is strongly-concave and \( \hat{\mathbf{v}}^*_S \) maximizes
\( F_S(w, \cdot) \), we know
\[
\frac{\rho}{2} \| \hat{v}^*_S - \hat{v}^*_{S(i)} \|^2 \leq F_S(w, \hat{v}^*_S) - F_S(w, \hat{v}^*_{S(i)}).
\]
Combining it with (2.59) we get \( \| \hat{v}^*_S - \hat{v}^*_{S(i)} \|^2 \leq \frac{4G}{\rho n} \). By (2.58), the following inequality holds for any \( z \)
\[
| f(w, \hat{v}^*_S; z) - f(w, \hat{v}^*_{S(i)}; z) | \leq \frac{4G^2}{\rho n}.
\]
Since \( z_i \) and \( z'_i \) are i.i.d., we have
\[
E[F(w, \hat{v}^*_S)] = E[F(w, \hat{v}^*_{S(i)})] = \frac{1}{n} \sum_{i=1}^{n} E[f(w, \hat{v}^*_{S(i)}; z_i)],
\]
where the last identity holds since \( z_i \) is independent of \( \hat{v}^*_{S(i)} \). Therefore
\[
E[F_S(w, \hat{v}^*_S) - F(w, \hat{v}^*_S)] = \frac{1}{n} \sum_{i=1}^{n} E[f(w, \hat{v}^*_S; z_i) - f(w, \hat{v}^*_{S(i)}; z_i)] \leq \frac{4G^2}{\rho n}.
\]
The proof is complete.

**Corollary 2.1** Let \( \beta_1, \rho > 0 \). Let Assumptions 2.1, 2.2, 2.3 with \( \beta_1(S) \geq \beta_1, \beta_2(S) \geq \rho \) and 2.4 hold. Assume for any \( w \) and any \( S \), the functions \( v \mapsto F(w, v) \) and \( v \mapsto F_S(w, v) \) are \( \rho \)-strongly concave. Let \( \{w_t, v_t\} \) be the sequence produced by (2.54) with \( \eta_{w,t} \approx 1/(\beta t) \) and \( \eta_{v,t} \approx 1/(\beta_1 \rho^2 t) \). Then for \( T \approx \left( \frac{n}{\beta_1 \rho^2} \right)^{\frac{2L+1}{2L}} \), we have
\[
E[R(w_T) - R(w^*)] = O\left( n^{-\frac{2L+1}{2L+1}} \beta_1^{-\frac{2L}{2L+1}} \rho^{-\frac{2L+1}{2L+1}} \right),
\]
where \( c \approx 1/(\beta_1 \rho^2) \).

**Proof**: We have the error decomposition
\[
R(w_T) - R(w^*) = (R(w_T) - R_S(w_T)) + (R_S(w_T) - R_S(w^*)) + (R_S(w^*) - R(w^*)). \tag{2.60}
\]
First we consider the term \( R(w_T) - R_S(w_T) \). Analogous to the proof of Theorem 2.9 (i.e., the only difference is to replace the conditional expectation of function values in (2.53) with
the conditional expectation of \( \mathbb{E}[\|w_T - w'_T\|_2 + \|v_T - v'_T\|_2] \), one can show that AGDA is \( O(n^{-1}T^{\epsilon}) \)-argument stable (note the step sizes satisfy \( \eta_{w,t} + \eta_{v,t} \leq c/t \)). This together with Part (b) of Theorem 2.1 implies that

\[
\mathbb{E}[R(w_T) - R_S(w_T)] = O((\rho n)^{-1}T^{\epsilon}).
\] (2.61)

For the term \( R_S(w_T) - R_S(w^*) \), the optimization error bounds in [Yang et al. 2020a] show that

\[
\mathbb{E}[R_S(w_T) - R_S(w^*)] = O\left(\frac{1}{\beta_1^2 \rho^4 T}\right).
\] (2.62)

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

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

where \( \hat{v}^*_S = \text{arg max}_v F_S(w^*, v) \). The second term \( \mathbb{E}[F(w^*, \hat{v}^*_S) - F(w^*, v^*)] \leq 0 \) since \((w^*, v^*)\) is a saddle point of \( F \). Therefore by Lemma 2.15 we have

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

We can plug the above inequality, (2.61), (2.62) into (2.60), and get

\[
\mathbb{E}[R(w_T) - R(w^*)] = O((\rho n)^{-1}T^{\epsilon}) + O\left(\frac{1}{\beta_1^2 \rho^4 T}\right) + O\left(\frac{1}{\rho n}\right).
\]

We can choose \( T = \left(\frac{n}{\beta_1^2 \rho^4}\right)^{\frac{\epsilon + 1}{\epsilon + 4}} \) to get the stated excess primal population risk bounds. The proof is complete.

**Lemma 2.16** Let Assumption 2.3 hold. For any \( u = (w, v) \) and any stationary point \( u(S) = (w(S), v(S)) \) of \( F_S \), we have

\[
-\frac{\|\nabla_v F_S(w, v)\|^2}{2\beta_2(S)} \leq F_S(u) - F_S(u(S)) \leq \frac{\|\nabla_w F_S(w, v)\|^2}{2\beta_1(S)}.
\]

**Proof:** Since \( u(S) \) is a stationary point, it is also a saddle point under the PL condition [Yang]
projection of $u$ onto the set of stationary points of $S$, which means that

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

It then follows that

$$F_S(u) - F_S(u(S)) = F_S(w, v) - F_S(w(S), v) + F_S(w(S), v') - F_S(w(S), v(S))$$

$$\leq F_S(w, v) - F_S(w(S), v)$$

$$\leq F_S(w, v) - \inf_{w' \in \mathcal{W}} F_S(w', v)$$

$$\leq \frac{1}{2\beta_1(S)} \|\nabla_w F_S(w, v)\|_2^2,$$

where in the last inequality we have used Assumption 2.3. In a similar way, we know

$$F_S(u) - F_S(u(S)) = F_S(w, v) - F_S(w, v(S)) + F_S(w, v') - F_S(w(S), v(S))$$

$$\geq F_S(w, v) - F_S(w(S), v(S))$$

$$\geq F_S(w, v) - \sup_{v'} F_S(w, v')$$

$$\geq -\frac{1}{2\beta_2(S)} \|\nabla_v F_S(w, v)\|_2^2.$$

For gradient dominated problems, we further have the following error bounds. Note we do not need the smoothness assumption here.

**Theorem 2.10** Let Assumptions 2.1, 2.3 hold. Let $u_S = (A_w(S), A_v(S))$ and $u^{(S)}_S$ be the projection of $u_S$ onto the set of stationary points of $F_S$. Then,

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

**Proof:** Let $S' = \{z'_1, \ldots, z'_{n}\}$ be drawn independently from $\rho$. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$. Let $u_S = (A_w(S), A_v(S))$ and $u^{(S)}_S$ be the projection of $u_S$ onto the set of stationary points of $F_S$. For each $i \in [n]$, we denote $u_i = (A_w(S^{(i)}), A_v(S^{(i)}))$ and $u^{(i)}_i$ the projection of $u_i$ onto the set of stationary points of $F_{S^{(i)}}$. Then $\nabla F_{S^{(i)}}(u^{(i)}_i) = 0$. 

71
We decompose \( f(u_i; z_i) - f(u_S; z_i) \) as follows

\[
f(u_i; z_i) - f(u_S; z_i)
\]

\[
= (f(u_i; z_i) - f(u_i^{(i)}; z_i)) + (f(u_i^{(i)}; z_i) - f(u_S^{(i)}; z_i)) + (f(u_S^{(i)}; z_i) - f(u_S; z_i)). \tag{2.63}
\]

We now address the above three terms separately. We first address \( f(u_i^{(i)}; z_i) - f(u_S^{(i)}; z_i) \). According to the definition of \( F_S, S, S^{(i)} \), we know

\[
f(u_i^{(i)}; z_i) = nF_S(u_i^{(i)}) - nF_{S^{(i)}}(u_i^{(i)}) + f(u_i^{(i)}; z_i').
\]

Since \( z_i \) and \( z_i' \) follow from the same distribution, we know \( \mathbb{E}[f(u_i^{(i)}; z_i')] = \mathbb{E}[f(u_S^{(i)}; z_i)] \) and further get

\[
\mathbb{E}[f(u_i^{(i)}; z_i)] = n\mathbb{E}[F_S(u_i^{(i)})] - n\mathbb{E}[F_{S^{(i)}}(u_i^{(i)})] + \mathbb{E}[f(u_S^{(i)}; z_i)].
\]

It then follows that

\[
\mathbb{E}[f(u_i^{(i)}; z_i) - f(u_S^{(i)}; z_i)] = n\mathbb{E}[F_S(u_i^{(i)}) - F_{S^{(i)}}(u_i^{(i)})] = n\mathbb{E}\left[F_S(u_i^{(i)}) - F_S(u_S^{(i)})\right], \tag{2.64}
\]

where we have used the following identity due to the symmetry between \( z_i \) and \( z_i' \):

\[
\mathbb{E}[F_{S^{(i)}}(u_i^{(i)})] = \mathbb{E}[F_S(u_S^{(i)})].
\]

By the PL condition of \( F_S \), it then follows from (2.64) and Lemma 2.16 that

\[
\mathbb{E}[f(u_i^{(i)}; z_i) - f(u_S^{(i)}; z_i)] \leq \frac{n}{2} \mathbb{E}\left[\frac{1}{\beta_1(S)} \| \nabla_w F_S(u_i^{(i)}) \|_2^2 \right]. \tag{2.65}
\]

According to the definition of \( u_i^{(i)} \) we know \( \nabla_w F_{S^{(i)}}(u_i^{(i)}) = 0 \) and therefore \((a + b)^2 \leq 2a^2 + 2b^2\)

\[
\| \nabla_w F_S(u_i^{(i)}) \|_2^2 = \left\| \nabla_w F_{S^{(i)}}(u_i^{(i)}) - \frac{1}{n} \nabla_w f(u_i^{(i)}; z_i') + \frac{1}{n} \nabla_w f(u_i^{(i)}; z_i) \right\|_2^2
\]

\[
\leq \frac{2}{n^2} \| \nabla_w f(u_i^{(i)}; z_i') \|_2^2 + \frac{2}{n^2} \| \nabla_w f(u_i^{(i)}; z_i) \|_2^2 \leq \frac{4G^2}{n^2}, \tag{2.66}
\]

72
where we have used Assumption 2.1. This together with (2.65) gives
\[
\mathbb{E}[f(u_i^{(i)}; z_i) - f(u_S^{(S)}; z_i)] \leq \frac{2G^2}{n} \mathbb{E}\left[\frac{1}{\beta_1(S)}\right].
\] (2.67)

We then address \(f(u_i; z_i) - f(u_i^{(i)}; z_i)\). Since \(u_i\) and \(u_i^{(i)}\) are independent of \(z_i\), we know
\[
\mathbb{E}[f(u_i; z_i) - f(u_i^{(i)}; z_i)] = \mathbb{E}[F(u_i) - F(u_i^{(i)})] = \mathbb{E}[F(u_S) - F(u_S^{(S)})],
\] (2.68)

where we have used the symmetry between \(z_i\) and \(z'_i\). Finally, we address \(f(u_S^{(S)}; z_i) - f(u_S; z_i)\). By the definition of \(u_S^{(S)}\) we know
\[
\sum_{i=1}^{n} (f(u_S^{(S)}; z_i) - f(u_S; z_i)) = n(F_S(u_S^{(S)}) - F_S(u_S)).
\] (2.69)

Plugging (2.67), (2.68) and the above inequality back into (2.63), we derive
\[
\sum_{i=1}^{n} \mathbb{E}[f(u_i; z_i) - f(u_S; z_i)] \leq \frac{2G^2}{n} \mathbb{E}\left[\frac{1}{\beta_1(S)}\right] + n\mathbb{E}[F(u_S) - F(u_S^{(S)})] + n\mathbb{E}[F_S(u_S^{(S)}) - F_S(u_S)].
\]

Since \(z_i\) and \(z'_i\) are drawn from the same distribution, we know
\[
\mathbb{E}[F(u_S) - F_S(u_S)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[F(u_i) - F_S(u_S)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(u_i; z_i) - f(u_S; z_i)]
\]
\[
\leq \frac{2G^2}{n} \mathbb{E}\left[\frac{1}{\beta_1(S)}\right] + \mathbb{E}[F(u_S) - F(u_S^{(S)})] + \mathbb{E}[F_S(u_S^{(S)}) - F_S(u_S)],
\] (2.70)

where the second identity holds since \(z_i\) is independent of \(u_i\). It then follows that
\[
\mathbb{E}[F(u_S^{(S)}) - F_S(u_S^{(S)})] \leq \frac{2G^2}{n} \mathbb{E}\left[\frac{1}{\beta_1(S)}\right].
\] (2.71)

According to the Lipschitz continuity we know
\[
|F(u_S) - F(u_S^{(S)})| \leq G\|u_S - u_S^{(S)}\|_2 \quad \text{and} \quad |F_S(u_S) - F_S(u_S^{(S)})| \leq G\|u_S - u_S^{(S)}\|_2.
\]
Plugging the above inequality back into (2.70), we derive the following inequality

\[ \mathbb{E} [F(u_S) - F_S(u_S)] \leq \frac{2G^2}{n} \mathbb{E} \left[ \frac{1}{\beta_1(S)} \right] + 2G \mathbb{E} \left[ \|u_S - u_S^{(S)}\|_2 \right]. \tag{2.72} \]

By Lemma 2.16 and (2.64), we can also have

\[ \mathbb{E} [ f(u_i^{(i)}; z_i) - f(u_S^{(S)}; z_i) ] \geq -\frac{n}{2} \mathbb{E} \left[ \frac{1}{\beta_2(S)} \right] \| \nabla_F S(u_i^{(i)}) \|_2^2. \]

Using this inequality, one can analyze analogously to (2.72) and derive the following inequality

\[ \mathbb{E} [F(u_S) - F_S(u_S)] \geq -\frac{2G^2}{n} \mathbb{E} \left[ \frac{1}{\beta_2(S)} \right] - 2G \mathbb{E} \left[ \|u_S - u_S^{(S)}\|_2 \right]. \]

The stated inequality follows from the above inequality and (2.72).

\[ \square \]

**Remark 2.7** Note \( \|u_S - u_S^{(S)}\|_2 \) measures how far the point found by \( A \) is from the set of stationary points of \( F_S \), and can be interpreted as an optimization error. Therefore, Theorem 2.10 gives a connection between generalization error and optimization error. For a variant of AGDA with noiseless stochastic gradients, it was shown that \( \| (w_T, v_T) - (w_T, v_T)^{(S)} \|_2 \) decays linearly w.r.t. \( T \) [Yang et al., 2020a]. We can plug this linear convergent optimization bound into Theorem 2.10 to directly get generalization bounds. If \( A \) returns a saddle point of \( F_S \), then \( \|u_S - u_S^{(S)}\|_2 = 0 \) and therefore \( \| \mathbb{E} [F(u_S) - F_S(u_S)] \| = O(\frac{n^{-1}}{\max\{E[1/\beta_1(S)], E[1/\beta_2(S)]\}}) \). Generalization errors of this particular ESP were studied in [Zhang et al., 2021] for SC-SC minimax problems, which were extended to more general gradient-dominated problems in Theorem 2.10. Furthermore, Theorem 2.10 applies to any optimization algorithm instead of the specific ESP. It should be mentioned that [Zhang et al., 2021] addressed PD population risks, while we consider plain generalization errors.
CHAPTER 3
Differential Privacy and Utility for Minimax Problems

3.1 Problem Formulation

Our main results in this chapter are based on Yang et al. [2022]. In this section, we introduce necessary assumptions, notations and the DP-SGDA algorithm. Some assumptions are carried over from Chapter 2.

Assumption 3.1 For any \( j \in [n] \), the gradients \( \nabla_w f(w, v; z_j) \) and \( \nabla_v f(w, v; z_j) \) have bounded variances \( B_w \) and \( B_v \) respectively. Let \( B = \max\{B_w, B_v\} \).

Our aim is to design a randomized algorithm satisfying \((\epsilon, \delta)\)-DP which solves the empirical minimax problem

\[
\min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^{n} f(w, v; z_i).
\] (3.1)

Notice that in the standard ERM problem, which involves the minimization only with respect to \( w \), DP-SGD [Bassily et al., 2019; Song et al., 2013; Wang et al., 2020a] 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 [Xie et al., 2018; Zhang et al., 2018] adds Gaussian noises per iteration to the randomized gradient mapping

\[
(g_{w,t}, g_{v,t}) = \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_j}), -\frac{1}{m} \sum_{j=1}^{m} \nabla_v f(w_t, v_t; z_{i_j}) \right)
\]

where the index of example \( z_{i_j} \) is from the mini-batch \( I_t \). Then, the primal variable \( w \) is updated by gradient descent based on the noisy gradient \( g_{w,t} + \xi_t \) and the dual variable \( v \) is updated by gradient ascent based on the noisy gradient \( g_{v,t} + \zeta_t \). The pseudo-code for DP-SGDA is given in Algorithm [1]. The noise levels \( \sigma_w, \sigma_v \) are given by (3.2) which will be
specified soon in Section 3.2 in order to guarantee \((\epsilon, \delta)-\text{DP}\). The notations \(\Pi_W(\cdot)\) and \(\Pi_V(\cdot)\) denote the projections to \(W\) and \(V\), respectively. From now on, the notation \(\mathcal{A}\) denotes the DP-SGDA algorithm and its output is denoted by \(\mathcal{A}(S) = (\mathcal{A}_w(S), \mathcal{A}_v(S))\).

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

**Method**

1: **Inputs:** data \(S = \{z_i : i \in [n]\}\), privacy budget \(\epsilon, \delta\), number of iterations \(T\), learning rates \(\{\eta_w,t, \eta_v,t\}_{t=1}^T\), and initialize \((w_0,v_0)\)

2: Compute noise parameters \(\sigma_w^2\) and \(\sigma_v^2\) based on Eq. (3.2)

3: for \(t = 1\) to \(T\) do

4: Sample a mini-batch \(I_t = \{i^1_t, \cdots, i^m_t \in [n]\}\) uniformly with replacement

5: Sample independent noises \(\xi_t \sim \mathcal{N}(0, \sigma_w^2 I_{d_1})\) and \(\zeta_t \sim \mathcal{N}(0, \sigma_v^2 I_{d_2})\)

6: \(w_{t+1} = \Pi_W(\mathcal{A}_w(t) = (\frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{i^j_t}) + \xi_t))\)

7: \(v_{t+1} = \Pi_V(\mathcal{A}_v(t) = (\frac{1}{m} \sum_{j=1}^m \nabla_v f(w_t, v_t; z_{i^j_t}) + \zeta_t))\)

8: end for

9: Outputs: \((\bar{w}_T, \bar{v}_T) = \frac{1}{T} \sum_{t=1}^T (w_t, v_t)\) or \((w_T, v_T)\)

### 3.2 Privacy and Utility Guarantees

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.

**Lemma 3.1** (Abadi et al. [2016]) Consider a sequence of mechanisms \(\{\mathcal{A}_t\}_{t \in [T]}\) and the composite mechanism \(\mathcal{A} = (\mathcal{A}_1, \cdots, \mathcal{A}_T)\).

a) **[Composability]** For any \(\lambda\),

\[ \alpha_{\mathcal{A}}(\lambda) = \sum_{t=1}^T \alpha_{\mathcal{A}_t}(\lambda). \]

b) **[Tail bound]** For any \(\epsilon\), the mechanism \(\mathcal{A}\) is \((\epsilon, \delta)\)-differentially private for

\[ \delta = \min_{\lambda} \alpha_{\mathcal{A}}(\lambda) - \lambda \epsilon. \]
Lemma 3.2 (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 \( g_t : \mathbb{Z}^m \to \mathbb{R}^d \) has \( \ell_2 \)-sensitivity of 1. And each \( S_t \) is a subsample of size \( m \) obtained by uniform sampling without replacement\footnote{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. 2019].} from \( S \), i.e. \( S_t \sim (\text{Unif}(S))^m \). Then

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

Theorem 3.1 Let Assumption 2.1 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 \((\epsilon, \delta)\)-differentially private for any \( \delta > 0 \) if we choose

\[
\sigma_w = \frac{c_2 G_w \sqrt{T \log(1/\delta)}}{n \epsilon}, \quad \sigma_v = \frac{c_3 G_v \sqrt{T \log(1/\delta)}}{n \epsilon}. \tag{3.2}
\]

Proof: Let \( S = \{z_1, \ldots, z_n\} \) and \( S' = \{z'_1, \ldots, z'_n\} \) be two neighboring datasets. At iteration \( t \), we first focus on \( A^w_t = \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z'_{i_j}) + \xi_t \). Since \( f(\cdot, v; z) \) is \( G_w \)-Lipschitz continuous, it implies for any neighboring datasets \( S, S' \),

\[
\left\| \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z'_{i_j}) - \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z'_{i_j}) \right\|_2 \leq \frac{2G_w}{m}.
\]

Therefore we can define \( g_t(S_t) = \frac{1}{2G_w} \sum_{j=1}^m \nabla_w f(w_t, v_t; z'_{i_j}) \) such that \( \Delta(g_t) = 1 \). By Lemma 3.1 b) and 3.2 the log moment of the composite mechanism \( A^w = (A^w_1, \ldots, A^w_T) \) can be bounded as follows

\[
\alpha_{A^w}(\lambda) \leq \frac{m^2 T \lambda^2}{n^2 \tilde{\sigma}_w^2}.
\]

where \( \tilde{\sigma}_w = \sigma_w / 2G_w \). Similarly, since \( A^v_t = \nabla_v f(w_t, v_t; z_{i_t}) + \xi_t \) has \( \ell_2 \)-sensitivity \( 2G_v / m \), then the log moment of the final output \( A = (A^w_1, A^v_1, \ldots, A^w_T, A^v_T) \) can be bounded as follows

\[
\alpha_A(\lambda) \leq \alpha_{A^v}(\lambda) + \alpha_{A^w}(\lambda) \leq \frac{m^2 T \lambda^2}{n^2 \tilde{\sigma}_w^2} + \frac{m^2 T \lambda^2}{n^2 \tilde{\sigma}_v^2}.
\]
By Lemma 3.1 a), to guarantee \( A \) to be \((\epsilon, \delta)\)-differentially private, it suffices that
\[
\frac{\lambda^2 m^2 T}{n^2 \tilde{\sigma}_w^2} \leq \frac{\lambda \epsilon}{4}, \quad \frac{\lambda^2 m^2 T}{n^2 \tilde{\sigma}_v^2} \leq \frac{\lambda \epsilon}{4}, \quad \exp(-\frac{\lambda \epsilon}{4}) \leq \delta, \quad \lambda \leq \tilde{\sigma}_w^2 \log\left(\frac{n}{m \tilde{\sigma}_w}\right) \quad \text{and} \quad \lambda \leq \tilde{\sigma}_v^2 \log\left(\frac{n}{m \tilde{\sigma}_v}\right)
\]

It is now easy to verify that when \( \epsilon = c_1 m^2 T/n^2 \), we can satisfy all these conditions by setting
\[
\tilde{\sigma}_w \geq \frac{c_2 \sqrt{T \log(1/\delta)}}{n \epsilon} \quad \text{and} \quad \tilde{\sigma}_v \geq \frac{c_3 \sqrt{T \log(1/\delta)}}{n \epsilon}
\]
for some explicit constants \( c_1, c_2 \) and \( c_3 \).

\[\square\]

**Remark 3.1** In practice, given privacy budget \( \epsilon, \delta \) and parameters \( m, T \), the constant \( c_2 \) and hence \( \sigma \) 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 3.1. 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 \( \sigma_w = \frac{8G_w \sqrt{T \log(1/\delta)}}{n \epsilon}, \sigma_v = \frac{8G_v \sqrt{T \log(1/\delta)}}{n \epsilon} \).

**Proof:** Without loss of generality, we consider with only one \( \sigma \) in the the proof of Theorem 3.1. Then algorithm \( A \) is guaranteed to be \((\epsilon, \delta)\)-DP if one can find \( \lambda > 0 \) such that
\[
\frac{\lambda^2 m^2 T}{n^2 \sigma^2} \leq \frac{\lambda \epsilon}{2}, \quad \exp(-\frac{\lambda \epsilon}{2}) \leq \delta, \quad \text{and} \quad \lambda \leq \sigma^2 \log\left(\frac{n}{m \sigma}\right)
\]

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 \( \lambda \) satisfies the third inequality, which is straightforward by the choice of \( m \) and \( \epsilon \leq 1 \).

\[\square\]

**Remark 3.2** Our Algorithm 1 allows the application of independent noises \( \xi_t, \zeta_t \) with different \( \sigma_w, \sigma_v \), respectively. In [Boob and Guzmán, 2021], a uniform \( \sigma \) is used (Theorem 5.4 or 7.4 there) for both primal and dual variables. In many examples, the primal and dual gradients \( \nabla_w f(w_t, v_t, z_{i_t}), \nabla_v f(w_t, v_t, z_{i_t}) \) enjoy different Lipschitz constants (\( \ell_2 \)-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, 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.

### 3.2.1 Convex-Concave Case

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:

\[
\Delta^w(\bar{w}_T, \bar{v}_T) = \Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^S(\bar{w}_T, \bar{v}_T) + \Delta^S(\bar{w}_T, \bar{v}_T),
\]

(3.3)

where the term \( \Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^S(\bar{w}_T, \bar{v}_T) \) is the generalization error and \( \Delta^S(\bar{w}_T, \bar{v}_T) \) is the optimization error.

The estimation for the optimization error can be conducted by standard techniques [Nemirovski et al., 2009]. We give a self-contained proof. The generalization error is estimated using a concept of weak stability defined in Chapter 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 (3.3) can yield the bound in Theorem 3.2.

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 3.3** Suppose Assumption 2.1 holds, and \( F_S \) is convex-concave. Let the stepsizes \( \eta_{w,t} = \eta_{v,t} = \eta \), \( t \in [T] \) for some \( \eta > 0 \). Then Algorithm 1 satisfies

\[
\sup_{v \in V} \mathbb{E}_{\mathcal{A}}[F_S(\bar{w}_T, v)] - \inf_{w \in W} \mathbb{E}_{\mathcal{A}}[F_S(w, \bar{v}_T)] \\
\leq \frac{\eta(G_w^2 + G_v^2)}{2} + \frac{D_w^2 + D_v^2}{\eta T} + \frac{(D_w G_w + D_v G_v)}{\sqrt{mT}} + \eta d(\sigma_w^2 + \sigma_v^2).
\]

**Proof:** According to the non-expansiveness of projection and update rule of Algorithm
for any \( w \in \mathcal{W} \), we have

\[
\begin{align*}
\| w_{t+1} - w \|^2 &\leq \| w_t - w - \eta \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \eta \xi_t \|^2 \\
&\leq \| w_t - w \|^2 + 2\eta \langle w - w_t, \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) \rangle + \eta^2 \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) \right)^2 \\
&\quad + \eta^2 \| \xi_t \|^2 + 2\eta^2 \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) \right) , \xi_t \rangle \\
&\leq \| w_t - w \|^2 + 2\eta \langle w - w_t, \nabla_w F_S(w_t, v_t) \rangle \\
&\quad + 2\eta \left( w - w_t, \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right) \\
&\quad + \eta^2 G_w^2 + \eta^2 \| \xi_t \|^2 + 2\eta^2 \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) , \xi_t \right) + 2\eta \langle w - w_t, \xi_t \rangle,
\end{align*}
\]

where in the last inequality we have used \( f(\cdot, v, z_{i_t}) \) is \( G_w \)-Lipschitz continuous. According to the convexity of \( F_S(\cdot, v_i) \) we know

\[
2\eta (F_S(w_t, v_i) - F_S(w, v_t)) \leq \| w_t - w \|^2 - \| w_{t+1} - w \|^2 \\
\quad + 2\eta \left( w - w_t, \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right) \\
\quad + \eta^2 G_w^2 + \eta^2 \| \xi_t \|^2 + 2\eta^2 \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) , \xi_t \right) + 2\eta \langle w - w_t, \xi_t \rangle.
\]

Taking a summation of the above inequality from \( t = 1 \) to \( T \) we derive

\[
2\eta \sum_{t=1}^{T} (F_S(w_t, v_t) - F_S(w, v_t)) \leq \| w_1 - w \|^2 \\
\quad + 2\eta \sum_{t=1}^{T} \left( w - w_t, \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right) \\
\quad + T \eta^2 G_w^2 + \eta^2 \sum_{t=1}^{T} \| \xi_t \|^2 + 2\eta^2 \sum_{t=1}^{T} \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) , \xi_t \right) + 2\eta \langle w - w_t, \xi_t \rangle.
\]
It then follows from the concavity of \( F_S(w, \cdot) \) and Schwartz’s inequality that

\[
2 \sum_{t=1}^{T} \eta (F_S(w_t, v_t) - F_S(w, v_T)) \\
\leq 2D_w^2 - 2\eta \sum_{t=1}^{T} \left\langle w_t, \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right\rangle \\
+ 2D_w \eta \left\| \sum_{t=1}^{T} \left( \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right) \right\|_2 \\
+ T\eta^2 G_w^2 + \eta^2 \left\| \xi_t \right\|_2^2 + 2\eta^2 \sum_{t=1}^{T} \left\langle \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}), \xi_t \right\rangle + 2\eta (w - w_t, \xi_t). \tag{3.4}
\]

We can take expectations on the randomness of \( A \) over both sides of (3.4) and get

\[
2\eta \sum_{t=1}^{T} \mathbb{E}_A[F_S(w_t, v_t) - F_S(w, v_T)] \\
\leq 2D_w^2 + 2D_w \eta \mathbb{E}_A \left[ \left\| \sum_{t=1}^{T} \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right\|_2 \right] + T\eta^2 G_w^2 \\
+ \eta^2 d_1 \sigma_w^2,
\]

where we used that the variance \( \mathbb{E}_A[||\xi_t||_2^2] = d_1 \sigma_w^2 \), the unbiasedness of the sampled gradient \( \mathbb{E}_A[(w_t, \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i}) - \nabla_w F_S(w_t, v_t))] = 0 \), the independence between random variables \( \mathbb{E}_A[(\frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}), \xi_t)] = 0 \) and \( \mathbb{E}_A[(w - w_t, \xi_t)] = 0 \). Since the above inequality holds for all \( w \), we further get

\[
2\eta \sum_{t=1}^{T} \mathbb{E}_A[F_S(w_t, v_t)] - \inf_{w \in \mathcal{W}} \mathbb{E}_A[F_S(w, v_T)] \\
\leq 2D_w^2 + 2D_w \eta \mathbb{E}_A \left[ \left\| \sum_{t=1}^{T} \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{i_t}) - \nabla_w F_S(w_t, v_t) \right\|_2 \right] + T\eta^2 G_w^2 + \eta^2 d_1 \sigma_w^2, \tag{3.5}
\]
According to Jensen’s inequality and $G_w$-Lipschitz continuity we further derive

\[
\left( \mathbb{E}_A \left[ \left\| \sum_{t=1}^T \left( \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w F_S(w_t, v_t) \right) \right\|_2 \right] \right)^2 \\
\leq \mathbb{E}_A \left[ \left\| \sum_{t=1}^T \left( \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w F_S(w_t, v_t) \right) \right\|_2^2 \right] \\
= \sum_{t=1}^T \mathbb{E}_A \left[ \left\| \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w F_S(w_t, v_t) \right\|_2^2 \right] \\
\leq T G_w^2.
\]

Plugging the above estimate into (3.5) we arrive

\[
2\eta \sum_{t=1}^T \mathbb{E}_A[F_S(w_t, v_t)] - \inf_{w \in W} \mathbb{E}_A[F_S(w, \bar{v}_T)] \leq 2D_w^2 + \frac{2D_w \eta G_w \sqrt{T}}{\sqrt{m}} + T \eta^2 G_w^2 + T \eta^2 d_1 \sigma_w^2.
\]

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

\[
\frac{1}{T} \sum_{t=1}^T \mathbb{E}_A[F_S(w_t, v_t)] - \inf_{w \in W} \mathbb{E}_A[F_S(w, \bar{v}_T)] \leq \frac{D_w^2}{\eta T} + \frac{D_w G_w}{\sqrt{m} \sqrt{T}} + \frac{\eta G_w^2}{2} + \frac{\eta d_1 \sigma_w^2}{2}. \tag{3.6}
\]

In a similar way, we can show that

\[
\frac{1}{T} \sum_{t=1}^T \sup_{v \in V} \mathbb{E}_A[F_S(w_T, v)] - \mathbb{E}_A[F_S(w_t, v_t)] \leq \frac{D_v^2}{\eta T} + \frac{D_v G_v}{\sqrt{m} \sqrt{T}} + \frac{\eta G_v^2}{2} + \frac{\eta d_2 \sigma_v^2}{2}. \tag{3.7}
\]

The stated bound then follows from (3.6) and (3.7) and the fact that $d = \max\{d_1, d_2\}$.

Next we move on to the generalization error. Firstly, Chapter 2 bridges the generalization and the stability. The stability analysis is given in the following lemma. This lemma is an extension of the uniform argument stability results in Chapter 2 to the case of mini-batch DP-SGDA.

**Lemma 3.4** Suppose the function $F_S$ is convex-concave. Let the stepsizes $\eta_{w,t} = \eta_{v,t} = \eta$ for some $\eta > 0$. 


a) Assume Assumption 2.1 and Assumption 2.2 hold, then Algorithm I satisfies
\[
\Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^w_S(\bar{w}_T, \bar{v}_T) \leq \frac{4\sqrt{e(T + T^2/n)}(G_w + G_v)^2\eta \exp(L^2T\eta^2/2)}{\sqrt{n}}.
\]

b) Assume Assumption 2.1 holds, then Algorithm I satisfies
\[
\Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^w_S(\bar{w}_T, \bar{v}_T) \leq 4\sqrt{2\eta}(G_w + G_v)^2\left(\sqrt{T} + \frac{T}{n}\right).
\]

Proof: Without loss of generality, let \(S = \{z_1, \ldots, z_n\}, S' = \{z'_1, \ldots, z'_n\}\) be neighboring datasets differing by the last element, i.e. \(z_n \neq z'_n\). Let \(\{w_t, v_t\}, \{w'_t, v'_t\}\) be the sequence produced by Algorithm I w.r.t. \(S\) and \(S'\), respectively. We first prove Part a). In the case \(n \not\in I_t\), by the non-expansiveness of projection, we have
\[
\begin{align*}
&\left\| \begin{pmatrix} w_{t+1} - w'_{t+1} \\ v_{t+1} - v'_{t+1} \end{pmatrix} \right\|_2^2 \\
\leq & \left\| \begin{pmatrix} w_t - \frac{n}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{ij}) - \eta \xi_t - w'_t + \frac{n}{m} \sum_{j=1}^m \nabla_w f(w'_t, v'_t; z'_{ij}) + \eta \xi_t \\ v_t + \frac{n}{m} \sum_{j=1}^m \nabla_v f(w_t, v_t; z_{ij}) + \eta \xi_t - v'_t - \frac{n}{m} \sum_{j=1}^m \nabla_v f(w'_t, v'_t; z'_{ij}) - \eta \xi_t \end{pmatrix} \right\|_2^2 \\
= & \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2 + \eta \frac{n}{m} \sum_{j=1}^m \left\langle \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix}, \begin{pmatrix} \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w f(w'_t, v'_t; z'_{ij}) \\ \nabla_v f(w_t, v_t; z_{ij}) - \nabla_v f(w'_t, v'_t; z'_{ij}) \end{pmatrix} \right\rangle \\
+ & \left\| \begin{pmatrix} \frac{n}{m} \sum_{j=1}^m (\nabla_w f(w_t, v_t; z_{ij}) - \nabla_w f(w'_t, v'_t; z'_{ij})) \\ \frac{n}{m} \sum_{j=1}^m (\nabla_v f(w_t, v_t; z_{ij}) - \nabla_v f(w'_t, v'_t; z'_{ij})) \end{pmatrix} \right\|_2^2 \\
\leq & (1 + L^2\eta^2) \left\| \begin{pmatrix} w_t - w'_t \\ v_t - v'_t \end{pmatrix} \right\|_2^2,
\end{align*}
\]
where the last inequality follows from Lemma 2.5 and the \(L\)-smoothness assumption. If
\[ n \in I_t, \text{ then it follows that} \]
\[
\begin{aligned}
&\left\| \left( w_{t+1} - w'_{t+1} \right) \right\|^2 + \left\| \left( v_{t+1} - v'_{t+1} \right) \right\|^2 \\
&\quad \leq \left\| \left( w_t - m \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{i_j}) - \eta \xi_t - w'_t + \frac{n}{m} \sum_{j=1}^m \nabla_w f(w'_t, v'_t; z'_{i_j}) + \eta \xi_t \right) \right\|^2 \\
&\quad + \left\| \left( v_t + \frac{n}{m} \sum_{j=1}^m \nabla_v f(w_t, v_t; z_{i_j}) + \eta \xi_t - v'_t - \frac{n}{m} \sum_{j=1}^m \nabla_v f(w'_t, v'_t; z'_{i_j}) - \eta \xi_t \right) \right\|^2 \\
&\quad \leq \frac{1}{m} \sum_{i_j \in I_t, i_j \neq n} \left\| \left( w_t - \eta \nabla_w f(w_t, v_t; z_{i_j}) - w'_t + \eta \nabla_w f(w'_t, v'_t; z'_{i_j}) \right) \right\|^2 \\
&\quad + \frac{1}{m} \left\| \left( v_t + \eta \nabla_v f(w_t, v_t; z_{i_j}) - v'_t - \eta \nabla_v f(w'_t, v'_t; z'_{i_j}) \right) \right\|^2 \\
&\quad \leq \left( 1 + \frac{1}{p} \right) \left( \frac{1}{m} \right)^2 \left\| \left( w_t - w'_t \right) \right\|^2 + \frac{1}{m} \left\| \left( v_t - v'_t \right) \right\|^2 \\
&\quad + \frac{1 + 1/p}{m} \left\| \left( \nabla_w f(w_t, v_t; z_{n}) - \nabla_w f(w'_t, v'_t; z'_{n}) \right) \right\|^2 + \frac{1}{m} \left\| \left( \nabla_v f(w_t, v_t; z_{n}) - \nabla_v f(w'_t, v'_t; z'_{n}) \right) \right\|^2,
\end{aligned}
\] (3.8)

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 \(n \not\in I_t\) happens with probability \(1 - m/n\) and the event \(n \in I_t\) happens with probability \(m/n\). Therefore, we know
\[
\mathbb{E}_{i_t} \left\| \left( w_{t+1} - w'_{t+1} \right) \right\|^2 + \left\| \left( v_{t+1} - v'_{t+1} \right) \right\|^2 \leq \left( 1 + L^2 \eta^2 + p/n \right) \left\| \left( w_t - w'_t \right) \right\|^2 + \frac{4(1 + 1/p)}{n} \eta^2 (G_w^2 + G_v^2) + \frac{m}{n} \left( \frac{1}{m} \right)^2 \left\| \left( v_t - v'_t \right) \right\|^2.
\]
Applying this inequality recursively, we derive

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

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

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

\[
= \frac{4(1 + 1/p)}{n} \left( G_w^2 + G_v^2 \right) \sum_{k=1}^{t} \eta^2 \exp \left( L^2 \sum_{j=k+1}^{t} \eta^2 + p(t - k)/n \right)
\]

\[
\leq \frac{4(1 + 1/p)}{n} \left( G_w^2 + G_v^2 \right) \exp \left( L^2 \sum_{j=1}^{t} \eta^2 + pt/n \right) \sum_{k=1}^{t} \eta^2.
\]

By taking \(p = n/t\) we get

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

Now by the Lipschitz continuity and Jensen’s inequality we have

\[
\sup_{z} \mathbb{E}_A[f(A_w(S), v; z) - f(A_w(S'), v; z)]
\]

\[
+ \sup_{w \in \mathcal{W}} \mathbb{E}_A[f(w, A_v(S); z) - f(w, A_v(S'); z)]
\]

\[
\leq G_w \mathbb{E}_A[\|\tilde{w}_T - w'_{T}\|_2] + G_v \mathbb{E}_A[\|\tilde{v}_T - v'_{T}\|_2]
\]

\[
\leq 4\sqrt{e(T + T^2/n)}(G_w + G_v)^2 \eta \exp(L^2 T \eta^2 / 2).
\]

According to Lemma 2.1 we know

\[
\Delta^w_{t}(\tilde{w}_T, \tilde{v}_T) - \Delta^w_{t}(\tilde{w}'_T, \tilde{v}'_T) \leq 4\sqrt{e(T + T^2/n)}(G_w + G_v)^2 \eta \exp(L^2 T \eta^2 / 2).
\]

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 Chapter 2 we can show

\[
\begin{align*}
\left\| \left( \begin{array}{c}
w_{t+1} - w'_{t+1} \\
v_{t+1} - v'_{t+1}
\end{array} \right) \right\|_2^2 \\
\leq \left\| \left( \begin{array}{c}
w_t - \frac{n}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t, z_{ij}) - \eta \xi_t - w' + \frac{n}{m} \sum_{j=1}^{m} \nabla_v f(w_t, v_t, z_{ij}) + \eta \xi_t \nabla_v f(w'_{t+1}, v'_{t+1}, z_{ij}) - \eta \xi_t \nabla_v f(w'_{t+1}, v'_{t+1}, z_{ij}) - \eta \xi_t \\
v_t + \frac{n}{m} \sum_{j=1}^{m} \nabla_v f(w_t, v_t, z_{ij}) + \eta \zeta_t - v' + \frac{n}{m} \sum_{j=1}^{m} \nabla_v f(w'_{t+1}, v'_{t+1}, z_{ij}) - \eta \zeta_t
\end{array} \right) \right\|_2^2 \\
\leq \left\| \left( \begin{array}{c}
w_t - w' \\
v_t - v'
\end{array} \right) \right\|_2^2 \\
+ 4(G_w^2 + G_v^2) \eta^2.
\end{align*}
\]

Combining the preceding inequality with (3.8) and using the probability of \( n \notin I_t \), we derive

\[
\mathbb{E}_A \left[ \left\| \left( \begin{array}{c}
w_{t+1} - w'_{t+1} \\
v_{t+1} - v'_{t+1}
\end{array} \right) \right\|_2^2 \right] \leq \frac{n - 1}{n} \left( \left\| \left( \begin{array}{c}
w_t - w' \\
v_t - v'
\end{array} \right) \right\|_2^2 + 4(G_w^2 + G_v^2) \eta^2 \right) \\
+ \frac{1 + p}{n} \left\| \left( \begin{array}{c}
w_t - w' \\
v_t - v'
\end{array} \right) \right\|_2^2 + \frac{4(1 + 1/p)}{n} (G_w^2 + G_v^2) \eta^2 \\
= (1 + p/n) \left\| \left( \begin{array}{c}
w_t - w' \\
v_t - v'
\end{array} \right) \right\|_2^2 + 4(G_w^2 + G_v^2) \eta^2 (1 + 1/(np)).
\]

Applying this inequality recursively implies that

\[
\mathbb{E}_A \left[ \left\| \left( \begin{array}{c}
w_{t+1} - w'_{t+1} \\
v_{t+1} - v'_{t+1}
\end{array} \right) \right\|_2^2 \right] \leq 4(G_w^2 + G_v^2) \eta^2 (1 + 1/(np)) \sum_{k=1}^{t} \left( 1 + \frac{p}{n} \right)^{t-k} \\
= 4(G_w^2 + G_v^2) \eta^2 \left( 1 + \frac{1}{np} \right) \frac{n}{p} \left( \left( 1 + \frac{p}{n} \right)^t - 1 \right) = 4(G_w^2 + G_v^2) \eta^2 \left( \frac{n}{p} + \frac{1}{p^2} \right) \left( \left( 1 + \frac{p}{n} \right)^t - 1 \right).
\]

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

\[
\mathbb{E}_A \left[ \left\| \left( \begin{array}{c}
w_{t+1} - w'_{t+1} \\
v_{t+1} - v'_{t+1}
\end{array} \right) \right\|_2^2 \right] \leq 16(G_w^2 + G_v^2) \eta^2 \left( t + \frac{t^2}{n^2} \right).
\]
Now by the Lipschitz continuity and Jensen’s inequality we have

\[
\sup_z \left( \sup_{v \in V} \mathbb{E}_A[f(A_w(S), v; z) - f(A_w(S'), v; z)] + \sup_{w \in W} \mathbb{E}_A[f(w, A_v(S); z) - f(w, A_v(S'); z)] \right) \\
\leq G_w \mathbb{E}_A[\| \bar{w}_T - \bar{w}'_T \|_2] + G_v \mathbb{E}_A[\| \bar{v}_T - \bar{v}'_T \|_2] \\
\leq 4\sqrt{2}(G_w + G_v)^2 \eta^2 \left( \sqrt{T} + \frac{T}{n} \right).
\]

According to Lemma 2.1 we know

\[
\Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^w_S(\bar{w}_T, \bar{v}_T) \leq 32(G_w + G_v)^2 \eta^2 \left( \sqrt{T} + \frac{T}{n} \right).
\]

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

**Theorem 3.2** Assume the function \(f\) is convex-concave. Assume \(W\) and \(V\) are bounded so that \(\max_{w \in W} \| w \|_2 \leq D_w\), \(\max_{v \in V} \| v \|_2 \leq D_v\). And let \(D = \max\{D_w, D_v\}\). Let the stepsizes \(\eta_{w,t} = \eta_{v,t} = \eta\) for all \(t \in [T]\) with some \(\eta > 0\). Under one of the conditions

a) if Assumption 2.1 2.2 hold true and we choose \(\eta \asymp \frac{1}{\max\{\sqrt{n}, \sqrt{d \log(1/\delta)/\epsilon}\}}\) and \(T \asymp n\),

b) or Assumption 2.1 holds true and we choose \(\eta \asymp \frac{1}{n \max\{\sqrt{n}, \sqrt{d \log(1/\delta)/\epsilon}\}}\) and \(T \asymp n^2\),

then Algorithm 1 satisfies

\[
\Delta^w(\bar{w}_T, \bar{v}_T) = \mathcal{O}\left( \max \left\{ \frac{1}{\sqrt{n}}, \frac{\sqrt{d \log(1/\delta)}}{n \epsilon} \right\} \right).
\]

**Proof:** We first focus on Part a). According to Part a) of Lemma 3.4 we know

\[
\Delta^w(\bar{w}_T, \bar{v}_T) - \Delta^w_S(\bar{w}_T, \bar{v}_T) \leq \frac{4\sqrt{e(T + T^2/n)}(G_w + G_v)^2 \eta \exp(L^2T\eta^2/2)}{\sqrt{n}}.
\]
and by Lemma 3.3 we know
\[ \Delta_w^S(\bar{w}_T, \bar{v}_T) \leq \frac{\eta(G_w^2 + G_v^2)}{2} + \frac{D_w^2 + D_v^2}{2\eta T} + \frac{D_w G_w + D_v G_v}{\sqrt{mT}} + \eta d(\sigma_w^2 + \sigma_v^2). \]

Combining the above two quantities we have
\[ \Delta_w(\bar{w}_T, \bar{v}_T) \leq 4\sqrt{e(T + T^2/n)}(G_w + G_v)^2\eta \exp(L^2T\eta^2/2) + \frac{\eta(G_w^2 + G_v^2)}{2} + \frac{D_w^2 + D_v^2}{2\eta T} \]
\[ + \frac{D_w G_w + D_v G_v}{\sqrt{mT}} + \eta d(\sigma_w^2 + \sigma_v^2). \]

Furthermore, by Theorem 3.1 we know
\[ \sigma_w^2 = \mathcal{O}\left(\frac{G_w^2 T \log(1/\delta)}{n^2 \epsilon^2}\right), \quad \sigma_v^2 = \mathcal{O}\left(\frac{G_v^2 T \log(1/\delta)}{n^2 \epsilon^2}\right). \]

Plugging it back into (3.10) we have
\[ \Delta_w(\bar{w}_T, \bar{v}_T) = \mathcal{O}\left(\sqrt{(T + T^2/n)(G_w + G_v)^2\eta \exp(L^2T\eta^2)} \right) \]
\[ + \frac{\eta(G_w^2 + G_v^2)}{2} + \frac{D_w^2 + D_v^2}{2\eta T} + \frac{D_w G_w + D_v G_v}{\sqrt{mT}} + \eta d(\sigma_w^2 + \sigma_v^2) \frac{T d \log(1/\delta)}{n^2 \epsilon^2}. \]

By picking \( T \approx n \) and \( \eta \approx 1/(L \max\{\sqrt{n}, \sqrt{d \log(1/\delta)}/\epsilon\}) \) we have
\[ \exp(L^2T\eta^2) = \mathcal{O}\left(\min\{1, \frac{n \epsilon^2}{d \log(1/\delta)}\}\right) = \mathcal{O}(1) \]
and
\[ \Delta_w^S(\bar{w}_T, \bar{v}_T) \]
\[ = \mathcal{O}\left(\max\{G_w^2 + G_v^2, (G_w + G_v)^2, D_w^2 + D_v^2, D_w G_w + D_v G_v\} \max\left\{\frac{1}{\sqrt{n}}, \frac{\sqrt{d \log(1/\delta)}}{n \epsilon}\right\}\right). \]

We now turn to Part b). According to Lemma 3.4 Part b) we know
\[ \Delta_w^S(\bar{w}_T, \bar{v}_T) \leq 4\sqrt{2}\eta(G_w + G_v)^2 \left(\sqrt{T + \frac{T^2}{n}}\right). \]
Similar to Part a) we have

$$
\Delta^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T) = O\left( \eta (G_w + G_v)^2 \left( \sqrt{T} + \frac{T}{n} \right) + \frac{\eta(G_w^2 + G_v^2)}{2} + \frac{D_w^2 + D_v^2}{2\eta T} + \frac{D_w G_w + D_v G_v}{\sqrt{mT}} + \frac{\eta (G_w^2 + G_v^2) T d \log(1/\delta)}{n^2 \epsilon^2} \right).
$$

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

$$
\Delta^w(\bar{\mathbf{w}}_T, \bar{\mathbf{v}}_T)
= O\left( \max\{G_w^2 + G_v^2, (G_w + G_v)^2, D_w^2 + D_v^2, D_w G_w + D_v G_v\} \max\left\{ \frac{1}{\sqrt{n}}, \frac{\sqrt{d \log(1/\delta)}}{n \epsilon} \right\} \right).
$$

**Remark 3.3** The same optimal utility bound $O\left( \max\left\{ \frac{1}{\sqrt{n}}, \frac{\sqrt{d \log(1/\delta)}}{n \epsilon} \right\} \right)$ was claimed in [Boob and Guzmán [2021]]. Yet our results also possess two theoretical gains compared to theirs. Firstly, when the smoothness assumption holds, Part a) in our Theorem 3.2 shows the optimal utility with $T = O(n)$ iterations and $O(n^{3/2})$ gradient computations by Remark 3.1, while their single-looped algorithm (Algorithm 1 there) requires $O(n^2)$ gradient computations in their Theorem 5.4. They further improved the gradient complexity to $O(n^{3/2} \log(n))$ 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]].

### 3.2.2 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., $R(\mathbf{w}_T) - \min_{\mathbf{w} \in W} R(\mathbf{w})$, where $\mathbf{w}_T$ is the last iterate of Algorithm 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 $\mathbf{w} \in W$ and $\mathbf{v} \in V$,

$$
F_S(\hat{\mathbf{w}}_S, \mathbf{v}) \leq F_S(\hat{\mathbf{w}}_S, \hat{\mathbf{v}}_S) \leq F_S(\mathbf{w}, \hat{\mathbf{v}}_S) = f(\mathbf{w}) - \min_{\mathbf{v} \in V} f(\mathbf{v}) + F_S(\mathbf{w}, \mathbf{v})
$$
To estimate the primal excess risk, we denote \( R^*_S = \min_{w \in W} R_S(w), \) and \( R^* = \min_{w \in W} R(w). \) Then, for any \( w^* \in \arg \min_w R(w) \) we have the error decomposition:

\[
\mathbb{E}[R(w_T) - R^*] = \mathbb{E}[R(w_T) - R_S(w_T)] + \mathbb{E}[R_S(w_T) - R^*_S] \\
+ \mathbb{E}[R^*_S - R_S(w^*)] + \mathbb{E}[R_S(w^*) - R(w^*)] \\
\leq \mathbb{E}[R(w_T) - R_S(w_T)] + \mathbb{E}[R_S(w^*) - R(w^*)] + \mathbb{E}[R_S(w_T) - R^*_S],
\]

where the last inequality follows from the fact that \( R^*_S - R_S(w^*) \leq 0 \) since it holds \( R^*_S = \min_{w \in W} R_S(w). \) The term \( \mathbb{E}[R_S(w_T) - R^*_S] \) 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 \( \mathbb{E}[R(w_T) - R_S(w_T)] + \mathbb{E}[R_S(w^*) - R(w^*)] \) 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.

The next several statements will characterize the primal empirical risk of DP-SGDA under the PL-SC assumption.

**Lemma 3.5 (Lin et al. [2020])** Assume Assumption 2.2 holds and \( F_S(w, \cdot) \) is \( \rho \)-strongly concave. Assume \( V \) is a convex and bounded set. Then the function \( R_S(w) \) is \( L + L^2/\rho \)-smooth and \( \nabla R_S(w) = \nabla_w F_S(w, \hat{v}_S(w)) \), where \( \hat{v}_S(w) = \arg \max_{v \in V} F_S(w, v) \). And \( \hat{v}_S(w) \) is \( L/\rho \) Lipschitz continuous.

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

**Lemma 3.6** Assume Assumption 2.2 holds. Assume \( F_S(\cdot, v) \) satisfies PL condition with constant \( \mu \) and \( F_S(w, \cdot) \) is \( \rho \)-strongly concave. Then the function \( R_S(w) \) satisfies the PL condition with \( \mu \).

**Proof:** From Lemma 3.5 \( \| \nabla R_S(w) \|_2^2 = \| \nabla_w F_S(w, \hat{v}_S(w)) \|_2^2. \) Since \( F_S \) satisfies PL condition with constant \( \mu \), we get

\[
\| \nabla R_S(w) \|_2^2 \geq 2\mu (F_S(w, \hat{v}_S(w)) - \min_{w' \in V} F_S(w', \hat{v}_S(w))).
\]
Also, since \( F_S(w', \hat{v}_S(w)) \leq \max_{v \in V} F_S(w', v) \), we have

\[
\min_{w' \in W} F_S(w', \hat{v}_S(w)) \leq \min_{w' \in W} \max_{v \in V} F_S(w', v) = \min_{w' \in W} R_S(w')
\]  

(3.13)

Combining equation (3.12) and (3.13), we have

\[
\|\nabla R_S(w)\|_2^2 \geq 2\mu (R_S(w) - \min_{w' \in W} R_S(w')).
\]

Now we present two key lemmas for the convergence analysis. The next lemma characterizes the descent behavior of \( R_S(w_t) \).

**Lemma 3.7** Assume Assumption 3.1 and Assumption 2.2 hold. Assume \( F_S(\cdot, v) \) satisfies the \( \mu \)-PL condition and \( F_S(w, \cdot) \) is \( \rho \)-strongly concave. For Algorithm 1, the iterates \( \{w_t, v_t\}_{t \in [T]} \) satisfies the following inequality

\[
E[R_S(w_{t+1}) - R^*_S] \leq (1 - \mu \eta_{w,t})E[R_S(w_t) - R^*_S] + \frac{L^2 \eta_{w,t}}{2} E[\|\hat{v}_S(w_t) - v_t\|_2^2]
\]

\[
+ \frac{(L + L^2/\rho) \eta_{w,t}^2}{2} (B_w^2/m + d\sigma_w^2).
\]

**Proof:** Because \( R_S \) is \( L + L^2/\rho \)-smooth by Lemma 3.5, we have

\[
R_S(w_{t+1}) - R^*_S \leq R_S(w_t) - R^*_S + \langle \nabla R_S(w_t), w_{t+1} - w_t \rangle + \frac{L + L^2/\rho}{2} \|w_{t+1} - w_t\|_2^2
\]

\[
=R_S(w_t) - R^*_S - \eta_{w,t} \nabla R_S(w_t), \frac{1}{m} \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{ij}) + \xi_t
\]

\[
+ \frac{(L + L^2/\rho) \eta_{w,t}^2}{2} \|m \sum_{j=1}^m \nabla_w f(w_t, v_t; z_{ij}) + \xi_t\|_2^2.
\]

We denote \( E_t \) as the conditional expectation of given \( w_t \) and \( v_t \). Taking this conditional
Assume Assumption 3.1 and Assumption 2.2 hold. Assume $E$ satisfies PL condition with constant $\mu$. Taking expectation of both sides, we get

$$
\mathbb{E}_t[R_S(w_{t+1}) - R^*_S] = R_S(w_t) - R^*_S - \eta_{w,t}\langle \nabla R_S(w_t), \nabla F_S(w_t, v_t) \rangle
$$

$$
+ \frac{(L + L^2/\rho)\eta^2_{w,t}}{2} \left\| \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w F_S(w_t, v_t) + \nabla F_S(w_t, v_t) - \xi_t \right\|_2^2
$$

$$
\leq R_S(w_t) - R^*_S - \eta_{w,t}\langle \nabla R_S(w_t), \nabla F_S(w_t, v_t) \rangle
$$

$$
+ \frac{(L + L^2/\rho)\eta^2_{w,t}}{2} \left\| \nabla_w F_S(w_t, v_t) \right\|_2^2 + \frac{(L + L^2/\rho)\eta^2_{w,t}}{2} \left\| \nabla R_S(w_t) - \nabla F_S(w_t, v_t) \right\|_2^2
$$

$$
+ \frac{(L + L^2/\rho)\eta^2_{w,t}}{2} \left( \frac{B^2_w}{m} + d\sigma^2_w \right),
$$

where first inequality holds since

$$
\mathbb{E}_t\left[ \left\| \frac{1}{m} \sum_{j=1}^{m} \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w F_S(w_t, v_t) \right\|_2^2 \right] = \frac{1}{m} \sum_{j=1}^{m} \mathbb{E}_t\left[ \left\| \nabla_w f(w_t, v_t; z_{ij}) - \nabla_w F_S(w_t, v_t) \right\|_2^2 \right] \leq \frac{B^2_w}{m} \quad (3.14)
$$

and $\mathbb{E}_t[\|\xi_t\|^2_2] = d_t\sigma^2_w \leq d\sigma^2_w$. And the last inequality we use $\eta_w \leq 1/(L + L^2/\rho)$. Because $R_S$ satisfies PL condition with $\mu$ by Lemma 3.6, we have

$$
\mathbb{E}_t[R_S(w_{t+1}) - R^*_S] \leq (1 - \mu\eta_{w,t})(R_S(w_t) - R^*_S) + \frac{\eta_{w,t}}{2} \left\| \nabla R_S(w_t) - \nabla F_S(w_t, v_t) \right\|_2^2
$$

$$
+ \frac{(L + L^2/\rho)\eta^2_{w,t}}{2} \left( \frac{B^2_w}{m} + d\sigma^2_w \right)
$$

$$
\leq (1 - \mu\eta_{w,t})(R_S(w_t) - R^*_S) + \frac{L^2\eta_{w,t}}{2} \left\| \hat{v}_S(w_t) - v_t \right\|_2^2
$$

$$
+ \frac{(L + L^2/\rho)\eta^2_{w,t}}{2} \left( \frac{B^2_w}{m} + d\sigma^2_w \right),
$$

where the second inequality we use $F_S$ is $L$-smooth. Now taking expectation of both sides yields the claimed bound.

The next lemma characterizes the descent behavior of $v_t$.

**Lemma 3.8** Assume Assumption 3.1 and Assumption 2.2 hold. Assume $F_S(\cdot, v)$ satisfies PL condition with constant $\mu$ and $F_S(w, \cdot)$ is $\rho$-strongly concave. Denote $\hat{v}_S(w) =$
arg max  \_\_  \_F_S(w, v). For Algorithm \ref{alg:1} and any \( \epsilon > 0 \), the iterates \( \{w_t, v_t\} \) satisfies the following inequality

\[
\mathbb{E}[\|v_{t+1} - \hat{v}_S(w_{t+1})\|^2_2] \leq ((1 + \frac{1}{\epsilon})2L^4/\rho^2\eta_{w,t}^2 \epsilon + (1 + \epsilon)(1 - \rho\eta_{v,t}))(1 + \frac{1}{\epsilon})2L^2/\rho^2(L + 2L/\rho)\eta_{w,t}^2\mathbb{E}[R_S(w_t) - R_S^*] + (1 + \epsilon)\eta_{v,t}^2(B_w^2 + m + d\sigma_w^2).
\]

**Proof:** By Young’s inequality, we have

\[
\|v_{t+1} - \hat{v}_S(w_{t+1})\|^2_2 \leq (1 + \epsilon)\|v_{t+1} - \hat{v}_S(w_t)\|^2_2 + (1 + \epsilon)\|\hat{v}_S(w_t) - \hat{v}_S(w_{t+1})\|^2_2.
\]

For the term \( \|\hat{v}_S(w_t) - \hat{v}_S(w_{t+1})\|^2_2 \), since \( \hat{v}_S(\cdot) \) is \( L/\rho \)-Lipschitz by Lemma \ref{lem:3.5}, taking conditional expectation, we have

\[
\mathbb{E}_t[\|\hat{v}_S(w_{t+1}) - \hat{v}_S(w_t)\|^2_2] \leq L^2/\rho^2\mathbb{E}_t[\|w_{t+1} - w_t\|^2_2]
\]

\[
= L^2/\rho^2\eta_{w,t}^2\mathbb{E}_t[\|\frac{1}{m}\sum_{j=1}^m \nabla_w f(w_t, v_t; z_{j,t}) + \zeta_t\|^2_2]
\]

\[
\leq L^2/\rho^2\eta_{w,t}^2\|\nabla_w F_S(w_t, v_t)\|^2_2 + L^2/\rho^2\eta_{w,t}^2(B_w^2 + m + d\sigma_w^2)
\]

\[
\leq 2L^2/\rho^2\eta_{w,t}^2\|\nabla R_S(w_t) - \nabla_w F_S(w_t, v_t)\|^2_2 + L^2/\rho^2\eta_{w,t}^2\|\nabla R_S(w_t)\|^2_2
\]

\[
+ L^2/\rho^2\eta_{w,t}^2(B_w^2 + m + d\sigma_w^2)
\]

\[
\leq 2L^4/\rho^2\eta_{w,t}^2\|\hat{v}_S(w_t) - v_t\|^2_2 + 2L^2/\rho^2\eta_{w,t}^2\|\nabla R_S(w_t)\|^2_2 + L^2/\rho^2\eta_{w,t}^2(B_w^2 + m + d\sigma_w^2),
\]

where the last step uses the fact that \( F_S \) is \( L \)-smooth. Because \( R_S \) is \( L + 2L/\rho \)-smooth by Lemma \ref{lem:3.5}, we have \( \|\nabla R_S(w_t)\|^2_2 \leq R_S(w_t) - R_S^* \). Therefore

\[
\mathbb{E}_t[\|\hat{v}_S(w_{t+1}) - \hat{v}_S(w_t)\|^2_2] \leq 2L^4/\rho^2\eta_{w,t}^2\|\hat{v}_S(w_t) - v_t\|^2_2
\]

\[
+ 4L^2/\rho^2(L + 2L/\rho)\eta_{w,t}^2(R_S(w_t) - R_S(w^*)) + L^2/\rho^2\eta_{w,t}^2(B_w^2 + m + d\sigma_w^2). \tag{3.15}
\]
For the term $\|v_{t+1} - \hat{v}_S(w_t)\|_2^2$, by the contraction of projection, we have

$$
\mathbb{E}_t[\|v_{t+1} - \hat{v}_S(w_t)\|_2^2] \\
\leq \mathbb{E}_t[\|v_t + \eta_{t,2}(\frac{1}{m} \sum_{j=1}^m \nabla_v f(w_t, v_i, z_i) + \zeta_t) - \hat{v}_S(w_t)\|_2^2] \\
\leq \|v_t - \hat{v}_S(w_t)\|_2^2 + 2\eta_{t,2} \mathbb{E}_t[\langle v_t - \hat{v}_S(w_t), \nabla_v F_S(w_t, v_t) \rangle + \eta_{t,2}^2 \|\nabla_v F_S(w_t, v_t)\|_2^2 + \eta_{t,2}^2 (\frac{B_v^2}{m} + d\sigma_w^2) \\
\leq (1 - \rho \eta_{t,2}) \|v_t - \hat{v}_S(w_t)\|_2^2 + 2\eta_{t,2} (F_S(w_t, v_t) - F_S(w_t, \hat{v}_S(w_t)) + \eta_{t,2}^2 \|\nabla_v F_S(w_t, v_t)\|_2^2 \\
+ \eta_{t,2}^2 (\frac{B_v^2}{m} + d\sigma_w^2),
$$

where the third inequality we use the $F_S(w, \cdot)$ is $\rho$-strongly concave. Since $F_S$ is $L$-smooth, by choosing $\eta_{t,2} \leq 1/L$, we have

$$
\mathbb{E}_t[\|v_{t+1} - \hat{v}_S(w_t)\|_2^2] \leq (1 - \rho \eta_{t,2}) \|v_t - \hat{v}_S(w_t)\|_2^2 - \frac{\eta_{t,2}^2}{L} \|\nabla_v F_S(w_t, v_t)\|_2^2 \\
+ \eta_{t,2}^2 \|\nabla_v F_S(w_t, v_t)\|_2^2 + \eta_{t,2}^2 (\frac{B_v^2}{m} + d\sigma_w^2) \\
\leq (1 - \rho \eta_{t,2}) \|v_t - \hat{v}_S(w_t)\|_2^2 + \eta_{t,2}^2 (\frac{B_v^2}{m} + d\sigma_w^2). \quad (3.16)
$$

Combining (3.16) and (3.15) we have

$$
\mathbb{E}_t[\|v_{t+1} - \hat{v}_S(w_{t+1})\|_2^2] \leq ((1 + \frac{1}{\epsilon})2L^4/\rho^2 \eta_{w,2}^2 + (1 + \epsilon)(1 - \rho \eta_{t,2})) \|v_t - \hat{v}_S(w_t)\|_2^2 \\
+ (1 + \frac{1}{\epsilon}) \|w_{t+1} - w_t\|_2^2 (\frac{B_w^2}{m} + d\sigma_w^2) \\
+ (1 + \frac{1}{\epsilon})4L^2/\rho^2 (L + L^2/\rho) \eta_{w,t}^2 (R_S(w_t) - R_S(w^*)) \\
+ (1 + \epsilon) \eta_{t,2}^2 (\frac{B_v^2}{m} + d\sigma_w^2).
$$

Taking expectation on both sides yields the desired bound. \qed
Lemma 3.9 Assume Assumption 3.1 and Assumption 2.2 hold. Assume \(F_S(\cdot, v)\) satisfies PL condition with constant \(\mu\) and \(F_S(w, \cdot)\) is \(\rho\)-strongly concave. Define \(a_t = \mathbb{E}[R_S(w_t) - R_S(w^*)]\) and \(b_t = \mathbb{E}[\|\hat{v}_S(w_t) - v_t\|^2]\). For Algorithm 1, if \(\eta_{w,t} \leq 1/(L+L^2/\rho)\) and \(\eta_{v,t} \leq 1/L\), then for any non-increasing sequence \(\{\lambda_t > 0\}\) and \(\epsilon > 0\), the iterates \([w_t, v_t]_{t \in [T]}\) satisfy the following inequality

\[
a_{t+1} + \lambda_{t+1}b_{t+1} \leq k_{1,t}a_t + k_{2,t}\lambda_tb_t + \frac{(L + L^2/\rho)\eta_{w,t}^2}{2} \left( \frac{B_w^2}{m} + d\sigma^2_w \right) + 2(1 + \frac{1}{\epsilon})\lambda_tL^2/\rho^2\eta_{w,t}^2 \left( \frac{B_w^2}{m} + d\sigma^2_w \right) + \lambda_t(1 + \epsilon)\eta_{v,t}^2 \left( \frac{B_v^2}{m} + d\sigma^2_v \right),
\]

where

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

\[
k_{2,t} = \frac{L^2\eta_{w,t}}{2\lambda_t} + (1 + \epsilon)(1 - \rho\eta_{v,t}) + (1 + \frac{1}{\epsilon})2L^4/\rho^2\eta_{w,t}^2.
\]
Proof: Combining Lemma 3.7 and Lemma 3.8 we have for any $\lambda_{t+1} > 0$, we have

\[ a_{t+1} + \lambda_{t+1} b_{t+1} \leq ((1 - \mu_t \eta_{w,t}) + \lambda_{t+1} (1 + \frac{1}{\epsilon}) 4 L^2 / \rho^2 (L + L^2 / \rho) \eta_{w,t}^2) a_t \]

\[ + \left( \frac{L^2 \eta_{w,t}}{2} + \lambda_{t+1} (1 + \epsilon) (1 - \rho \eta_{v,t}) + \lambda_{t+1} (1 + \frac{1}{\epsilon}) 2 L^4 / \rho^2 \eta_{w,t}^2 \right) b_t \]

\[ + \frac{(L + L^2 / \rho) \eta_{w,t}^2}{2} \left( \frac{B_w^2}{m} + d \sigma_w^2 \right) + 2(1 + \frac{1}{\epsilon}) \lambda_{t+1} L^2 / \rho^2 \eta_{w,t}^2 \left( \frac{B_w^2}{m} + d \sigma_w^2 \right) \]

\[ + \lambda_{t+1} (1 + \epsilon) \eta_{v,t}^2 \left( \frac{B_v^2}{m} + d \sigma_v^2 \right) \]

\[ \leq ((1 - \mu_t \eta_{w,t}) + \lambda_t (1 + \frac{1}{\epsilon}) 4 L^2 / \rho^2 (L + L^2 / \rho) \eta_{w,t}^2) a_t \]

\[ + \lambda_t \left( \frac{L^2 \eta_{w,t}}{2 \lambda_t} + (1 + \epsilon) (1 - \rho \eta_{v,t}) + (1 + \frac{1}{\epsilon}) 2 L^4 / \rho^2 \eta_{w,t}^2 \right) b_t \]

\[ + \frac{(L + L^2 / \rho) \eta_{w,t}^2}{2} \left( \frac{B_w^2}{m} + d \sigma_w^2 \right) + 2(1 + \frac{1}{\epsilon}) \lambda_t L^2 / \rho^2 \eta_{w,t}^2 \left( \frac{B_w^2}{m} + d \sigma_w^2 \right) \]

\[ + \lambda_t (1 + \epsilon) \eta_{v,t}^2 \left( \frac{B_v^2}{m} + d \sigma_v^2 \right). \]

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

\textbf{Theorem 3.3} Assume Assumptions Assumption 2.1 and Assumption 3.1 hold true, and the function $F_S(w, \cdot)$ is $\rho$-strongly concave and $F_S(\cdot, v)$ satisfies $\mu$-PL condition. Assume $V$ is bounded. Let $\kappa = L / \rho$. If we choose $\eta_{w,t} \asymp \frac{1}{\mu}$ and $\eta_{v,t} \asymp \frac{\kappa^{3.5}}{\mu^{2.5} T^{2/3}}$, then

\[ \mathbb{E}[R_S(w_{T+1}) - R^*_S] = O \left( \frac{\kappa^{3.5}}{\mu^{2.5}} \left( \frac{1}{m} + d(\sigma_w^2 + \sigma_v^2) \right) \right). \]

Proof: Since $\eta_{v,t} \leq 1 / L$, we can pick $\epsilon = \frac{\rho \eta_{w,t}}{2 (1 - \rho \eta_{v,t})}$. Then we have $(1 + \epsilon)(1 - \rho \eta_{v,t}) = \ldots$
If we choose $\lambda_t = \frac{4L^2\eta_{w,t}}{\rho \eta_{v,t}}$ and $\eta_{w,t} \leq \min\{\frac{\sqrt{\pi}}{8\kappa^2 \sqrt{L + L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\} \eta_{v,t}$, then further we have $k_{1,t} \leq 1 - \frac{\mu \eta_{w,t}}{2}$ and $k_{2,t} \leq 1 - \frac{\rho \eta_{v,t}}{2}$. By Lemma 3.9 we have

$$a_{t+1} + \lambda_{t+1} b_{t+1} \leq (1 - \min\{\frac{\mu}{2}, L^2\}) \eta_{w,t}(a_t + \lambda_t b_t) + \frac{(L + L^2/\rho)\eta_{w,t}^2}{2m} (B_w^2 m + d\sigma_w^2)$$

$$+ \frac{16L^4/\rho^3 \eta_{w,t}^3}{\rho \eta_{v,t}} (B_w^2 m + d\sigma_w^2) + \frac{4L^2(2 - \rho \eta_{v,t})\eta_{w,t}^2}{2\rho(1 - \rho \eta_{v,t})} (B_v^2 m + d\sigma_v^2)$$

$$\leq (1 - \frac{\mu \eta_{w,t}}{2})(a_t + \lambda_t b_t) + \frac{(L + L^2/\rho)\eta_{w,t}^2}{2m} (B_w^2 m + d\sigma_w^2)$$

$$+ \frac{16L^4/\rho^3 \eta_{w,t}^3}{\rho \eta_{v,t}^2} (B_w^2 m + d\sigma_w^2) + \frac{4L^2(2 - \rho \eta_{v,t})\eta_{w,t}^2}{2\rho(1 - \rho \eta_{v,t})} (B_v^2 m + d\sigma_v^2),$$

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

$$t(a_{t+1} + \lambda_{t+1} b_{t+1}) \leq (t - 1)(a_t + \lambda_t b_t) + \frac{2(L + L^2/\rho)}{\mu t} (B_w^2 m + d\sigma_w^2)$$

$$+ \frac{32L^4/\rho^3 \min\{\frac{\sqrt{\pi}}{8\kappa^2 \sqrt{L + L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}^2}{\mu t^{2/3}} (B_w^2 m + d\sigma_w^2)$$

$$+ \frac{16L^2 \max\{8\kappa^2 \sqrt{(L + L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}}{2\mu t^{2/3}} (B_v^2 m + d\sigma_v^2).$$

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

$$T(a_{T+1} + \lambda_{T+1} b_{T+1}) \leq \frac{2(L + L^2/\rho)}{\mu} (B_w^2 m + d\sigma_w^2) \log(T)$$

$$+ \frac{32L^4/\rho^3 \min\{\frac{\sqrt{\pi}}{8\kappa^2 \sqrt{L + L^2/\rho}}, \frac{1}{4\sqrt{2}\kappa^2}\}^2}{\mu t^{2/3}} (B_w^2 m + d\sigma_w^2) T^{1/3}$$

$$+ \frac{16L^2 \max\{8\kappa^2 \sqrt{(L + L^2/\rho)/\mu}, 4\sqrt{2}\kappa^2\}}{2\mu t^{2/3}} (B_v^2 m + d\sigma_v^2) T^{1/3}.$$
Consequently,
\[
\mathbb{E}[R_S(w_{T+1}) - R_S^*] \leq \alpha_{T+1} + \lambda_{T+1} b_{T+1} \\
\leq \frac{2(L + L^2 / \rho)(B_w^2 / m + d \sigma_w^2) \log(T)}{\mu^2} \\
+ \frac{32(B_w^2 / m + d \sigma_w^2)L^4 / \rho^3 \min\left\{ \frac{\sqrt{\mu}}{8 \kappa \sqrt{L + L^2 / \rho}}, \frac{1}{4 \sqrt{2} \kappa^2} \right\}^2 1}{T^{2/3}} \\
+ \frac{16(B_w^2 / m + d \sigma_w^2) L^2 \max\{8 \kappa^2 \sqrt{(L + L^2 / \rho) / \mu}, 4 \sqrt{2} \kappa^2 \} 1}{2 \mu^2 \rho} T^{2/3} \quad (3.17)
\]

Therefore, the following equation holds by calling \( \kappa = L / \rho \).
\[
\mathbb{E}[R_S(w_{T+1}) - R_S^*] = \mathcal{O}\left( \min\left\{ \frac{1}{L}, \frac{1}{\mu} \right\} \left( \frac{B_w^2 / m + d \sigma_w^2}{T^{2/3}} \right) + \max\left\{ 1, \sqrt{\frac{L \kappa^3}{\mu}} \right\} \left( \frac{B_v / m + d \sigma_v^2}{T^{2/3}} \right) \right).
\quad (3.18)
\]

The result in Theorem 3.3 follows by observing \( \max\left\{ 1, \sqrt{\frac{L \kappa^3}{\mu}} \right\} \frac{L \kappa^3}{\mu^2} \geq \min\left\{ \frac{1}{L}, \frac{1}{\mu} \right\} \). Substituting the values of \( \sigma_w, \sigma_v \), i.e., \( \sigma_w = \frac{c_3 G_w \sqrt{T \log(\frac{1}{\kappa})}}{\kappa \sqrt{\mu}} \) and \( \sigma_v = \frac{c_3 G_v \sqrt{T \log(\frac{1}{\kappa})}}{\kappa \sqrt{\mu}} \), into (3.18) yields the desired estimation.

In the non-private setting, i.e. \( \sigma_w = \sigma_v = 0 \), Theorem 3.3 implies that the convergence rate in terms of the primal empirical risk is of the order \( \mathcal{O}\left( \frac{\kappa^{3.5}}{\mu^2 T^{2/3}} \right) \), 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 \( \mathbb{E}_\tau[\|
abla R_S(w_\tau)\|^2] \) where \( \tau \) is chosen uniformly at random from the set \{1, 2, ..., T\}. Our analysis is much more involved since it proves the global convergence of the last iterate \( w_T \). Our main idea is to prove the coupled recursive inequalities for two terms, i.e., \( a_t = R_S(w_t) - R_S^* \) and \( b_t = \|v_t - \hat{v}_S(w_t)\|^2 \) where \( \hat{v}_S(w_t) = \arg \max_{v \in V} F_S(w_t, v) \), and then carefully derive the the convergence rate for \( a_t + \lambda_t b_t \) by choosing \( \lambda_t \) appropriately. The convergence rate and its proof can be of interest in their own right.

We present the bound for the generalization error which is proved again using the stability approach. We first focus on to the generalization error \( \mathbb{E}[R(w_T) - R_S(w_T)] \). We begin with a discussion of the saddle points. While the saddle point \( (\hat{w}_S, \hat{v}_S) \) may not
be unique, \( \hat{v}_S \) must be unique if \( F_S(w, v) \) is strongly-concave in \( v \). Therefore, we can define \( \pi_S(w) \) the projection of \( w \) to the set of saddle points, as \( \Omega_S = \{ \hat{w}_S : (\hat{w}_S, \hat{v}_S) \in \arg \min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} F_S(w, v) \} = \{ \hat{w}_S : \hat{w}_S \in \arg \min_{w \in \mathcal{W}} F_S(w, \hat{v}_S) \} \).

**Proposition 3.1** Assume \( F_S(w, \cdot) \) is \( \rho \)-strongly concave with \( \rho > 0 \). Let \((\hat{w}_S, \hat{v}_S)\) and \((\hat{w}_S', \hat{v}_S')\) be two saddle points of \( F_S \). Then we have \( \hat{v}_S = \hat{v}_S' \).

**Proof:** Given \( \hat{w}_S \), by the strong concavity, we have

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

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

\[
F_S(\hat{w}_S, \hat{v}_S) \geq F_S(\hat{w}_S, \hat{v}_S') + \frac{\rho}{2} \| \hat{v}_S - \hat{v}_S' \|_2^2 \geq F_S(\hat{w}_S', \hat{v}_S') + \frac{\rho}{2} \| \hat{v}_S - \hat{v}_S' \|_2^2, \tag{3.19}
\]

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

\[
F_S(\hat{w}_S', \hat{v}_S') \geq F_S(\hat{w}_S', \hat{v}_S) + \frac{\rho}{2} \| \hat{v}_S - \hat{v}_S' \|_2^2. \tag{3.20}
\]

Adding (3.19) and (3.20) together implies that \( \rho \| \hat{v}_S - \hat{v}_S' \|_2^2 \leq 0 \). This implies \( \hat{v}_S = \hat{v}_S' \) which completes the proof. \( \square \)

Recall that \( \pi_S : \mathcal{W} \to \mathcal{W} \) is the projection onto the set of saddle points \( \Omega_S = \{ \hat{w}_S : (\hat{w}_S, \hat{v}_S) \in \arg \min_{w \in \mathcal{W}} \max_{v \in \mathcal{V}} F_S(w, v) \} \), i.e. \( \pi_S(w) = \arg \min_{\hat{w}_S \in \Omega_S} \frac{1}{2} \| w - \hat{w}_S \|_2^2 \). Proposition 3.1 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 3.10** Suppose the function \( F_S(\cdot, v) \) satisfies \( \mu \)-PL condition. Then \( F_S \) satisfies the QG condition with respect to \( w \) with constant \( 4\mu \), i.e.

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

99
Recall that \( w_T \) is the iterate of DP-SGDA at time \( T \) based on the training data \( S \). Likewise, we denote by \( w'_T \) 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 3.2** For the (randomized) algorithm DP-SGDA, assume that \( \pi_S'(\pi_S(w_T)) = \pi_{S'}(w'_T) \) for any neighboring sets \( S \) and \( S' \).

Assumption 3.2 was introduced in Charles and Papailiopoulos [2018] for studying the stability of SGD in the non-convex case which only involves the minimization over \( w \). In our case, Assumption 3.2 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. \( \Omega_S \) and \( \Omega_{S'} \) do not change too much. Since our algorithm satisfies \((\varepsilon, \delta)\)-DP it means that the distributions of \( w_T \) and \( w'_T \) generated from two neighboring sets \( S \) and \( S' \) are “close”, which indicates \( \sup_{S, S'} \| \pi_{S'}(\pi_S(w_T)) - \pi_{S'}(w'_T) \|_2 \) can be small. Proving such statement serves as an interesting open problem.

With the help of Assumption 3.2 and the preceding lemmas, we can derive the uniform argument stability.

**Lemma 3.11** Assume Assumptions 2.1, 2.2 and 3.2 hold. Assume \( F_S(\cdot, v) \) satisfies PL condition with constant \( \mu \) and \( F_S(w, \cdot) \) is \( \rho \)-strongly concave. Let \( A \) be a randomized algorithm. If for any \( S \), \( E[\| A_w(S) - \pi_S(A_w(S)) \|_2] = O(\varepsilon_A) \), then we have

\[
E[\| A_w(S) - A_w(S') \|_2] \leq O(\varepsilon_A) + \frac{1}{n} \sqrt{\frac{G_w}{4\mu^2} + \frac{G_v^2}{\rho \mu}}.
\]

**Proof:** Let \((\pi_S(A_w(S)), \hat{v}_S) \in \arg\min_w \max_v F_S(w, v) \) and \((\pi_{S'}(A_w(S'))), \hat{v}_{S'} \) defined in the similar way. By triangle inequality we have

\[
E[\| A_w(S) - A_w(S') \|_2] \leq E[\| A_w(S) - \pi_S(A_w(S)) \|_2] + \| \pi_S(A_w(S)) - \pi_{S'}(A_w(S')) \|_2
\]

\[
+ E[\| A_w(S') - \pi_{S'}(A_w(S')) \|_2]
\]

\[
= \| \pi_S(A_w(S)) - \pi_{S'}(A_w(S')) \|_2 + O(\varepsilon_A).
\]

100
Since $\pi_S(\mathcal{A}_w(S)) \in \arg\min_{w \in W} F_S(w, \hat{v}_S)$ and by Assumption 3.2, we know that $\pi_S(\mathcal{A}_w(S))$ is the closest optimal point of $F_S$ to $\pi_{S'}(\mathcal{A}_w(S'))$. And since $\hat{v}_S$ is fixed, by Lemma 3.10, we have

$$2\mu\|\pi_S(\mathcal{A}_w(S)) - \pi_{S'}(\mathcal{A}_w(S'))\|^2 \leq F_S(\pi_{S'}(\mathcal{A}_w(S'))), \hat{v}_S) - F_S(\pi_S(\mathcal{A}_w(S)), \hat{v}_S).$$

Similarly, we have

$$2\mu\|\pi_S(\mathcal{A}_w(S)) - \pi_{S'}(\mathcal{A}_w(S'))\|^2 \leq F_{S'}(\pi_{S'}(\mathcal{A}_w(S)), \hat{v}_{S'}) - F_{S'}(\pi_{S'}(\mathcal{A}_w(S'))), \hat{v}_{S'}).$$

Summing up the above two inequalities we have

$$4\mu\|\pi_S(\mathcal{A}_w(S)) - \pi_{S'}(\mathcal{A}_w(S'))\|^2 \leq F_S(\pi_{S'}(\mathcal{A}_w(S'))), \hat{v}_S) - F_S(\pi_S(\mathcal{A}_w(S)), \hat{v}_S)
+ F_{S'}(\pi_S(\mathcal{A}_w(S)), \hat{v}_{S'}) - F_{S'}(\pi_{S'}(\mathcal{A}_w(S'))), \hat{v}_{S'}). \quad (3.21)$$

On the other hand, by the $\rho$-strong concavity of $F_S(\cdot, v)$ as well as the notation $\hat{v}_S = \arg\max_{v \in V} F_S(\pi_S(\mathcal{A}_w(S)), v)$, we have

$$\frac{\rho}{2}\|\hat{v}_S - \hat{v}_{S'}\|^2 \leq F_S(\pi_S(\mathcal{A}_w(S)), \hat{v}_S) - F_S(\pi_S(\mathcal{A}_w(S)), \hat{v}_{S'}).$$

Similarly, we have

$$\frac{\rho}{2}\|\hat{v}_S - \hat{v}_{S'}\|^2 \leq F_{S'}(\pi_{S'}(\mathcal{A}_w(S'))), \hat{v}_{S'}) - F_{S'}(\pi_{S'}(\mathcal{A}_w(S')), \hat{v}_{S'}).$$

Summing up the above two inequalities we have

$$\rho\|\hat{v}_S - \hat{v}_{S'}\|^2 \leq F_S(\pi_S(\mathcal{A}_w(S)), \hat{v}_S) - F_S(\pi_S(\mathcal{A}_w(S)), \hat{v}_{S'})
+ F_{S'}(\pi_{S'}(\mathcal{A}_w(S'))), \hat{v}_{S'}) - F_{S'}(\pi_{S'}(\mathcal{A}_w(S')), \hat{v}_{S'}). \quad (3.22)$$
Summing up (3.21) and (3.22) rearranging terms, we have

\[ 4\mu \|\pi_S(A_w(S)) - \pi_{S'}(A_w(S'))\|_2^2 + \rho \|\hat{v}_S - \hat{v}_{S'}\|_2^2 \leq F_S(\pi_{S'}, A_w(S')) - F_S(\pi_S(A_w(S)), \hat{v}_{S'}) - F_S(\pi_S(A_w(S)), \hat{v}_S) \]

\[ = \frac{1}{n} \left( f(\pi_{S'}, A_w(S')), \hat{v}_{S'}; z) - f(\pi_S(A_w(S)), \hat{v}_S; z) + f(\pi_S(A_w(S)), \hat{v}_{S'}; z) - f(\pi_S(A_w(S)), \hat{v}_S; z) \right) \]

\[ \leq \frac{2G_w}{n} \|\pi_S(A_w(S)) - \pi_S(A_w(S'))\|_2 + \frac{2G_v}{n} \|\hat{v}_S - \hat{v}_{S'}\|_2 \]

\[ \leq \frac{1}{n} \sqrt{\frac{G_w^2}{\mu} + \frac{4G_v^2}{\rho}} \times \sqrt{4\mu \|\pi_S(A_w(S)) - \pi_{S'}(A_w(S'))\|_2^2 + \rho \|\hat{v}_S - \hat{v}_{S'}\|_2^2}, \]

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

\[ 2\sqrt{\mu} \|\pi_S(A_w(S)) - \pi_{S'}(A_w(S'))\|_2 \leq \sqrt{4\mu \|\pi_S(A_w(S)) - \pi_{S'}(A_w(S'))\|_2^2 + \rho \|\hat{v}_S - \hat{v}_{S'}\|_2^2} \]

\[ \leq \frac{1}{n} \sqrt{\frac{G_w^2}{\mu} + \frac{4G_v^2}{\rho}}. \]

Now we can state the results on the generalization error.

**Theorem 3.4 (Generalization Error)** Assume Assumptions Assumption 2.1, Assumption 2.2, and Assumption 3.2 hold true, and assume the function \( f(w, \cdot; z) \) is \( \rho \)-strongly concave and \( \rho_S(\cdot, v) \) satisfies \( \mu \)-PL condition. Let \( \kappa = L/\rho \). If \( \mathbb{E}[R_S(w_T) - R^*_S] \leq \varepsilon_T \), then

\[ \mathbb{E}[R(w_T) - R_S(w_T)] \leq (1 + \kappa)G_w \left( \sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n} \sqrt{\frac{G_w^2}{4\mu^2} + \frac{G_v^2}{\rho \mu}} \right), \]

and

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

**Proof:** Since \( R_S \) satisfies \( \mu \)-PL, by Lemma 3.10 and Theorem 3.3 we have

\[ \mathbb{E}[\|w_T - \pi(w_T)\|_2] \leq \sqrt{\mathbb{E}[\|w_T - \pi(w_T)\|_2^2]} \leq \sqrt{\mathbb{E}[\frac{1}{2\mu}(R_S(w_T) - R_S)]} \leq \sqrt{\frac{\varepsilon_T}{2\mu}}. \]
By Lemma 3.11 we have

$$\mathbb{E}[\|w_T - w'_T\|_2] \leq \sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n} \sqrt{\frac{G^2_w}{4\mu^2} + \frac{G^2_{\varphi}}{\rho\mu}}.$$  

By Chapter 2 we have

$$\mathbb{E}[R(w_T) - R_S(w_T)] \leq (1 + \frac{L}{\rho})G_w \left( \sqrt{\frac{\varepsilon_T}{2\mu}} + \frac{1}{n} \sqrt{\frac{G^2_w}{4\mu^2} + \frac{G^2_{\varphi}}{\rho\mu}} \right).$$

Next we prove the second equation. We decompose the term $\mathbb{E}[R_S(w^*) - R(w^*)]$ as

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

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

where $\hat{v}_S = \arg \max_v F_S(w^*, v)$. The second term $\mathbb{E}[F(w^*, \hat{v}_S^*) - F(w^*, v^*)] \leq 0$ since $(w^*, v^*)$ is a saddle point of $F$. Hence it suffices to bound $\mathbb{E}[F_S(w^*, \hat{S}_S) - F_S(w^*, \hat{v}_S^*)]$. Let $S' = \{z'_1, \ldots, z'_n\}$ be drawn independently from $\rho$. For any $i \in [n]$, define $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$. Denote $\hat{v}_{S(i)}^* = \arg \max_{v \in V} F_{S(i)}(w^*, v)$. Then

$$F_S(w^*, \hat{v}_S) - F_S(w^*, \hat{v}_{S(i)})$$

$$= \frac{1}{n} \sum_{j \neq i} \left( f(w^*, \hat{v}_S; z_j) - f(w^*, \hat{v}_{S(i)}; z_j) \right) + \frac{1}{n} \left( f(w^*, \hat{v}_S; z_i) - f(w^*, \hat{v}_{S(i); z_i}) \right)$$

$$= \frac{1}{n} \left( f(w^*, \hat{v}_{S(i); z_i}) - f(w^*, \hat{v}_S; z_i) \right) + \frac{1}{n} \left( f(w^*, \hat{v}_S; z_i) - f(w^*, \hat{v}_{S(i); z_i}) \right)$$

$$+ F_{S(i)}(w^*, \hat{v}_S) - F_{S(i)}(w^*, \hat{v}_{S(i)})$$

$$\leq \frac{1}{n} \left( f(w^*, \hat{v}_{S(i); z_i}) - f(w^*, \hat{v}_S; z_i) \right) + \frac{1}{n} \left( f(w^*, \hat{v}_S; z_i) - f(w^*, \hat{v}_{S(i); z_i}) \right)$$

$$\leq \frac{2G^2_{\varphi}}{n} \|\hat{v}_S - \hat{v}_{S(i)}\|_2^2, \quad (3.23)$$

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

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

103
Combining it with (3.23) we get $\|\hat{v}_S^* - \hat{v}_{S(i)}^*\|_2 \leq 4G_v/(\rho n)$. By Lipschitz continuity, the following inequality holds for any $z$

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

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

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

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

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

\[ \text{Remark 3.4} \] The generalization error bounds given in Theorem 3.4 indicate that if the optimization error $\mathbb{E}[R_S(w_T) - R_S^*]$ 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, 2021a] 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 3.4 and 3.3.

**Theorem 3.5** Under the same assumptions of Theorem 3.4, if we choose $T \asymp n$, $\eta_{w,t} \asymp \frac{1}{\mu t}$ and $\eta_{v,t} \asymp \frac{\kappa^{2.5}}{\mu^{1.75} n^{2/3}}$, then

$$\mathbb{E}[R(w_{T+1}) - R^*] = \mathcal{O}\left(\kappa^{2.75} \left(\frac{1}{n^{1/3}} + \frac{\sqrt{d \log(1/\delta)}}{n^{5/6} \epsilon}\right)\right).$$

**Proof:** For any $w^* \in \arg\min_w R(w)$, recall that we have the error decomposition
which is

$$
\mathbb{E}[R(w_T) - R^*] = \mathbb{E}[R(w_T) - R_S(w_T)] + \mathbb{E}[R_S(w_T) - R^*_S] + \mathbb{E}[R^*_S - R_S(w^*)] \\
+ \mathbb{E}[R_S(w^*) - R(w^*)]
$$

$$
\leq \mathbb{E}[R(w_T) - R_S(w_T)] + \mathbb{E}[R_S(w_T) - R^*_S] + \mathbb{E}[R_S(w^*) - R(w^*)],
$$

where the inequality is by $R^*_S - R_S(w^*) \leq 0$. By Theorem 3.4, we have

$$
\mathbb{E}[R(w_T) - R_S(w_T)] \leq (1 + \frac{L}{\rho})G_w \left( \sqrt{\frac{\varepsilon T}{2\mu}} + \frac{1}{n} \sqrt{\frac{G^2_w}{4\mu^2} + \frac{G^2_v}{\rho \mu}} \right).
$$

And by the second statement, we have

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

We can plug the above two inequalities into (3.11), and get

$$
\mathbb{E}[R(w_T) - R^*] = \mathcal{O}(\varepsilon_T + (1 + \frac{L}{\rho})G_w \left( \sqrt{\frac{\varepsilon T}{2\mu}} + \frac{1}{n} \sqrt{\frac{G^2_w}{4\mu^2} + \frac{G^2_v}{\rho \mu}} \right) + \frac{4G^2_v}{\rho n}).
$$

Now by the choice of $\eta_{w,t}, \eta_{v,t}$, and Theorem 3.3, we have $\varepsilon_T = \mathcal{O}(\frac{\kappa^3}{\mu^2 \sigma^2 + \sigma^2\frac{d(\sigma^2 + \frac{\sigma^2}{2})}{T^{2/3}}})$. Assume $m$ is a constant. Plugging $\varepsilon_T$ into the preceding inequality and letting $T = \mathcal{O}(n)$ yields the second statement. 

$\square$
CHAPTER 4
Stability and Differential Privacy of Non-Smooth Pairwise Learning Problems

4.1 Problem Set-up

Our main results in this chapter are based on Yang et al. [2021a]. Before stating our main results, we first introduce necessary materials and notations. Given a pairwise loss function $f: \mathcal{W} \times \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, we aim to minimize the following population risk

$$F(w) = \mathbb{E}_{z, z'}[f(w, z, z')] ,$$

where $z$ and $z'$ are drawn independently from the population distribution $D$ on $\mathcal{Z}$. The population distribution is often unknown and we only have access to a set of i.i.d. training data $S = \{z_1, z_2, \ldots, z_n\} \in \mathcal{Z}^n$. The task then reduces to minimize the empirical risk

$$\min_{w \in \mathcal{W}} F_S(w) := \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j}^n f(w, z_i, z_j) .$$

Randomized optimization algorithm $A: \mathcal{Z}^n \rightarrow \mathcal{W}$ provides an efficient approach to find an approximate solution to problem (1.3), which takes $S$ as input and produces an output $A(S) \in \mathcal{W}$. The randomized algorithm $A$ here can be either SGD for pairwise learning or its noisy variant for differential privacy. The performance of $A$ is quantified by the excess population risk: $\epsilon_{\text{risk}}(A(S)) = F(A(S)) - \inf_{w \in \mathcal{W}} F(w)$. We can decompose $\epsilon_{\text{risk}}(A(S))$ as follows:

$$\epsilon_{\text{risk}}(A(S)) = [F(A(S)) - F_S(A(S))] + [F_S(w_*) - F(w_*)] + [F_S(A(S)) - F_S(w_*)] , \quad (4.1)$$

where $w_* \in \arg \min_{w \in \mathcal{W}} F(w)$. The first term on the right hand side of (4.1) is called the estimation error. Since $w_*$ is fixed, the term $F_S(w_*) - F(w_*)$ can be trivially handled by the standard Hoeffding inequality [Hoeffding, 1963]. As a comparison, the estimation of the term
\( F(\mathcal{A}(S)) - F_S(\mathcal{A}(S)) \), also called the generalization error, is much more challenging since \( \mathcal{A}(S) \) depends on \( S \). We will develop novel stability analysis to handle this term. The last term \( F_S(\mathcal{A}(S)) - F_S(w^*) \) is called the optimization error and we can bound it by applying optimization theory.

We now introduce some necessary assumptions and definitions.

**Assumption 4.1** The function \( f \) is said to be Lipschitz continuous if there exist \( G > 0 \) such that, for any \( z, z' \) and \( w \in W \),

\[
| f(w; z, z') - f(w'; z, z') | \leq G \| w - w' \|_2.
\]

**Assumption 4.2** The function \( f \) is said to be smooth if there exist \( L > 0 \) such that, for any \( z, z' \) and \( w \in W \),

\[
f(w; z, z') - f(w'; z, z') - \langle \nabla f(w'; z, z'), w - w' \rangle \leq \frac{L}{2} \| w - w' \|_2^2.
\]

**Assumption 4.3** The function \( f \) is said to be strongly convex if there exist \( \alpha > 0 \) such that, for any \( z, z' \) and \( w \in W \),

\[
f(w; z, z') - f(w'; z, z') - \langle \nabla f(w'; z, z'), w - w' \rangle \geq \frac{\alpha}{2} \| w - w' \|_2^2.
\]

The case of \( \alpha = 0 \) is identical to convexity.

### 4.2 SGD for Pairwise Learning

Throughout this chapter, we assume that the (possibly non-smooth) nonnegative loss function \( f \) satisfies Assumption 4.3 and 4.1.

#### 4.2.1 Stability and Excess Risk Analysis

In this subsection, we consider the stability and generalization of the SGD algorithms for pairwise learning. The SGD algorithm is described in Algorithm 2 which has been widely discussed in Lin et al. [2017], Wang et al. [2012], Ying and Zhou [2016]. Note that \( \Pi_W(\cdot) \) is the projection onto the parameter space \( W \) and \( [n] = \{1, \ldots, n\} \). In this subsection, the notation \( \mathcal{A} \) denotes Algorithm 2.

In particular, we will use the uniform argument stability (UAS) [Liu et al. 2017] where its original concept was stated in expectation w.r.t. the internal randomness of \( \mathcal{A} \). We will use its probabilistic version here. Specifically, let \( S = \{z_1, \ldots, z_n\} \) and \( S' = \{z'_1, \ldots, z'_n\} \) be two neighborhood datasets that differ only in one single example. For any \( \gamma \in (0, 1) \), \( \mathcal{A} \) is...
**Algorithm 2** SGD for Pairwise Learning

**Input:** Data set $S = \{z_1, \cdots, z_n\}$, step size $\eta$, number of iterations $T$, initial point $w_1 = 0$ and initial sample $i_1 \in [n]$ from uniform distribution

**for** $t = 1$ to $T$ **do**

Select $i_{t+1} \in [n]$ by uniform distribution

$w_{t+1} = \Pi_W (w_t - \frac{\eta}{t} \sum_{k=1}^t \partial f(w_t, z_{i_{t+1}}, z_{i_k}))$

**end for**

**Output:** $\bar{w}_T = \frac{1}{T} \sum_{t=1}^T w_t$

called $\epsilon_{stab}$-UAS with probability $1 - \gamma$ if for any neighborhood datasets $S$ and $S'$,

$$\mathbb{P}_A[\|A(S) - A(S')\|_2 > \epsilon_{stab}] \leq \gamma.$$  

We emphasize the probability here is taken over the internal randomness of $A$, i.e. the uniform distribution of generating $i_t$’s.

The following theorem states a high-probability UAS result for Algorithm 2 with non-smooth losses. To prove Theorem 4.1 we need the following Chernoff’s bound for a summation of independent Bernoulli random variables [Chernoff, 1952; Wainwright, 2019].

**Lemma 4.1** Let $X_1, \ldots, X_t$ be independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{j=1}^t X_j$ and $\mu = \mathbb{E}[X]$. Then for any $\tilde{\gamma} > 0$, with probability at least $1 - \exp \left( - \mu \tilde{\gamma}^2 / (2 + \tilde{\gamma}) \right)$ we have $X \leq (1 + \tilde{\gamma})\mu$.

**Theorem 4.1** Suppose that we run Algorithm 2 under random selection with replacement for $t$ iterations based on $S$ and $S'$. Then, with probability $1 - \gamma$ w.r.t. the internal randomness of $A$, we have, for any $S$ and $S'$, that

$$\|w_{t+1} - w'_{t+1}\|_2^2 \leq 4\epsilon^2 G^2 \left[ t + \ln^2(et) \times \left( \frac{t}{n} + \ln(1/\gamma) + \sqrt{\frac{t \ln(1/\gamma)}{n}} \right) \right]$$  \hspace{1cm} (4.2)

In particular, if $T \geq n$, then the output of Algorithm 2 is $\epsilon_{stab}$-UAS with high probability where

$$\epsilon_{stab} = \tilde{O} \left( \eta \sqrt{T} + \frac{\eta T \ln(T)}{n} \right).$$

**Proof:** We let $\hat{L}_{t+1}(w_t)$ denote the accumulated loss until $z_{i_{t+1}}$ is revealed. i.e. $\hat{L}_{t+1}(w_t) = \frac{1}{t} \sum_{k=1}^t \ell(w_t, z_{i_{t+1}}, z_{i_k})$. Without loss of generality, assume that $S$ and $S'$ differs in $n$-th po-
sition. Denote \( \delta_{t+1,k} = \partial f(w_t, z_{i_{t+1}}, z_{i_k}) - \partial f(w_t', z_{i_{t+1}}, z_{i_k}) \) and \( \delta_{t+1,k}' = \partial f(w_t, z_{i_{t+1}}, z_{i_k}) - \partial f(w_t', z_{i_{t+1}}, z_{i_k}') \). The following recursive inequality holds

\[
\|w_{t+1} - w_{t+1}'\|^2 = \|w_t - \eta \partial L_{t+1}(w_t) - w_t' + \eta \partial L_{t+1}'(w_t')\|^2 \\
= \|w_t - w_t' - \frac{\eta}{t} \sum_{k=1}^{t} \delta_{t+1,k}'\|^2 \\
\leq \frac{1}{t} \sum_{k=1}^{t} \|w_t - w_t' - \eta \delta_{t+1,k}'\|^2. \tag{4.3}
\]

Now we estimate the term on the right hand side of (4.3) by considering two cases. For the case \( i_{t+1} \neq n \) and \( i_k \neq n \), we have \( z_{i_{t+1}} = z_{i_{t+1}}' \) and \( z_{i_k} = z_{i_k}' \). Then

\[
\|w_t - w_t'\|^2 = \|w_t - w_t'\|^2 + \eta^2 \|\delta_{t+1,k}'\|^2 - 2\eta \langle w_t - w_t', \delta_{t+1,k}' \rangle \\
\leq \|w_t - w_t'\|^2 + 4\eta^2 G^2,
\]

where the last inequality holds because \( f \) is \( G \)-Lipschitz and convex. If \( i_{t+1} = n \) or \( i_k = n \), then \( z_{i_{t+1}} \neq z_{i_{t+1}}' \) or \( z_{i_k} \neq z_{i_k}' \). It follows from the Young’s inequality that for any \( p > 0 \)

\[
\|w_t - w_t'\|^2 \leq (1 + p)\|w_t - w_t'\|^2 + (1 + 1/p)\eta^2 \|\delta_{t+1,k}'\|^2 \\
\leq (1 + p)\|w_t - w_t'\|^2 + 4(1 + 1/p)\eta^2 G^2.
\]

Combining the above two inequalities together and let \( Y_t = \frac{1}{t} \sum_{k=1}^{t} \mathbb{I}_{[i_{t+1} = n \vee i_k = n]} \), we have

\[
\|w_{t+1} - w_{t+1}'\|^2 \leq (1 + pY_t) \|w_t - w_t'\|^2 + 4(1 + Y_t/p)\eta^2 G^2.
\]
Applying the above inequality recursively we have

$$\|w_{t+1} - w'_{t+1}\|_2^2 \leq \sum_{j=1}^{t} \prod_{j,j+1} (1 + pY_t)(4 + 4Y_j/p)\eta_j^2G^2$$

$$\leq \sum_{j=1}^{t} \prod_{j,j+1} (1 + p)Y_t(4 + 4Y_j/p)\eta_j^2G^2$$

$$\leq (1 + p)\sum_{l=1}^{t} Y_l\eta_l^2G^2(4t + 4\sum_{l=1}^{t} Y_l/p), \quad (4.4)$$

where (a) is due to the recursive relation, (b) is due to $1 + ax \leq (1 + a)x$ for $a > 0$ and $x \geq 0$ and (c) inequality is due to $\prod_{i=a}^{b} x^i \leq x^\sum_{i=a}^{b} i$ for $a \geq 1$. We note that $Y_1, \cdots, Y_t$ are dependent variables, but the sum of $Y_l$’s has the following decomposition:

$$t \sum_{l=1} Y_l = \sum_{l=1} \frac{1}{l} \sum_{k=1}^{l} \mathbb{I}_{[i_{l+1}=n \land i_k=n]} \leq \sum_{l=1} \frac{1}{l} \sum_{k=1}^{l} (\mathbb{I}_{[i_{l+1}=n]} + \mathbb{I}_{[i_k=n]})$$

$$= \sum_{l=1} \mathbb{I}_{[i_{l+1}=n]} + \sum_{l=1} \frac{1}{l} \sum_{k=1}^{l} \mathbb{I}_{[i_k=n]}$$

$$\leq \sum_{l=1} \mathbb{I}_{[i_{l+1}=n]} + \ln(t) \sum_{k=1} \mathbb{I}_{[i_k=n]} \leq \ln(et) \sum_{k=1} \mathbb{I}_{[i_k=n]}.$$

Applying Lemma [4.1] with $X_k = \mathbb{I}_{[i_k=n]}$ and $X = \sum_{k=1}^{t} X_k$, with probability at least $1 - \gamma$, we have

$$\sum_{k=1}^{t+1} \mathbb{I}_{[i_k=n]} \leq \frac{t + 1}{n} + \ln(1/\gamma) + \sqrt{\frac{(t + 1) \ln(1/\gamma)}{n}}.$$

For the simplicity of notation, let $c_{\gamma,t} = \ln(et)((t + 1)/n + \ln(1/\gamma) + \sqrt{(t + 1) \ln(1/\gamma)/n})$. Plugging the above inequality back into (4.4), we derive the following inequality with probability $1 - \gamma$

$$\|w_{t+1} - w'_{t+1}\|_2^2 \leq 4\eta^2G^2(1 + p)^{c_{\gamma,t}}(t + c_{\gamma,t}/p).$$

By selecting $p = 1/c_{\gamma,t}$ in the above equality, we have $(1 + p)^{c_{\gamma,t}} \leq e$. Therefore we have proved (4.2) in Theorem [4.1]. Now, since the bound on left hand side of (4.2) is monotonically
increasing, with probability $1 - \gamma$, we have

$$
\|\bar{w}_T - \bar{w}_T'\|_2^2 \leq \frac{1}{T} \sum_{t=1}^{T} \|w_T - w_T'\|_2^2 \\
\leq 4\eta^2 G^2 \left( T + \frac{3T^2 \ln^2(eT) \ln^2(1/\gamma)}{n^2} \right),
$$

where we have used the fact that $T \geq n$. Therefore the $\epsilon_{\text{stab}}$-UAS bound holds by calling the convexity of $f_{2}$-norm.

This bound matches the result in the pointwise learning with non-smooth losses \cite{Bassily2020,Lei2020} up to a logarithmic term of $T$. The proof is motivated by \cite{Lei2020} in the pointwise case but more involved in pairwise learning. Indeed, the key challenge, in comparison with pointwise learning, is that the sub-gradient estimator at the $t$-th step depends not only on the current example $z_{t+1}$ but also on previous examples $\{z_k : k = 1, \cdots, t\}$.

To our best knowledge, \cite{Shen2020} is the only available work which considered the stability of SGD in pairwise learning. However, their work required the loss to be Lipschitz continuous and strongly smooth to ensure the non-expansiveness of the gradient update, which is very critical for the proof of the main results there. The non-smoothness assumption in our paper makes the corresponding gradient update no longer non-expansive, and therefore the arguments in \cite{Shen2020} no longer apply. We bypass this obstacle by a refined control of the expansiveness between adjacent steps. To address this dependence issue, the work of \cite{Shen2020} counts the number $m$ of different examples $z_i \neq z'_i$ encountered by SGD until iteration $t$, which obeys a binomial distribution. In contrast, high-probability analysis here for non-smooth loss is more challenging and involved because directly applying concentration inequality to similar binomial distribution yields an undesired estimation. We overcome this hurdle by decomposing the sub-gradients into sum of $t$ pairs of dependent random variables first, and then upper bound this sum by two sums of independent random variables. From this new decomposition, we can apply the Chernoff-type tail bounds to these two sums of independent random variables to get the desired estimation.

Here we simply assume bounds for $\|w\|$. A simple lemma indicates that if $w$ is bounded, then $f(w, \cdot, \cdot)$ is also bounded. In the subsequent sections, we will characterize the bounds...
for the iterates \( \{w_t\} \) whenever the parameter space \( \mathcal{I} \) is bounded or unbounded.

**Lemma 4.2** Let \( M = \sup_{z,z'} f(0,z,z') \). For any \( w \in W \) that \( w \leq B \) for some \( 0 \leq B < \infty \), then \( \sup_{z,z'} f(w,z,z') \leq M + GB \).

**Proof:** By convexity of \( f \), we have for any \( z,z' \)

\[
f(w,z,z') \leq \sup_{z,z'} f(0,z,z') + \langle w, \partial f(w,z,z') \rangle \leq M + \| w \| \| \partial f(w,z,z') \|_2 \leq M + GB
\]

where the second inequality is due to Cauchy-Schwarz inequality [Schwarz, 1890]. The proof is complete by taking the supremum.

The next theorem in this section is the high probability generalization bound of UAS algorithms in pairwise learning. This theorem is an extension of Theorem 1 in Lei et al. [2020] for generalization bound of uniformly stable algorithms in pairwise learning.

**Theorem 4.2** Suppose \( f \) is nonnegative, convex and \( G \)-Lipschitz. Let \( A \) be a \( \epsilon \)-UAS randomized algorithm for pairwise learning. Suppose the output of \( A \) is bounded by \( B \) and let \( M = \sup_{z,z'} f(0,z,z') \). Then we have for any \( \gamma \in (0,1) \), with probability at least \( 1 - \gamma \) with respect to the sample \( S \) and the internal randomness of \( A \),

\[
F(A(S)) - F_S(A(S)) \leq 4\epsilon + 48\sqrt{6eG\epsilon[\ln(n)] \ln(e/\gamma)} + 12\sqrt{2e(M + GB)}\sqrt{\frac{\ln(e/\gamma)}{\ln(2\ln(n))}}.
\]

**Proof:** According to Theorem 1 in Lei et al. [2020], we only need to check the expected boundedness of \( f(A(S), \cdot, \cdot) \) and the uniform stability of \( A \). For the boundedness part, by Lemma 4.2 we know

\[
\mathbb{E}[f(A(S), z, z')] \leq M + GB
\]

for any \( z, z' \). For the uniform stability, since \( A \) is \( \epsilon \)-UAS, by the Lipschitz continuity of \( f \) we have

\[
\sup_{z,z'}|f(A(S), z, z') - f(A(S'), z, z')| \leq G\|A(S) - A(S')\|_2 \leq G\epsilon.
\]

The proof is complete.

The next corollary is a direct application of Theorem 4.2, which states if UAS holds with high probability, then so is the generalization.
Corollary 4.1 Let $A$ be a randomized algorithm for pairwise learning. If for any $\gamma_0 \in (0, 1)$, we have, for any neighborhood datasets $S, S'$,

$$
P_A\left[\|A(S) - A(S')\|_2 > \epsilon\right] \leq \gamma_0.
$$

Suppose $f$ is nonnegative, convex and $G$-Lipschitz. Suppose the output of $A$ is bounded by $B$ and let $M = \sup_{z, z'} f(0, z, z')$. Then we have for any $\gamma \in (0, 1),$

$$
P_{S, A}\left[F(A(S)) - F_{S}(A(S)) > 4\epsilon + 48\sqrt{6}G\epsilon\ln(n) \ln(e/\gamma) + 12\sqrt{2}e(M + GB)\sqrt{\frac{\ln(e/\gamma)}{n}}\right] 
\leq \gamma + \gamma_0.
$$

Proof: Denote $E = \{A|\|A(S) - A(S')\|_2 > \epsilon\}$ and $F = \{S, A|F(A(S)) - F_{S}(A(S)) > 4\epsilon + 48\sqrt{6}G\epsilon\ln(n) \ln(e/\gamma) + 12\sqrt{2}e(M + GB)\sqrt{\ln(e/\gamma)/n}\}$. Then by assumption we have $P_A[A \in E] \leq \gamma_0$. By Theorem 4.2 for any $\gamma \in (0, 1)$, we have $P_{S, A}[S, A \in F | A \notin E] \leq \gamma$. Then the following identity holds

$$
P_{S, A}[S, A \in F] = P_{S, A}[S, A \in F \cap A \in E] + P_{S, A}[S, A \in F \cap A \notin E] 
= P_{S, A}[S, A \in F | A \in E]P[A \in E] + P_{S, A}[S, A \in F | A \notin E]P[A \notin E] 
\leq \gamma_0 + \gamma.
$$

The proof is completed.

Combining Corollary 4.1 and the stability result in Theorem 4.1, we arrive at the following generalization bound for Algorithm 2.

Corollary 4.2 Suppose $f$ is nonnegative, convex and $G$-Lipschitz. Let $B_T = \|\bar{w}_T\|$ and $M = \sup_{z, z'} f(0, z, z')$. If we run Algorithm 2 for $T \geq n$ iterations under random selection with replacement rule. For any $\gamma \in (0, 1)$, with probability at least $1 - \gamma$ with respect to the sample $S$ and the internal randomness of Algorithm 2, we have

$$
F(\bar{w}_T) - F_{S}(\bar{w}_T) \leq 2\sqrt{\eta G(4 + 48\sqrt{6}G\ln(n) \ln(2e/\gamma))\left(\sqrt{T} + \frac{\sqrt{3T\ln(eT)\ln(2/\gamma)}}{n}\right)}
+ 12\sqrt{2}e(M + GB_T)\sqrt{\frac{\ln(2e/\gamma)}{n}}.
$$

113
Proof: By Theorem 4.1, elementary inequality and the fact that stability is monotonically increasing, we have with probability at least $1 - \gamma/2$,

$$\|\bar{w}_T - \bar{w}_T'\|^2 \leq 4\eta^2 G^2 \left( T + \frac{3T^2 \ln^2(eT) \ln^2(2/\gamma)}{n^2} \right).$$

The proof is completed by convexity of $\| \cdot \|_2$ and applying Theorem 4.2 with probability $1 - \gamma/2$. \qed

The next theorem gives a bound on $F_S(w_*) - F(w_*)$ by Hoeffding inequality of U-statistics [Hoeffding, 1963].

**Theorem 4.3** Suppose $f$ is convex and $G$-Lipschitz. Let $M = \sup_{z,z'} f(0,z,z')$ and $B = \|w_*\|_2$. For any $\gamma \in (0,1)$, with probability at least $1 - \gamma$ with respect to the sample $S$, we have

$$F_S(w_*) - F(w_*) \leq (M + GB) \sqrt{\frac{\ln(1/\gamma)}{n}}.$$

Proof: The result is derived by applying Hoeffding inequality since $f(w_*, z, z') \leq M + GB$ for any $z, z'$ according to Lemma 4.2. \qed

To bound the optimization error, we need the following variant of Rademacher average [Bartlett and Mendelson, 2002].

$$\mathcal{R}_t(f \circ \mathcal{W}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{w \in \mathcal{W}} \frac{1}{t} \sum_{k=1}^{t} \sigma_k f(w, z_i, z_{i_k}) \right]. \quad (4.6)$$

Here $\sigma_k$ are Rademacher random variables taking values in $\{\pm 1\}$ with equal probability $1/2$, and the expectation is taken over $z_i, z_{i_k}$ and $\sigma_k$.

In order to prove the optimization error, we decompose $F_S(\bar{w}_T) - F_S(w_*)$ as in [Kar et al., 2013] and bound each part separately. In particular, recall $\hat{L}_{t+1}(w) = \frac{1}{t} \sum_{k=1}^{t} f(w, z_{i_{t+1}}, z_{i_k})$. We have the following lemmas.

**Lemma 4.3** Let $\mathcal{W}_t = \{ w \in \mathcal{W} \ | \|w\|_2 \leq B_t \}$ and let $M = \sup_{z,z'} f(0,z,z')$. With probab-
ity $1 - \gamma$, we have
\[
\frac{1}{T} \sum_{t=1}^{T} F_S(w_t) - \hat{L}_{t+1}(w_t) \leq \frac{2}{T} \sum_{t=1}^{T} R_t(f \circ W_t) + 3(M + GB_T) \sqrt{\frac{\ln(T/\gamma)}{T}}.
\]

**Proof:** For any $w$, denote $\hat{L}_{t+1}(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} f(w, z_i, z_{ik})$. This allows us to decompose the risk as follows
\[
\frac{1}{T} \sum_{t=1}^{T} F_S(w_t) - \hat{L}_{t+1}(w_t) = \frac{1}{T} \sum_{t=1}^{T} \left( F_S(w_t) - \tilde{L}_{t+1}(w_t) + \bar{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_t) \right)
\]

By construction, we have $E_{z_{t+1}}[Q_{t+1}|z_{i_1}, \ldots, z_{it}] = 0$ and hence the sequence $Q_2, \ldots, Q_T$ forms a martingale difference sequence. By Lemma 4.2 we have $Q_{t+1}$ lies in $[-M - GB_t, M + GB_t] \subseteq [-M - GB_T, M + GB_T]$ as $B_t$'s are non-decreasing. An application of the Azuma-Hoeffding inequality shows that with probability at least $1 - \gamma$,
\[
\frac{1}{T} \sum_{t=1}^{T} Q_t \leq (M + GB_T) \sqrt{\frac{2\ln(1/\gamma)}{T}}.
\]

We now analyze each term $P_t$ individually. Let us start by introducing a ghost sample $\{z'_{i_1}, \ldots, z'_{it}\}$, where each $z'_{ik}$ follows the same distribution as $z_{ik}$. By linearity of expectation, we have
\[
F_S(w_t) = E\left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} f(w_t, z_i, z'_{ik}) \right],
\]
where the expectation is taken over $\{z'_{ik}\}_{k=1}^{t}$. It allows us to write $P_t$ as follow
\[
P_t = E\left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} f(w_t, z_i, z'_{ik}) \right] - \hat{L}_{t+1}(w_t) \leq \sup_{w \in \mathcal{W}_t} E\left[ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{t} \sum_{k=1}^{t} f(w, z_i, z'_{ik}) \right] - \hat{L}_{t+1}(w)
\]
\[\triangleq g_{t+1}(z_{i_1}, \ldots, z_{it}).\]

Since $f$ is bounded by $A_t$, the expression $g_{t+1}(z_{i_1}, \ldots, z_{it})$ can have a variation of at most $(M + GB_t)/t$ when changing any of its $t$ variables. Hence an application of McDiarmid’s
inequality [McDiarmid, 1989] gives us, with probability at least 1 − γ,

\[ g_{t+1}(z_1, \ldots, z_t) \leq \mathbb{E}_{z_1, \ldots, z_t} [g_{t+1}(z_1, \ldots, z_t)] + (M + GB_t) \sqrt{\frac{\ln(1/\gamma)}{2t}}. \]

For any \( w \in W_t \), let \( f(w, z') = \frac{1}{t} \sum_{i=1}^{n} f(w, z_i, z'). \) Then we can write the expectation \( \mathbb{E}_{z_1, \ldots, z_t} [g_{t+1}(z_1, \ldots, z_t)] \) as follow

\[
\mathbb{E}_{z_k} [g_{t+1}(z_i, \ldots, z_t)] = \mathbb{E}_{z_k} \left[ \max_{w \in W_t} \mathbb{E}_{z_k'} \left[ \sum_{k=1}^{t} f(w, z_k') - \sum_{k=1}^{t} f(w, z_k) \right] \right]
\]

\[
\leq \mathbb{E}_{z_k, z_k'} \left[ \max_{w \in W_t} \sum_{k=1}^{t} f(w, z_k') - \sum_{k=1}^{t} f(w, z_k) \right]
\]

\[
= \mathbb{E}_{z_k, z_k'} \left[ \max_{w \in W_t} \sum_{k=1}^{t} \sigma_k (f(w, z_k') - f(w, z_k)) \right]
\]

\[
\leq \frac{2}{t} \mathbb{E}_{z_k, \sigma_k} \left[ \max_{w \in W_t} \sum_{k=1}^{t} \sigma_k \sigma_k \sum_{i=1}^{n} f(w, z_i, z_k) \right]
\]

\[
\leq \frac{2}{t} \sum_{i=1}^{n} \mathbb{E}_{z_k, \sigma_k} \left[ \max_{w \in W_t} \sum_{k=1}^{t} \sigma_k f(w, z_i, z_k) \right] = 2R_t(f \circ W_t).
\]

Thus we have, with probability at least 1 − γ,

\[ P_t \leq 2R_t(f \circ W_t) + (M + GB_t) \sqrt{\frac{\ln(1/\gamma)}{2t}}. \]

The Lemma holds by applying a union bound on \( P_t \) and taking the average over \( t \).

**Lemma 4.4** Let \( W_B = \{ w \in W \mid \|w\|_2 \leq B \} \) and let \( M = \sup_{z, z'} f(0, z, z') \). With probability 1 − γ, we have

\[
\frac{1}{T} \sum_{t=1}^{T} \hat{L}_{t+1}(w_*) - F_S(w_*) \leq \frac{2}{T} \sum_{t=1}^{T} R_t(f \circ W_B) + 3(M + GB) \sqrt{\frac{\ln(T/\gamma)}{T}}.
\]

**Proof:** Similar to the proof of Lemma 4.3 by replacing \( w_t \) with \( w_* \).

**Lemma 4.5** Suppose \( f \) is nonnegative, convex and G-Lipschitz. Suppose \( \|w_*\|_2 \leq B \). Sup-
pose we run Algorithm 2 for $T$ iterations, then we have

$$\frac{1}{T} \sum_{t=1}^{T} \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_*) \leq \frac{B^2}{2T\eta} + \frac{\eta G^2}{2}$$

**Proof:** By the update rule of Algorithm 2, we have

$$\|w_{t+1} - w_*\|_2^2 = \|w_t - \eta \partial \hat{L}_{t+1}(w_t) - w_*\|_2^2$$

$$= \|w_t - w_*\|_2^2 + \eta^2 \|\partial \hat{L}_{t+1}(w_t)\|_2^2 - 2\eta \langle w_t - w_*, \partial \hat{L}_{t+1}(w_t) \rangle$$

$$\leq \|w_t - w_*\|_2^2 + \eta^2 G^2 - 2\eta \langle w_t - w_*, \partial \hat{L}_{t+1}(w_t) \rangle.$$ 

Therefore, by the convexity of $\hat{L}_{t+1}$, we have

$$\sum_{t=1}^{T} \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_*) \leq \sum_{t=1}^{T} \langle w_t - w_*, \partial \hat{L}_{t+1}(w_t) \rangle$$

$$\leq \sum_{t=1}^{T} \frac{\|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2}{2\eta} + \frac{T\eta G^2}{2}$$

$$\leq \frac{\|w_*\|_2^2}{2\eta} + \frac{\eta G^2}{2},$$

the Lemma holds by dividing $T$ over both sides.

Next we give an upper bound on the optimization error $F_S(w_T) - F_S(w_*)$. The results are inspired by Kar et al. [2013], where they consider the online-to-batch generalization bound for pairwise learning. Our bound in the next theorem is given for optimization bound on finite sample.

**Theorem 4.4** Suppose there are some non-decreasing sequence $0 \leq B_t < \infty$ such that $\|w_t\|_2 \leq B_t$, and let $M = \sup_{z,z'} f(0, z, z')$ and $B = \|w_*\|_2$. Suppose we run Algorithm 2 for $T$ iterations, then with probability at least $1 - \gamma$ with respect to the sample $S$ and the internal
randomness of Algorithm 2, we have

\[ F_S(\bar{w}_T) - F_S(w_*) \leq \frac{2}{T} \sum_{t=1}^{T} R_t(f \circ W_t) + \frac{2}{T} \sum_{t=1}^{T} R_t(f \circ W_B) + \frac{B^2}{2T\eta} + \eta G^2 + \left(6M + 3GB\right)\sqrt{\frac{\ln(2T/\gamma)}{T}} + 3GBT\sqrt{\frac{\ln(2T/\gamma)}{T}}, \]

where \( W_t = \{w \in W ||w||_2 \leq B_t\} \) and \( W_B = \{w \in W ||w||_2 \leq B\} \) are subspaces of \( W \).

Proof: By the convexity of the empirical loss \( F_S \), we have

\[ F_S(\bar{w}_T) - F_S(w_*) \leq \frac{1}{T} \sum_{t=1}^{T} F_S(w_t) - F_S(w_*) \]

\[ = \frac{1}{T} \sum_{t=1}^{T} \left( F_S(w_t) - \hat{L}_{t+1}(w_t) + \hat{L}_{t+1}(w_*) - F_S(w_*) + \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_*) \right). \tag{4.7} \]

The conclusion follows from Lemma 4.3 and Lemma 4.4 both with probability \( 1 - \gamma/2 \) and Lemma 4.5.

Based on Theorem 4.1 and the error decomposition (4.1), we derive the excess risk bounds for bounded (Theorem 4.5) and unbounded domains (Theorem 4.6).

**Theorem 4.5** Suppose \( W \) is bounded with diameter \( D \). Denote \( M = \sup_{z,z'} f(0,z,z') \). Assume we run Algorithm 2 for \( T \geq n \) iterations under random selection with replacement rule. Then for any \( \gamma \in (0,1) \), with probability at least \( 1 - \gamma \) w.r.t. the sample \( S \) and the internal randomness of \( A \), we have

\[ \epsilon_{\text{risk}}(\bar{w}_T) \leq \frac{4}{T} \sum_{t=1}^{T} R_t(f \circ W) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + c_2 \sqrt{\frac{\ln(6T/\gamma)}{n}} \]

\[ + c_1 \eta \ln(n) \left( \sqrt{T} + \sqrt{3T \ln(6T/\gamma)} \right), \]

where \( c_1 = 100\sqrt{6e^{3/2}G} \max\{1, G\} \ln(6e/\gamma) \) and \( c_2 = (6 + 19e)(M + GD) \).

In particular, if \( R_t(f \circ W) = \mathcal{O}(1/\sqrt{t}) \) and we choose \( T = n^2 \) and \( \eta = \mathcal{O}(n^{-3/2}) \) then with high probability we have

\[ \epsilon_{\text{risk}}(\bar{w}_T) = \tilde{O}\left(\frac{\ln^2(n)}{\sqrt{n}}\right). \]
Proof: Since $W$ is bounded by $D$, we have $B = B_t = D$. Furthermore, by Lemma 4.2 we have $\sup_{z,z'} f(w, z, z') \leq M + GD$ and $\sup_{z,z'} f(w_t, z, z') \leq M + GD$. The proof is completed by recalling the error decomposition (4.1), applying Corollary 4.2, Theorem 4.3 and 4.4 each with probability $1 - \gamma/3$.

Using standard technique [Bartlett and Mendelson, 2002], the Rademacher complexity estimation of $R_t(f \circ W) = O(1/\sqrt{t})$ holds true in many cases when $\mathcal{X}$ and $W$ are bounded (e.g. see Section 4.3 for concrete examples of AUC maximization and similarity metric learning). It is worthy of mentioning that the choice of $T = n^2$ is consistent with pointwise learning with non-smooth loss [Bassily et al., 2020, Lei and Ying, 2020].

Although the boundedness assumption on the parameter space $W$ is removed, the next lemma characterizes the bound of the iterates $w_t$ by the sum of stepsizes.

Lemma 4.6 Suppose $f$ is nonnegative, convex and $G$-Lipschitz. Let $M = \sup_{z,z'} f(0, z, z')$. Let $\{w_t\}$ be the sequence of iterates by Algorithm 2 with $\eta \leq 1$. Then

$$\|w_{t+1}\|^2 \leq (G^2 + 2M)\eta t.$$ 

Proof: By the update rule of Algorithm 2 we have

$$\|w_{t+1}\|^2 = \|w_t - \eta \partial \hat{L}_{t+1}(w_t)\|^2 = \|w_t\|^2 + \eta^2 \|\partial \hat{L}_{t+1}(w_t)\|^2 - 2\eta \langle w_t, \partial \hat{L}_{t+1}(w_t) \rangle$$

$$\leq \|w_t\|^2 + \eta G^2 - 2\eta \langle w_t, \partial \hat{L}_{t+1}(w_t) \rangle \leq \|w_t\|^2 + \eta G^2 + 2\eta (\hat{L}_{t+1}(0) - \hat{L}_{t+1}(w_t))$$

$$\leq \|w_t\|^2 + \eta (G^2 + 2M),$$

where the first inequality holds since $f$ is $G$-Lipschitz and $\eta \leq 1$, the second inequality is due to the convexity of $f$ and the last inequality is due to the nonnegativity of $f$ and the definition of $M$.

We can also derive excess generalization bounds for Algorithm 2 even when $W$ is unbounded. Specifically, let $D = \|w_*\|_2$ and $W_D = \{w \in \mathcal{W} | \|w\|_2 \leq D\}$. The main idea is to show that the iterate $w_t$ from Algorithm 2 has an adaptive bound, i.e. $w_t \in W_t = \{w \in W | \|w\|_2^2 \leq (G^2 + M)\eta t\}$.

Theorem 4.6 Denote $M = \sup_{z,z'} f(0, z, z')$ and $D = \|w_*\|_2$. Suppose we run Algorithm 2
for $T \geq n$ iterations. For any $\gamma \in (0, 1)$, with probability at least $1 - \gamma$ w.r.t. the sample $S$ and the internal randomness of $A$, we have

$$
\epsilon_{\text{risk}}(\bar{w}_T) \leq \frac{2}{T} \sum_{t=1}^{T} (\mathcal{R}_t(f \circ \mathcal{W}_t) + \mathcal{R}_t(f \circ \mathcal{W}_D)) + \frac{D^2}{2T} + \frac{\eta G^2}{2}
+ c_4 \sqrt{\frac{\eta \ln(6T/\gamma)}{n}} + c_5 \sqrt{\frac{\eta T \ln(6e/\gamma)}{n}}
+ c_3 \eta \sqrt{n} \left(\sqrt{T} + \frac{4T \ln(eT) \sqrt{\ln(6n/\gamma)}}{n}\right),
$$

where $c_1 = 100 \sqrt{6e^{3/2}} G \max\{1, G\} \ln(6e/\gamma)$, $c_3 = (7 + 12 \sqrt{2e}) M + 4GD + 16eG$, $c_4 = 3G \sqrt{G^2 + 2M}$ and $c_5 = 12 \sqrt{2eG} \sqrt{G^2 + 2M}$.

In particular, if $\mathcal{R}_t(f \circ \mathcal{W}_t) = \mathcal{O}(\eta \sqrt{t})$ and $\mathcal{R}_t(f \circ \mathcal{W}_D) = \mathcal{O}(1/\sqrt{t})$ and we choose $T = n^{4/3}$ and $\eta = \mathcal{O}(n^{-1})$, then with high probability we have

$$
\epsilon_{\text{risk}}(\bar{w}_T) = \tilde{\mathcal{O}}\left(\frac{\ln^2(n)}{n^{1/3}}\right).
$$

Proof: By assumption and Lemma 4.6 we have $B = D$ and $B_t = \sqrt{(G^2 + 2M)\eta t}$. Therefore, by Lemma 4.2 we also get $\sup_{w, z, z'} f(w_t, z, z') \leq M + GD$ and $\sup_{w, z, z'} f(w_t, z, z') \leq M + G \sqrt{(G^2 + 2M)\eta t}$. The proof is completed by recalling the error decomposition (4.1), applying Corollary 4.2 Theorem 4.3 and 4.4 with probability $1 - \gamma/3$ each.

In particular, one can show that the Rademacher complexity can be estimated using standard technique [Bartlett and Mendelson, 2002] such that $\mathcal{R}_t(f \circ \mathcal{W}_D) = \mathcal{O}(D/\sqrt{t})$ when $\mathcal{X}$ is a bounded domain. Therefore by the definition of $\mathcal{W}_t$ one can similarly show that $\mathcal{R}_t(f \circ \mathcal{W}_t) = \mathcal{O}(\eta t/\sqrt{t}) = \mathcal{O}(\eta \sqrt{t})$. One can see more discussion on such estimation in Section 4.3. Therefore, Theorem 4.6 mainly differs from Theorem 4.5 in the additional $\tilde{\mathcal{O}}(\sqrt{\eta T/n})$ term where $T \geq n$. This is due to the unboundedness of $\mathcal{W}$. Our excess risk bound is consistent with the results in [Lin et al., 2016] in the pointwise setting (up to a logarithmic term), where the authors studied SGD for non-smooth loss functions in the pointwise setting using uniform convergence. However, the bound there is given in expectation while we have provided a high-probability bound.
4.2.2 Differentially Private Pairwise Learning

We show the implication of stability analysis in analyzing differentially private SGD in pairwise learning. There are other forms of differential privacy such as Gaussian differential privacy [Dong et al., 2021]. In this paper we restrict our attention to the standard DP mentioned above. In particular, we consider Gaussian mechanism [Dwork et al., 2006], i.e. given any query function \( q : S^n \rightarrow \mathbb{R}^d \), let \( A(S) = q(S) + u \) where \( u \sim \mathcal{N}(0, \sigma^2 I_d) \) with \( I_d \) being the identical matrix. For all neighborhood datasets \( S, S' \) that differ by one example, the \( f_2 \)-sensitivity \( \Delta \) of the query function \( q \) is defined as \( \Delta(q) = \sup_{S, S'} \| q(S) - q(S') \|_2 \).

**Algorithm 3** Private SGD for Pairwise Learning with Output Perturbation

**Input:** Private dataset \( S = \{z_1, \ldots, z_n\} \), privacy parameter \( \epsilon, \delta \), stepsize \( \eta \), number of iterations \( T \), initial point \( w_1 = 0 \) and initial sample \( i_1 \in [n] \) from uniform distribution

\[
\text{for } t = 1 \text{ to } T \text{ do}
\]
\[
\text{Select } i_{t+1} \in [n] \text{ from uniform distribution}
\]
\[
w_{t+1} = \Pi_W \left( w_t - \frac{\eta}{t} \sum_{k=1}^{t} \partial f(w_{t}, z_{i_{t+1}}, z_{i_k}) \right)
\]

\[
\text{end for}
\]

\[
w_T = \frac{1}{T} \sum_{t=1}^{T} w_t
\]

Sample \( u \sim \mathcal{N}(0, \sigma^2 I_d) \) with \( \sigma^2 \) being given by (4.8)

**Output:** \( w_{\text{priv}} = \Pi_W(w_T + u) \)

We develop a private version of SGD for pairwise learning. In this subsection, the notation \( A \) denotes Algorithm 3. The idea is to add Gaussian noise to the output of the non-private Algorithm 2. In return, Algorithm 3 is guaranteed to be \((\epsilon, \delta)\)-DP by properly choosing \( \sigma \) as shown below.

In order to establish the privacy guarantee of Algorithm 3, we need the following lemmas. The first lemma characterizes the necessary scale of \( \sigma \) of Gaussian mechanism [Dwork et al., 2014].

**Lemma 4.7 (Gaussian mechanism)** For a Gaussian mechanism \( A(S) = q(S) + u \) with \( u \sim \mathcal{N}(0, \sigma^2 I_d) \), if \( q \) has \( f_2 \)-sensitivity \( \Delta(q) \) and assume that \( \sigma \geq \sqrt{2 \ln(1.25/\delta) \Delta(q)/\epsilon} \), then \( A \) yields \((\epsilon, \delta)\)-DP.

The next lemma indicates that differential privacy is immune to post-processing [Dwork et al., 2014].
Lemma 4.8 (Post-processing) Let $\mathbf{A} : \mathcal{Z}^n \rightarrow \mathcal{W}$ be a (randomized) algorithm that is $(\epsilon, \delta)$-DP. Let $f : \mathcal{W} \rightarrow \mathcal{W}$ be an arbitrary randomized mapping. Then $f \circ \mathbf{A} : \mathcal{Z}^n \rightarrow \mathcal{W}$ is $(\epsilon, \delta)$-DP.

Theorem 4.7 Given the total number of iterations $T$, for any privacy budget $\epsilon > 0$ and $\delta > 0$, Algorithm $\mathcal{A}$ satisfies $(\epsilon, \delta)$-differential privacy with

$$\sigma^2 = \frac{8en^2G^2 \ln(2.5/\delta)}{\epsilon^2} \left( T + \frac{3T^2 \ln^2(\epsilon T) \ln^2(2/\delta)}{n^2} \right).$$

(4.8)

Proof: Consider the mechanism $\mathbf{A}'_T = \bar{w}_T + \mathbf{u}$ and for any $S, S'$, consider the $f_2$-sensitivity $\Delta_T = \|\bar{w}_T - \bar{w}'_T\|_2$. Let $I = \{i_1, \ldots, i_T\}$ be the sequence of sampling after $T$ iterations in Algorithm $\mathcal{A}$. Choosing $\gamma = \delta' / 2$ in Equation (4.5), then the event

$$E = \left\{ I | \Delta_T^2 \leq 4en^2G^2 \left( T + \frac{3T^2 \ln^2(\epsilon T) \ln^2(2/\delta)}{n^2} \right) \right\}$$

satisfies $\mathbb{P}[I \in E] \geq 1 - \delta/2$. When $I \in E$, Lemma 4.7 implies $\mathbf{A}'_T$ satisfies $(\epsilon, \delta'/2)$-DP when

$$\sigma = \frac{\sqrt{2 \ln(2.5/\delta)}}{\epsilon} \Delta_T.$$

Furthermore, by Lemma 4.8, the final output $\mathbf{w}_{\text{priv}} = \Pi_{\mathcal{W}}(\mathbf{A}'_T)$ also satisfies $(\epsilon, \delta'/2)$-DP. Therefore, for any $\epsilon > 0$ and any event $O$ in the output space of $\mathbf{w}_{\text{priv}},$

$$\mathbb{P}[\mathbf{w}_{\text{priv}}(S) \in O] = \mathbb{P}[\mathbf{w}_{\text{priv}}(S) \in O \cap I \in E] + \mathbb{P}[\mathbf{w}_{\text{priv}}(S) \in O \cap I \notin E]$$

$$= \mathbb{P}[\mathbf{w}_{\text{priv}}(S) \in O | I \in E] \mathbb{P}[I \in E] + \mathbb{P}[\mathbf{w}_{\text{priv}}(S) \in O | I \notin E] \mathbb{P}[I \notin E]$$

$$\leq \left( e^{\epsilon} \mathbb{P}[\mathbf{w}_{\text{priv}}(S') \in O | I \in E] + \frac{\delta}{2} \right) \mathbb{P}[I \in E] + \frac{\delta}{2}$$

$$\leq e^{\epsilon} \mathbb{P}[\mathbf{w}_{\text{priv}}(S') \in O \cap I \in E] + \frac{\delta}{2} + \frac{\delta}{2}$$

$$\leq e^{\epsilon} \mathbb{P}[\mathbf{w}_{\text{priv}}(S') \in O] + \delta$$

where the first inequality is because when $I \in E$, $\mathbf{w}_{\text{priv}}$ satisfies $(\epsilon, \delta'/2)$-DP and the fact $\mathbb{P}[I \notin E] \leq \delta/2$, the second inequality is by the definition of conditional probability. The
The proof is complete.

The goal here is to guarantee privacy with the added noise being as small as possible. The key observation is that the UAS of the non-private output \( \bar{w}_T \) can be used to quantify the high-probability sensitivity of the query function \( q(S) = \bar{w}_T \). Specifically, subsampling forms an event of probability measure \( 1 - \delta/2 \) under which a small sensitivity \( \tilde{O}(\eta \sqrt{T} + \eta T \ln(T)/n) \) holds true. Hence, under this event, we only need to add noise with \( \sigma = \tilde{O}((\eta \sqrt{T} + \eta T \ln(T)) \ln(2/\delta)/(n\epsilon)) \) to guarantee a slightly restrictive \((\epsilon, \delta/2)\)-DP. Therefore, the algorithm is \((\epsilon, \delta)\)-DP over the whole event space. \[\text{Wu et al. } [2017]\] studied differential private SGD by output perturbation method in the pointwise learning setting and they also utilized the idea of bounding sensitivity by UAS. However, they considered the stability and sensitivity regardless of the randomness of the algorithm, which is not suitable for high probability analysis of utility bound later. In contrast, our technique can also be applied to derive privacy guarantee and high probability utility in pointwise learning. \[\text{Huai et al. } [2020]\] also studied the sensitivity of SGD for Pairwise learning. However, they focused on the online setting where the data arrives in a streaming manner, and hence the different example between \( S \) and \( S' \) will only appear once in the algorithm. While in our stochastic setting the different example can be used more than once by subsampling, it is more challenging to measure the sensitivity. Moreover, their analysis depends on the strong smoothness of the loss function while we allow the loss function to be non-smooth.

In order to derive the utility bound of Algorithm 3, we need a new error decomposition scheme as follow

\[
\epsilon_{\text{risk}}(w_{\text{priv}}) = F(w_{\text{priv}}) - F(w_*) = F(w_{\text{priv}}) - F(\bar{w}_T) + F(\bar{w}_T) - F(w_*),
\]

where \( F(\bar{w}_T) - F(w_*) \) measures the excess risk incurred by the non-private output \( \bar{w}_T \) (Algorithm 2) and \( F(w_{\text{priv}}) - F(\bar{w}_T) \) measures the effect of perturbation by adding random noises. Some necessary lemmas are given below. The first one is the Chernoff bound for the \( f_2 \) norm of a Gaussian vector.

**Lemma 4.9** Let \( X_1, \cdots , X_d \) be i.i.d standard Gaussian random variables. And denote \( X = [X_1, \cdots , X_d] \in \mathcal{R}^d \). Then for any \( \tilde{\gamma} \in (0, 1) \), with probability at least \( 1 - \exp(-d\tilde{\gamma}^2/8) \) there holds \( \|X\|_2^2 \leq (1 + \tilde{\gamma})d \).
The next lemma tells us the error incurred by $F(w_{\text{priv}}) - F(\bar{w}_T)$ is bounded by the added noise $u$.

**Lemma 4.10** Suppose $f$ is nonnegative, convex and $G$-Lipschitz. Consider $w_{\text{priv}}$ and $\bar{w}_T$ from Algorithm 3. For any $\gamma > 0$, and for any $\gamma \in (\exp(-d/8), 1)$, with probability at least $1 - \gamma$, we have

$$F(w_{\text{priv}}) - F(\bar{w}_T) \leq 2G\sigma\sqrt{d \ln^{1/4}(1/\gamma)}.$$  

**Proof:** By the definition of $R$, we have

$$F(w_{\text{priv}}) - F(\bar{w}_T) = \mathbb{E}_{z,z'}[f(w_{\text{priv}}, z, z') - f(\bar{w}_T, z, z')]$$

$$\leq \mathbb{E}_{z,z'}[(w_{\text{priv}} - \bar{w}_T, \partial f(w_{\text{priv}}, z, z'))]$$

$$\leq \mathbb{E}_{z,z'}[\|\Pi_W(\bar{w}_T + u) - \bar{w}_T\|_2\|\partial f(w_{\text{priv}}, z, z')\|_2]$$

$$\leq \mathbb{E}_{z,z'}[\|u\|_2\|\partial f(w_{\text{priv}}, z, z')\|_2]$$

$$\leq G\|u\|_2 \quad (4.10)$$

where the first inequality is due to the convexity of $f$, the second inequality is by Cauchy-Schwarz inequality, the third inequality is by the non-expansiveness of projection and the last inequality is because $f$ is $G$-Lipschitz for any $w, z, z'$. Now, since $u \sim \mathcal{N}(0, \sigma^2 I_d)$, then by Lemma 4.9 for $\gamma \in (\exp(-d/8), 1)$ we have with probability $1 - \gamma$,

$$\|u\|_2 \leq \sigma\sqrt{d} \left(1 + \left(\frac{8\ln(1/\gamma)}{d}\right)^{1/4}\right).$$

Plugging the above inequality back into Equation (4.10) we get the desired result. \qed

The utility bound is given as follow.

**Theorem 4.8** Suppose $W$ is bounded with diameter $D$. Consider Algorithm 3 for $T$ iterations under random selection with replacement rule. For any privacy budget $\epsilon > 0$, $\delta > 0$, \ldots
and for any \( \gamma \in (\max\{4\delta, \exp(-d/8)\}, 1) \), with probability at least \( 1 - \gamma \), we have

\[
\epsilon_{\text{risk}}(w_{\text{priv}}) \leq \frac{4}{T} \sum_{t=1}^{T} \mathcal{R}_t(f \circ \mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + c_2 \sqrt{\frac{\ln(6T/\gamma)}{n}} + 2G\sigma \sqrt{d} \ln^{1/4}(4/\gamma).
\]

\[
+ c_1 \eta \ln(n) \left( \sqrt{T} + \frac{\sqrt{3T \ln(eT) \ln(2/\delta)}}{n} \right),
\]

where \( c_1 = 100\sqrt{6e^{3/2}}G \max\{1, G\} \ln(6e/\gamma) \) and \( c_2 = (6+19e)(M+GD) \).

In particular, letting \( \sigma \) satisfy (4.8) and choosing \( T = n^2 \) and \( \eta = \mathcal{O}(n^{-3/2}) \), then with high probability we have

\[
\epsilon_{\text{risk}}(w_{\text{priv}}) = \tilde{\mathcal{O}}\left( \frac{\sqrt{d}}{\sqrt{n\epsilon}} \right).
\]

**Proof:** For any neighborhood datasets \( S \) and \( S' \), Theorem 4.1 implies with probability least \( 1 - \delta/2 \) that

\[
\|\bar{w}_T - \bar{w}'_T\|_2 \leq 2\sqrt{eG\eta} \left( \sqrt{T} + \frac{\sqrt{3T \ln(eT) \ln(2/\delta)}}{n} \right).
\] (4.11)

Since \( \gamma \geq 4\delta \), we know the (4.11) holds with probability at least \( 1 - \gamma/8 \). Applying Corollary 4.2 with (4.11) we know with probability at least \( 1 - \gamma/4 \) we have

\[
F(\bar{w}_T) - F_S(\bar{w}_T) \leq 2\sqrt{eG\eta} \left( 4 + 48\sqrt{6G\ln(n)} \ln(8e/\gamma) \right) \left( \sqrt{T} + \frac{\sqrt{3T \ln(eT) \ln(2/\delta)}}{n} \right)
\]

\[
+ 12\sqrt{2e(M+GD)} \sqrt{\frac{\ln(8e/\gamma)}{n}}.
\] (4.12)

Recalling the error decomposition (4.9) and applying Theorem 4.3, Theorem 4.4 and Lemma 4.10 each with probability \( 1 - \gamma/4 \) together with (4.12), we have the desired bound.

The difference compared to Theorem 4.5 is the additional \( \tilde{\mathcal{O}}(\sigma \sqrt{d}) \) term caused by \( F(w_{\text{priv}}) - F(\bar{w}_T) \) in (4.9). The utility bound \( \tilde{\mathcal{O}}(\sqrt{d}/(\sqrt{n}\epsilon)) \) matches that of the output perturbation for pairwise learning studied in [Huai et al. 2020] which, however, requires the loss to be both strongly smooth and Lipschitz continuous. Our analysis only needs the loss to be Lipschitz continuous.
4.3 Applications

In this section, we illustrate our main results in the above sections by considering two concrete examples of pairwise learning, namely AUC maximization and similarity metric learning. According to Theorems 4.5 and 4.8, the key here is to estimate the Rademacher complexity defined by (4.6).

**AUC Maximization.** AUC maximization aims to learn a ranking function \( h_w \) defined by

\[
h_w(x, x') = w^\top (x - x').
\]

One expects \( h_w \) will rank positive examples higher than negative examples, i.e. \( w^\top (x - x') \geq 0 \) for \( y = 1 \) and \( y' = -1 \). Using the hinge loss \( f(w, z, z') = (1 - h_w(x, x')) + \mathbb{I}[y=1 \land y'=-1] \), AUC maximization can be formulated as

\[
\min_{w \in W} \mathbb{E}_{z, z'}[(1 - w^\top (x - x')) + \mathbb{I}[y=1 \land y'=-1]].
\]  

(4.13)

Denote \( \kappa = \sup_x \|x\|_2 \). The Rademacher complexity defined by (4.6) for AUC maximization is given in the following lemma.

**Lemma 4.11** Given the parameter space \( W = \{ w \in \mathcal{R}^d : \|w\|_2 \leq D \} \), the Rademacher complexity of \( H = \{ h_w : w \in W \} \) can be upper bounded by

\[
\mathcal{R}_t(H) \leq 2D\kappa/\sqrt{t}.
\]

**Proof:** Starting with the definition, the Rademacher complexity can be upper bounded by

\[
\mathcal{R}_t(H) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[ \sup_{h \in H} \frac{1}{t} \sum_{k=1}^t \sigma_k h_w(x_i, x_{ik}) \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[ \sup_{w \in \mathcal{X}_t} \frac{1}{t} \sum_{k=1}^t \sigma_k \langle w, x_i - x_{ik} \rangle \right]
\]

\[
\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[ \sup_{w \in \mathcal{X}_t} \|w\|_2 \frac{1}{t} \sum_{k=1}^t \sigma_k (x_i - x_{ik}) \|_2 \right] \leq \frac{D}{nt} \sum_{i=1}^n \left( \mathbb{E}\left[ \| \sum_{k=1}^t \sigma_k (x_i - x_{ik}) \|_2 \right] \right)^2 \leq \frac{2DE\kappa}{\sqrt{t}}
\]

where the first inequality is due to Cauchy-Schwarz inequality, the third identity is due to \( \{ \sigma_k \}_{k=1}^t \) are independent random variables with mean zero.

Note in the case of (4.13), it is easy to check \( \mathcal{R}_t(f \circ H) \leq 4GD\kappa/\sqrt{t} \) by Ledoux-Talagrand inequality [Ledoux and Talagrand, 1991]. Combining this lemma with Theorems 4.5 and 4.8, one can derive the following excess risk and utility bound for Algorithms 2 and
in the context of non-smooth AUC maximization.

**Corollary 4.3** Consider the problem of AUC maximization \( (4.13) \). If one runs Algorithm 3 with \( T = n^2 \) and \( \eta = \mathcal{O}(n^{3/2}) \), then, with high probability we have

\[
\varepsilon_{\text{risk}}(\bar{w}_T) = \tilde{O}\left(\sqrt{\frac{\kappa}{n}}\right).
\]

If one runs Algorithm 3 with \( T = n^2, \eta = \mathcal{O}(n^{3/2}) \) and \( \sigma \) given by \( (4.8) \), then, with high probability we have

\[
\varepsilon_{\text{risk}}(w_{\text{priv}}) = \tilde{O}\left(\sqrt{\frac{\kappa d}{n \epsilon}}\right).
\]

**Similarity Metric Learning.** We now turn to another notable example of pairwise learning called similarity metric learning. It aims to learn a (squared) Mahalanobis distance metric which is defined by \( h_w(x, x') = (x - x')^\top w(x - x') \) parametrized by a positive semi-definite matrix \( w \in \mathbb{R}^{d \times d} \). The intuition behind similarity metric learning is that the distance between samples from the same class should be small and the distance between examples from distinct classes should be large. Using the hinge loss \( f(w, z, z') = (1 + \tau(y, y')h_w(x, x'))_+ \), it can be formulated as

\[
\min_{w \in \mathcal{W}} \mathbb{E}_{z, z'}[(1 + \tau(y, y')(x - x')^\top w(x - x'))_+], \tag{4.14}
\]

where \( \tau(y, y') = 1 \) if \( y = y' \) and \(-1\) otherwise.

For any \( p \geq 1 \), the Schatten-\( p \) norm of a matrix \( W \in \mathbb{R}^{d \times d} \) is defined as the \( f_p \)-norm of the vector of singular values \( \sigma(W) := (\sigma_1(W), \ldots, \sigma_d(W))^\top \) (the singular values are assumed to be sorted in non-increasing order), i.e., \( \|W\|_{S_p} := \|\sigma(W)\|_p \). Let \( \Sigma = \mathbb{E}[XX^\top] \). We assume \( d \geq 3 \).

The following Khintchine-Kahane inequality [Lust-Piquard and Pisier 1991] provides a powerful tool to control the \( q \)-th norm of the summation of Rademacher series. The following form can be found in [Qiu and Wicks 2014].

**Lemma 4.12 (Matrix Khintchine)** Let \( X_1, \ldots, X_n \) be a set of symmetric matrices of the same dimension and let \( \sigma_1, \ldots, \sigma_n \) be a sequence of independent Rademacher random vari-
ables. For all $q \geq 2$,

\[
\left( \mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i X_i \right\|_{S_q}^q \right)^{\frac{1}{q}} \leq 2^{-\frac{1}{4}} \sqrt{\frac{q \pi}{c}} \left\| \left( \sum_{i=1}^n X_i^2 \right)^{\frac{1}{2}} \right\|_{S_q}.
\] (4.15)

The following inequality is the Bernstein inequality for a summation of independent matrices [Tropp, 2015].

**Lemma 4.13 (Matrix Bernstein)** Let $Z_1, \ldots, Z_n$ be independent, mean-zero and symmetric random matrices in $\mathbb{R}^{d \times d}$. Assume that each one is uniformly bounded

\[
\mathbb{E}[Z_i] = 0 \quad \text{and} \quad \|Z_i\|_{S_\infty} \leq L \quad \text{for each } i = 1, \ldots, n.
\]

Introduce the sum $S = \sum_{i=1}^n Z_i$ and let $v(S)$ denote the matrix variance statistic of the sum

\[
v(S) = \left\| \sum_{i=1}^n \mathbb{E}[Z_i^2] \right\|_{S_\infty}.
\]

Then

\[
\mathbb{E}[\|S\|_{S_\infty}] \leq \sqrt{2v(S) \log(2d)} + \frac{L}{3} \log(2d).
\]

**Lemma 4.14** Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables. Then

\[
\frac{1}{n} \mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_{S_\infty} \leq 2^{\frac{1}{4}} \sqrt{\frac{\pi d}{n}} \left( \frac{\sqrt{\log(2d) \sup_x \|x\|_2^2}}{n} + \frac{2\mathbb{E}[\|X\|_2^2 XX^\top]^{\frac{1}{2}}}{\sqrt{n}} \right). \tag{4.16}
\]

Under the mild assumption $\sqrt{\log(2d) \sup_x \|x\|_2^2} \leq \sqrt{n} \mathbb{E}[\|X\|_2^2 XX^\top]^{\frac{1}{2}}_{S_\infty}$ we get

\[
\frac{1}{n} \mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_{S_\infty} = O \left( \frac{\sqrt{\log d} \mathbb{E}[\|X\|_2^2 XX^\top]^{\frac{1}{2}}_{S_\infty}}{\sqrt{n}} \right).
\]
Proof: By the concavity of the square-root function, we know
\[
\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_{S_\infty} \leq \left( \mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_{S_q}^q \right)^{\frac{1}{q}} \leq 2^{-\frac{q}{4}} \left( \mathbb{E}_\sigma \left\| \left( \sum_{i=1}^n x_i x_i^\top \right)^{\frac{1}{2}} \right\|_{S_q} \right)^{\frac{1}{2}}
\]
\[
\leq 2^{-\frac{q}{4}} \sqrt{\frac{q \pi}{e}} \left\| \left( \sum_{i=1}^n x_i x_i^\top \right)^{\frac{1}{2}} \right\|_{S_\infty},
\]
where we have used Lemma 4.12 and \( \|W\|_{S_\infty} \leq \|W\|_{S_q} \leq d^{\frac{1}{2}} \|W\|_{S_\infty} \) for all \( W \in \mathbb{R}^{d \times d} \). If we choose \( q = 2 \log d \) (\( d \geq 3 \)), then
\[
\sqrt{qd^\frac{1}{4}} = \sqrt{2 \log d \frac{1}{2 \log d}} = \sqrt{2e \log d}
\]
and therefore
\[
\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_{S_\infty} \leq 2^{\frac{1}{4}} \sqrt{\pi \log d} \left( \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_q} \right)^{\frac{1}{2}}
\]
\[
= 2^{\frac{1}{4}} \sqrt{\pi \log d} \left( \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} \right)^{\frac{1}{2}}.
\]
It then follows from the concavity of the square-root function that
\[
\mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i x_i^\top \right\|_{S_\infty} \leq 2^{\frac{1}{4}} \sqrt{\pi \log d} \left( \mathbb{E} \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} \right)^{\frac{1}{2}} \tag{4.17}
\]
It is clear
\[
\mathbb{E} \left[ \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} \right] \leq \mathbb{E} \left[ \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} - \mathbb{E} \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} \right] + \mathbb{E} \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty}
\]
\[
= \mathbb{E} \left[ \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} - \mathbb{E} \left\| \sum_{i=1}^n x_i x_i^\top \right\|_{S_\infty} \right] + n \mathbb{E} \left[ \|X\|_{2}^{\infty} XX^\top \right] \tag{4.18}
\]
For all \( i \in [n] \), denote \( Z_i = \|x_i\|_{2}^{\infty} x_i x_i^\top - \mathbb{E} \left[ \|x_i\|_{2}^{\infty} x_i x_i^\top \right] \). It is clear that
\[
\mathbb{E} \left[ \sum_{i=1}^n Z_i^2 \right] = \sum_{i=1}^n \mathbb{E} \left[ \|x_i\|_{2}^{\infty} x_i x_i^\top \right] - \sum_{i=1}^n \left( \mathbb{E} \left[ \|x_i\|_{2}^{\infty} x_i x_i^\top \right] \right) \left( \mathbb{E} \left[ \|x_i\|_{2}^{\infty} x_i x_i^\top \right] \right)
\]
\[
= n \mathbb{E} \left[ \|X\|_{2}^{\infty} XX^\top \right] - n \mathbb{E} \left[ \|X\|_{2}^{\infty} XX^\top \right] \mathbb{E} \left[ \|X\|_{2}^{\infty} XX^\top \right] \leq n \mathbb{E} \left[ \|X\|_{2}^{\infty} XX^\top \right]
\]
and therefore
\[ \left\| \mathbb{E} \left[ \sum_{i=1}^{n} Z_i^2 \right] \right\|_{s_{\infty}} \leq n \left\| \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \right\|_{s_{\infty}}. \] (4.19)

Furthermore,
\[ \|Z_i\|_{s_{\infty}} \leq \sup_{x_i} \|x_i x_i^\top x_i x_i^\top\|_{s_{\infty}} \leq \sup_x \|x\|_2^4. \] (4.20)

We can apply Lemma 4.13 with the above bound of variance (4.19) and magnitude (4.20), and derive
\[ \mathbb{E} \left[ \sum_{i=1}^{n} Z_i \right] \leq \frac{1}{2n} \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \frac{\log(2d)}{s_{\infty}} + \frac{1}{3} \sup_x \|x\|_2^2 \log(2d). \]

This together with the sub-additivity of the square-root function and (4.18) implies
\[ \left( \mathbb{E} \left[ \sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top \right] \right)^{\frac{1}{2}} \leq \left( \mathbb{E} \left[ \sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top \right] - \mathbb{E} \left[ \sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top \right] \right)^{\frac{1}{2}} + \left( n \left\| \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \right\|_{s_{\infty}} \right)^{\frac{1}{2}} \leq (2n \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \frac{\log(2d)}{s_{\infty}} + \sqrt{\frac{\log(2d)}{3n}} \sup_x \|x\|_2^2 + \sqrt{\frac{n \mathbb{E} \left[ \|X\|_2^2 XX^\top \right]}{s_{\infty}}} \right)^{\frac{1}{2}}. \]

We plug the above inequality back into (4.17), and get the inequality
\[ \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{n} \sigma_i x_i x_i^\top \right] \leq 2^{\frac{3}{2}} \sqrt{\pi \log(d)} \left( (2 \log(2d))^{\frac{1}{2}} n^{-\frac{2}{3}} \left\| \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \right\|_{s_{\infty}}^{\frac{1}{2}} \right. \]
\[ \left. + \sqrt{\frac{\log(2d)}{3n}} \sup_x \|x\|_2^2 + \frac{\mathbb{E} \left[ \|X\|_2^2 XX^\top \right]}{\sqrt{n}} \right). \] (4.21)

It is clear that
\[ \left\| \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \right\|_{s_{\infty}}^{\frac{1}{2}} \leq \sup_x \|x\|_2 \left\| \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \right\|_{s_{\infty}}^{\frac{1}{4}}. \]

This together with Cauchy-Schwartz inequality shows that
\[ (2 \log(2d))^{\frac{1}{2}} n^{-\frac{2}{3}} \left\| \mathbb{E} \left[ \|X\|_2^2 XX^\top \right] \right\|_{s_{\infty}}^{\frac{1}{4}} \leq \sqrt{\frac{n \mathbb{E} \left[ \|X\|_2^2 XX^\top \right]}{s_{\infty}}} + \sqrt{\frac{\log(2d) \sup_x \|x\|_2^2}{2^{\frac{3}{2}} n}}. \]

Plugging the above inequality back into (4.21) gives the stated bound (4.16) \((2^{-\frac{3}{2}} + 3^{-\frac{1}{2}} < 1)\).
The proof is complete.

Lemma 4.15  Consider the parameter space defined via the nuclear norm, i.e.

\[ \mathcal{W} = \{ \mathbf{w} \in \mathbb{R}^{d \times d}, \| \mathbf{w} \|_{S_1} \leq D \}, \]

where \( \| \mathbf{w} \|_{S_1} \) denotes the nuclear norm of a matrix \( \mathbf{w} \). The complexity of \( \mathcal{H} = \{ h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W} \} \) is bounded by

\[ \mathcal{R}_t(\mathcal{H}) = \mathcal{O} \left( \frac{D \| \mathbb{E}[\|X\|_2^2 XX^\top] \|_{S_\infty}^{\frac{1}{2}} \sqrt{\log d}}{\sqrt{t}} \right), \quad (4.22) \]

where \( \| \cdot \|_{S_\infty} \) denotes the largest singular value.

Proof:  The complexity of \( \mathcal{H} = \{ h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W} \} \) is bounded by

\[
\mathcal{R}_t(\mathcal{H}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{\mathbf{w} \in \mathcal{W}} \frac{1}{t} \sum_{k=1}^{t} \sigma_k \left( \mathbf{w}, (\mathbf{x}_i - \mathbf{x}_{i_k}) (\mathbf{x}_i - \mathbf{x}_{i_k})^\top \right) \right] \\
\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{\mathbf{w} \in \mathcal{W}} \| \mathbf{w} \|_{S_1} \| \frac{1}{t} \sum_{k=1}^{t} \sigma_k (\mathbf{x}_i - \mathbf{x}_{i_k}) (\mathbf{x}_i - \mathbf{x}_{i_k})^\top \|_{S_\infty} \right] \\
\leq \frac{D}{nt} \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{k=1}^{t} \sigma_k (\mathbf{x}_i - \mathbf{x}_{i_k}) (\mathbf{x}_i - \mathbf{x}_{i_k})^\top \right\|_{S_\infty} = \mathcal{O} \left( \frac{D \| \mathbb{E}[\|X\|_2^2 XX^\top] \|_{S_\infty}^{\frac{1}{2}} \sqrt{\log d}}{\sqrt{t}} \right),
\]

where \( \| \cdot \|_{S_\infty} \) denotes the largest singular value of a matrix and we have used Lemma 4.14 in the last step.

As direct corollaries of Lemma 4.15, we can derive generalization bounds for metric learning from Theorems 4.5 and 4.8. For brevity, denote \( \chi = \| \mathbb{E}[\|X\|_2^2 XX^\top] \|_{S_\infty} \). We derive the following results of SGD for pairwise learning in the context of non-smooth metric learning.

Corollary 4.4  Consider the similarity metric learning problem (4.14). If one runs Algorithm 2 for \( T = n^2 \) and \( \eta = \mathcal{O}(n^{-3/2}) \), then, with high probability we have

\[ \epsilon_{\text{risk}}(\bar{\mathbf{w}}_T) = \tilde{\mathcal{O}} \left( \sqrt{\frac{\chi \log(d)}{n}} \right). \]
If one runs Algorithm 3 with $T = n^2$, $\eta = \mathcal{O}(n^{-3/2})$ and $\sigma$ given by (4.8), then, with high probability we have

$$
\epsilon_{\text{risk}}(w_{\text{priv}}) = \tilde{\mathcal{O}}\left(\frac{\sqrt{d \log(d)}}{\sqrt{n\epsilon}}\right).
$$

**Remark 4.1** We now show the advantage of our result as compared to the existing results.

Based on the argument in Lei and Ying [2016], it can be shown

$$
\mathcal{R}_t(H) = \mathcal{O}\left(\frac{D \sup_x \|x\|^2 \sqrt{\log d}}{t}\right).
$$

The difference between (4.22) and (4.23) is that we replace the term $\sup_x \|x\|^2$ by the term $\|\mathbb{E}[\|X\|^2 XX^\top]\|_{S_\infty}^{1/2}$. Notice

$$
\|\mathbb{E}[\|X\|^2 XX^\top]\|_{S_\infty} \geq \frac{1}{d} tr\left(\mathbb{E}[XX^\top XX^\top]\right) = \frac{1}{d} \mathbb{E}\left[tr(XX^\top XX^\top)\right] = \frac{1}{d} \mathbb{E}\left[\|X\|_2^4\right].
$$

If we assume $\mathbb{E}[\|X\|_2^2] \geq d^2$, then the bound of (4.22) satisfies the relation $\geq \sqrt{d \log d} / \sqrt{t}$ and in the extreme case this lower bound can be achieved within a constant factor. As a comparison, the upper bound in (4.23) satisfies the relation $\geq d \sqrt{\log d} / t$. That is, our argument outperforms the existing results by enjoying a milder dependency on the dimensionality for using nuclear-norm constraints, which is appealing in the high-dimensional setting. If we use Frobenius-norm constraint in defining $W \{w \in \mathbb{R}^{d \times d}, \|w\|_F \leq D_2\}$, then one can show that $\mathcal{R}_t(H) = \mathcal{O}\left(D_2 \sup_x \|x\|^2 / \sqrt{t}\right)$ [Lei and Ying, 2016]. This matches the bound (4.23) within a logarithmic factor except that $D$ there is replaced by $D_2$. Since $\|w\|_F \leq \|w\|_{S_1}$, the argument in Lei and Ying [2016] leads to a misleading argument that Frobenius-norm constraint is always preferable to the nuclear-norm constraint. It was posed as an open question on whether one can derive a generalization bound for similarity metric learning showing the advantage of nuclear-norm constraint over Frobenius-norm constraint [Cao et al., 2016]. We provide an affirmative solution to this open question in Lemma 4.15.
CHAPTER 5
Pairwise Learning Problems Revisited

5.1 Online and Offline Learning for Pairwise Learning

Our main results in this chapter are based on [Yang et al. 2021b]. In this section, we describe the proposed algorithms in two common learning settings for pairwise learning: offline and online learning settings.

Algorithm 4 SGD for Pairwise Learning

1: Inputs: $S = \{z_i : i \in [n]\}$ and step sizes $\{\eta_t\}$
2: Initialize: $w_0 \in W$, let $w_{-1} = w_0$ and randomly select $i_0 \in [n]$
3: for $t = 1, 2, \ldots, T$ do
4: Randomly select $i_t \in [n]$
5: $w_t = \text{proj}(w_{t-1} - \eta_t \nabla f(w_{t-1}; z_{i_t}, z_{i_t-1}))$
6: end for
7: Outputs: $\bar{w}_T = \sum_{j=1}^{T} \eta_j w_{j-2} / \sum_{j=1}^{T} \eta_j$

Offline Learning (Finite-Sum) Setting. The first is the finite-sum setting where the training data $S = \{z_i = (x_i, y_i) \in Z : i \in [n]\}$ are drawn independently according to $D$. In this context, one aims to solve the following empirical risk minimization (ERM):

$$w_S = \arg\min_{w \in W} \left[ F_S(w) := \frac{1}{n(n-1)} \sum_{i,j \in [n], i \neq j} f(w; z_i, z_j) \right].$$  (5.1)

Our proposed algorithm to solve (5.1) is described in Algorithm 4. The notation proj($\cdot$) there denotes the projection operator to $W$. In particular, at iteration $t$, it randomly selects one instance $z_{i_t}$ from the uniform distribution over $[n]$ and pairs it only with the previous instance $z_{i_{t-1}}$, and then do the gradient descent based on $\nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}})$. This is in contrast to the classical SGD for pairwise learning in [Kar et al. 2013, Wang et al. 2012, Zhao et al. 2011] where the present instance $z_{i_t}$ is paired with all previous instances $\{z_{i_1}, z_{i_2}, \ldots, z_{i_{t-1}}\}$. Note $w_{t-1}$ depends on $z_{i_{t-1}}$ and then $\nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}})$ is not an unbiased estimate of
Therefore, the standard analysis of SGD does not apply. We introduce novel techniques to handle this dependency in our analysis (see more details in Section 5.2).

**Algorithm 5** OGD for Pairwise Learning

1: **Inputs:** learning rates \( \{\eta_t\} \)
2: **Initialize:** \( w_0 \in W \), let \( w_{-1} = w_0 \) and receiving datum \( z_0 \).
3: for \( t = 1, 2, \ldots, T \) do
4: Receive a data point \( z_t \)
5: \( w_t = \text{proj} \left( w_{t-1} - \eta_t \nabla f(w_{t-1}; z_t, z_{t-1}) \right) \)
6: end for
7: **Outputs:** \( \bar{w}_T = \sum_{j=1}^{T} \eta_j w_{j-2} / \sum_{j=1}^{T} \eta_j \)

**Online Learning Setting.** In the online learning setting where the data \( z_0, z_1, z_2, \ldots \) is assumed i.i.d. from an unknown distribution \( \mathcal{D} \) on \( \mathcal{Z} \), the number of iterations of an online algorithm is identical to the size of available data. In the same spirit to Algorithm 4, the pseudo code is given in Algorithm 5. Specifically, upon receiving a datum \( z_t \) at the current time \( t \), we pair it with \( z_{t-1} \) which was revealed at the previous time \( t-1 \) and then perform gradient descent based on the gradient \( \nabla f(w_{t-1}; z_t, z_{t-1}) \). It aims to minimize the population risk which is defined as \( F(w) = \mathbb{E}_{z,z'}[f(w; z, z')] \). Here \( \mathbb{E}_{z,z'} \) denotes the expectation with respect to (w.r.t.) \( z, z' \sim \mathcal{D} \).

It is worth pointing out that online learning \cite{Hazan2016, Orabona2019, Shalev-Shwartz2012} in general does not require the i.i.d. assumption on the data and study the regret bounds. In this paper, we mainly consider the statistical performance, measured by excess generalization bounds, of the output \( \bar{w}_T \) of Algorithm 5 where the streaming data \( \{z_0, z_1, z_2, \ldots\} \) is i.i.d. from the population distribution \( \mathcal{D} \).

**Remark 5.1** As discussed above, OGD for pairwise learning was proposed and studied in \cite{Kar2013, Wang2012} where the current instance \( z_t \) is paired with a buffering set \( B_t \subseteq \{z_1, \ldots, z_{t-1}\} \). However, the resultant excess generalization bound is in the form of \( O\left(\frac{1}{\sqrt{s}} + \frac{1}{\sqrt{n}}\right) \) which indicated that the buffer size \( s \) needs to be large enough in order to achieve good generalization. Their analysis does not apply to our case since the buffering set \( B_t = \{z_{t-1}\} \) with size \( s = 1 \) for Algorithm 5. As we show soon in Section 5.2, we can prove...
that Algorithm 5, which pairs the current instance $z_t$ with the previous instance $z_{t-1}$, still enjoys optimal statistical performance $O(1/\sqrt{n})$.

**Remark 5.2** The work by Lei et al. [2020] studied the stability and generalization of an SGD-type algorithm for pairwise learning by randomly generating pairs of instances. Specifically, at time $t$, randomly generating a pair $(z_{i_t}, z_{j_t})$ from a given set of training data $S = \{z_i : i \in [n]\}$ and the subsequent update is given by $w_t = w_{t-1} - \eta_t \nabla f(w_{t-1}; z_{i_t}, z_{j_t})$.

In contrast, our algorithm, i.e. Algorithm 4, updates the model parameter based on the pair of the current random instance and the random one generated at the previous time $t - 1$, i.e. $(z_{i_t}, z_{i_{t-1}})$. Furthermore, our work here significantly differs from Lei et al. [2020] in the following aspects. Firstly, the algorithm there by randomly selecting pairs of instances does not work in the online learning setting while ours can seamlessly deal with the streaming data as stated in Algorithm 5. Secondly, regarding the technical analysis, our algorithms are more challenging to analyze than the algorithm in Lei et al. [2020]. Indeed, $\nabla f(w_{t-1}; z_{i_t}, z_{j_t})$ is not an unbiased estimate of $\nabla F_S(w_{t-1})$ due to the independency between $w_{t-1}$ and $(i_t, j_t)$. Therefore, the optimization error analysis of the algorithm in Lei et al. [2020] is the same as the SGD for pointwise learning. As a comparison, $\nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}})$ is a biased estimate of $\nabla F_S(w_{t-1})$ due to the coupling between $w_{t-1}$ and $i_{t-1}$. We introduce novel techniques to handle this coupling for both the optimization and generalization analyses. Thirdly, we will soon see below that we provide generalization results for nonsmooth, nonconvex losses and also use Algorithm 4 to develop novel differentially private pairwise learning algorithms while Lei et al. [2020] focused on the smooth convex losses in the non-private setting.

**Remark 5.3** If we let $\xi_t = (i_t, i_{t-1})$ in Algorithm 4, then $\{\xi_t : t \in \mathbb{N}\}$ forms a Markov Chain as $\xi_t$ only depends on $\xi_{t-1}$ but not on $\{\xi_1, \ldots, \xi_{t-2}\}$. Hence, Algorithm 4 can be regarded as a Markov Chain SGD which was studied in Sun et al. [2018]. Despite this similarity, our results differ from Sun et al. [2018] in two important aspects. Firstly, we are mainly interested in stability and generalization of Algorithm 4 while Sun et al. [2018] focused on the convergence analysis of the Markov Chain SGD. One cannot apply the results in Sun et al. [2018] to obtain excess generalization bounds for Algorithm 4 in terms of the population risk as we will show soon in the next section. Secondly, directly applying Theorem 1 in Sun et al. [2018] only yields a convergence rate of $O(1/t^{1-q})$ with some $1/2 < q < 1$ in the convex setting. Our
proof for the convergence analysis of Algorithm 4 is much simpler and direct which can yield a faster convergence rate $O(1/\sqrt{t})$ as shown in Section 5.2.2.

5.2 Generalization Analysis

The aim for the generalization analysis of Algorithm 4 and Algorithm 5 is the same, i.e. to analyze the excess generalization error $F(w) - F(w^*)$ of a model $w$ measuring its relative behavior w.r.t. the best model $w^* = \arg\min_{w \in W} F(w)$.

For Algorithm 4 where the training data $S$ with $n$ datum is given beforehand, the excess generalization involves the generalization error and optimization error. Specifically, one has the following error decomposition for $\bar{w}$

$$
\mathbb{E}[F(\bar{w}_T)] - F(w^*) = \mathbb{E}[F(\bar{w}_T) - F_S(\bar{w}_T)] + \mathbb{E}[F_S(\bar{w}_T) - F_S(w^*)].
$$

(5.2)

Here, the expectation is taken w.r.t. the randomness of Algorithm 4, i.e., $\{i_t\}$ and the randomness of data $S$ which is i.i.d. from $D$ on $Z$. We refer to the first term $\mathbb{E}[F(\bar{w}_T) - F_S(\bar{w}_T)]$ as the generalization error and $\mathbb{E}[F_S(\bar{w}_T) - F_S(w^*)]$ as the optimization error. We will use algorithmic stability to handle its generalization error in Subsection 5.2.1 for smooth and nonsmooth losses. The estimation of the optimization error is given in Subsection 5.2.2 for both convex and nonconvex losses. For Algorithm 5, there is no generalization error as the data $\{z_1, z_2, \ldots, z_T\}$ is arriving in a sequential manner with $T$ increasing all the time which does not involve the training data. The randomness of Algorithm 5 is only from the i.i.d. data. Therefore, the optimization error in this setting is exactly the excess generalization error $F(\bar{w}_T) - F(w^*)$ which is estimated in Subsection 5.2.2.

5.2.1 Stability and Generalization Errors

We study generalization errors by algorithmic stability, which measures the sensitivity of the output of an algorithm w.r.t. the perturbation of the dataset. Below we give the definition of uniform argument stability. We say $S, S'$ are neighboring datasets if they differ at most by a single example.
Definition 5.1 A (randomized) algorithm \( \mathcal{A} \) for pairwise learning is called \( \varepsilon \)-uniformly argument stable if for all neighboring datasets \( S, S' \in \mathcal{Z}^n \) we have \( \mathbb{E}_{\mathcal{A}}[\|\mathcal{A}(S) - \mathcal{A}(S')\|_2] \leq \varepsilon \).

It is clear \( \varepsilon \)-uniform argument stability implies \( G\varepsilon \)-uniform stability i.e.,

\[
\sup_{z, z'} \mathbb{E}_{\mathcal{A}}[f(\mathcal{A}(S), z, z') - f(\mathcal{A}(S'), z, z')] \leq G\varepsilon
\]

for Lipschitz losses [Bousquet and Elisseeff, 2002]. The connection between the uniform stability for pairwise learning and its generalization has been established in the literature [Agarwal and Niyogi, 2009, Shen et al., 2020].

Lemma 5.1 If an algorithm \( \mathcal{A} \) for pairwise learning is \( \varepsilon \)-uniformly stable for some \( \varepsilon > 0 \), then we have \( \|\mathbb{E}_{S, A}[F_S(\mathcal{A}(S)) - F(\mathcal{A}(S))]\| \leq 2\varepsilon \).

We develop uniform argument stability bound of Algorithm 4 and apply it together with Lemma 5.1 to establish the following generalization bounds. Theorem 5.1 handles nonsmooth problems, while Theorem 5.2 handles smooth problems.

Theorem 5.1 Let \( w_{-1} = w_0 \) and \( \{w_j : j \in [t]\} \) be produced by Algorithm 4 with \( \eta_j = \eta \). Let Assumption 4.1 and 4.3 hold with \( \alpha = 0 \). Then, Algorithm 4 is \( 2\sqrt{eG\eta}(\sqrt{5t} + \frac{2n}{t}) \)-uniformly argument stable and

\[
\mathbb{E}_{S, A}[F(\bar{\mathcal{W}}_t) - F_S(\bar{\mathcal{W}}_t)] \leq 4\sqrt{eG^2\eta}\left(\sqrt{5t} + \frac{2t}{n}\right).
\]

Proof: We first investigate the uniform stability of Algorithm 4.

Let \( S' = \{z_1, \ldots, z_{n-1}, z'_n\} \), where \( z'_n \) is independently drawn from \( \mathcal{D} \), and \( \{w'_t\} \) be produced by Algorithm 4 w.r.t. data \( S' \). We consider two cases: i.e. the case of \( \{i_t \neq n \text{ and } i_{t-1} \neq n\} \) and the case of \( \{i_t = n \text{ or } i_{t-1} = n\} \).
If \( i_t \neq n \) and \( i_{t-1} \neq n \), then

\[
\| w_t - w_t' \|_2^2 \leq \| w_{t-1} - \eta_t \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - w_{t-1}' + \eta_t \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \|_2^2
\]

\[
= \| w_{t-1} - \eta_t \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - w_{t-1}' + \eta_t \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \|_2^2
\]

\[
= \| w_{t-1} - w_{t-1}' \|_2^2 + \eta_t^2 \| \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \|_2^2
\]

\[
- 2\eta_t \langle w_{t-1} - w_{t-1}', \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \rangle
\]

\[
\leq \| w_{t-1} - w_{t-1}' \|_2^2 + \eta_t^2 \| \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \|_2^2
\]

\[
\leq \| w_{t-1} - w_{t-1}' \|_2^2 + 4\eta_t^2 G^2,
\]

where the last second inequality follows from the inequality

\[
\langle w_{t-1} - w_{t-1}', \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \rangle \geq 0
\]

due to the convexity of \( f \) and the last inequality follows from the Lipschitz continuity of \( f \). If \( i_t = n \) or \( i_{t-1} = n \), it follows from the elementary inequality \((a + b)^2 \leq (1 + p)a^2 + (1 + 1/p)b^2\) and the Lipschitz condition that

\[
\| w_t - w_t' \|_2^2 \leq (1 + p)\| w_{t-1} - w_{t-1}' \|_2^2
\]

\[
+ (1 + 1/p)\eta_t^2 \| \nabla f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) - \nabla f(w_{t-1}'; z_{i_t}, z_{i_{t-1}}) \|_2^2
\]

\[
\leq (1 + p)\| w_{t-1} - w_{t-1}' \|_2^2 + 4(1 + 1/p)\eta_t^2 G^2.
\]

We can combine the above two cases together and derive

\[
\| w_t - w_t' \|_2^2 \leq \left( \| w_{t-1} - w_{t-1}' \|_2^2 + 4\eta_t^2 G^2 \right) \mathbb{I}_{i_t \neq n \text{ and } i_{t-1} = n}
\]

\[
+ \left( (1 + p)\| w_{t-1} - w_{t-1}' \|_2^2 + 4(1 + 1/p)\eta_t^2 G^2 \right) \mathbb{I}_{i_t = n \text{ or } i_{t-1} = n}
\]

\[
\leq (1 + p)\| w_{t-1} - w_{t-1}' \|_2^2 + 4\eta_t^2 G^2 \left( 1 + \mathbb{I}_{i_t = n \text{ or } i_{t-1} = n}/p \right)
\]

\[
= (1 + p)\| w_{t-1} - w_{t-1}' \|_2^2 + 4\eta_t^2 G^2 \left( 1 + \mathbb{I}_{i_t = n \text{ or } i_{t-1} = n}/p \right),
\]

138
where \( \mathbb{I}_{[i]} \) is the indicator function. We can apply the above inequality recursively and get

\[
\|w_t - w'_t\|_2^2 \leq 4G^2 \sum_{k=1}^{t} \eta_k^2 \left( 1 + \mathbb{I}_{[ik=n \text{ or } ik-1=n]} / p \right) \prod_{j=k+1}^{t} (1 + p) \mathbb{I}_{[ij=n \text{ or } i_{j-1}=n]}
\]

\[
\leq 4G^2 \prod_{j=1}^{t} (1 + p) \mathbb{I}_{[ij=n \text{ or } i_{j-1}=n]} \sum_{k=1}^{t} \eta_k^2 \left( 1 + \mathbb{I}_{[ik=n \text{ or } ik-1=n]} / p \right)
\]

\[
= 4G^2 \eta^2 \left( 1 + p \right) \sum_{j=1}^{t} \mathbb{I}_{[ij=n \text{ or } i_{j-1}=n]} \left( t + \sum_{k=1}^{t} \mathbb{I}_{[ik=n \text{ or } ik-1=n]} / p \right),
\]

where the last inequality follows from \( \eta_j = \eta \).

Now, we choose \( p = 1 / \left( \sum_{j=1}^{t} \mathbb{I}_{[ij=n \text{ or } i_{j-1}=n]} \right) \) and use the inequality \((1 + x)^{1/x} \leq e\) to derive the following inequality

\[
\|w_t - w'_t\|_2^2 \leq 4eG^2 \eta^2 \left( t + \left( \sum_{k=1}^{t} \mathbb{I}_{[ik=n \text{ or } ik-1=n]} \right)^2 \right).
\]

By the inequality \( \mathbb{I}_{[ik=n \text{ or } ik-1=n]} \leq \mathbb{I}_{[ik=n]} + \mathbb{I}_{[ik-1=n]} \) and \((a + b)^2 \leq 2(a^2 + b^2)\) we know

\[
\mathbb{E}_A \left[ \left( \sum_{k=1}^{t} \mathbb{I}_{[ik=n \text{ or } ik-1=n]} \right)^2 \right] \leq 2 \mathbb{E}_A \left[ \left( \sum_{k=1}^{t} \mathbb{I}_{[ik=n]} \right)^2 \right] + 2 \mathbb{E}_A \left[ \left( \sum_{k=1}^{t} \mathbb{I}_{[ik-1=n]} \right)^2 \right]
\]

\[
= 4 \mathbb{E} \left[ \left( \sum_{k=1}^{t} \mathbb{I}_{[ik=n]} \right)^2 \right] \leq 4t + 4 \sum_{j,k \in [t]: j \neq k} \mathbb{E} \left[ \mathbb{I}_{[ij=n]} \mathbb{I}_{[ik=n]} \right]
\]

\[
= 4t + 4 \sum_{j,k \in [t]: j \neq k} \frac{1}{n^2} \leq 4t + 4t^2 / n^2.
\]

We can combine the above two inequalities together and derive

\[
\mathbb{E}_A[\|w_t - w'_t\|_2^2] \leq 4eG^2 \eta^2 \left( 5t + \frac{4t^2}{n^2} \right)
\]

and by the convexity of \( \| \cdot \|_2^2 \) it follows

\[
\mathbb{E}_A[\|\bar{w}_t - \bar{w}'_t\|_2^2] \leq \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}_A[\|w_j - w'_j\|_2^2] \leq 4eG^2 \eta^2 \left( 5t + \frac{4t^2}{n^2} \right).
\]
This establishes the uniform stability of Algorithm 4. Furthermore, for any \(z, z'\), we have

\[
\mathbb{E}_{\mathcal{A}}[f(\bar{w}_t, z, z') - f(\bar{w}_t', z, z')] \leq G \mathbb{E}_{\mathcal{A}}[\|\bar{w}_t - \bar{w}_t'\|_2] = G \mathbb{E}_{\mathcal{A}}\left[\sqrt{\|\bar{w}_t - \bar{w}_t'\|^2}\right] \\
\leq G \sqrt{\mathbb{E}_{\mathcal{A}}[\|\bar{w}_t - \bar{w}_t'\|^2]} \leq 2\sqrt{e} G^2 \eta \left(\sqrt{5t} + \frac{2t}{n}\right)
\]

where the first inequality we used the \(G\)-Lipschitz continuity of \(f\) and the second inequality we used the Jensen’s inequality. Therefore, Algorithm 4 is uniformly stable and the generalization error satisfies

\[
\mathbb{E}_{\mathcal{A}}[f(\bar{w}_t, z, z') - f(\bar{w}_t', z, z')] \leq 2\sqrt{e} G^2 \eta \left(\sqrt{5t} + \frac{2t}{n}\right)
\]

which gives us the desired result.

To prove the smooth case we require the following lemma on the nonexpansiveness of gradient map \(\mathbf{w} \mapsto \mathbf{w} - \eta \nabla f(\mathbf{w}; z, z')\).

**Lemma 5.2 (Hardt et al. 2016)** Assume for all \(z \in \mathcal{Z}\), the function \(\mathbf{w} \mapsto f(\mathbf{w}; z, z')\) is convex and \(L\)-smooth. Then for all \(\eta \leq 2/L\) and \(z, z' \in \mathcal{Z}\) there holds

\[
\|\mathbf{w} - \eta \nabla f(\mathbf{w}; z, z') - \mathbf{w}' + \eta \nabla f(\mathbf{w}'; z, z')\|_2 \leq \|\mathbf{w} - \mathbf{w}'\|_2.
\]

**Theorem 5.2** Let \(\mathbf{w}_0 = \mathbf{w}_0\) and \(\{\mathbf{w}_j : j \in [t]\}\) be produced by Algorithm 4 with \(\eta_j = \eta \leq 2/L\). Let Assumption 4.1, 4.2 and 4.3 hold true with \(\alpha = 0\). Then, Algorithm 4 is \(\frac{4G}{n} \sum_{j=1}^{t} \eta_j\)-uniformly argument stable and the generalization error satisfies \(\mathbb{E}_{\mathcal{S}, \mathcal{A}}[F(\bar{w}_t) - F_S(\bar{w}_t)] \leq \frac{8G^2}{n} \sum_{j=1}^{t} \eta_j\).

**Proof**: Let \(S' = \{z_1, \ldots, z_{n-1}, z'_n\}\), where \(z'_n\) is independently drawn from \(\mathcal{D}\). Let \(\{\mathbf{w}'_j\}\) be produced by Algorithm 4 w.r.t. \(S'\). We consider two cases. If \(i_t \neq n\) and \(i_{t-1} \neq n\), then it follows from Lemma 5.2 that

\[
\|\mathbf{w}_t - \mathbf{w}'_t\|_2 \leq \|\mathbf{w}_{t-1} - \eta_t \nabla f(\mathbf{w}_{t-1}; z_{i_t}, z_{i_{t-1}}) - \mathbf{w}'_{t-1} + \eta_t \nabla f(\mathbf{w}'_{t-1}; z'_{i_t}, z'_{i_{t-1}})\|_2 \\
\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2.
\]
Otherwise, we know

\[ \|w_t - w'_t\|_2 \leq \|w_{t-1} - w'_{t-1}\|_2 + \eta_t \|\nabla f(w_{t-1}; z_{it}, z_{i_{t-1}}) - \nabla f(w'_{t-1}; z'_{it}, z'_{i_{t-1}})\|_2 \]

\[ \leq \|w_{t-1} - w'_{t-1}\|_2 + 2\eta G. \]

We can combine the above two cases together and derive the following inequality

\[ \|w_t - w'_t\|_2 \leq \|w_{t-1} - w'_{t-1}\|_2 \Pi_{[i \neq n \text{ and } i_{t-1} \neq n]} + (\|w_{t-1} - w'_{t-1}\|_2 + 2\eta G)\Pi_{[i = n \text{ or } i_{t-1} = n]} \]

\[ = \|w_{t-1} - w'_{t-1}\|_2 + 2\eta G\Pi_{[i = n \text{ or } i_{t-1} = n]}. \]

We can apply the above inequality recursively and get

\[ \|w_t - w'_t\|_2 \leq 2G \sum_{j=1}^{t} \eta_j \Pi_{[i_j = n \text{ or } i_{j-1} = n]} \leq 2G \sum_{j=1}^{t} \eta_j (\Pi_{[i_j = n]} + \Pi_{[i_{j-1} = n]}). \]

Taking expectations over both sides gives \( \mathbb{E}_A[\|w_t - w'_t\|_2] \leq \frac{4G}{n} \sum_{j=1}^{t} \eta_j. \) It then follows from the convexity of \( \| \cdot \|_2 \) that

\[ \mathbb{E}_A[\|w_t - w'_t\|_2] \leq \frac{4G}{n} \sum_{j=1}^{t} \eta_j. \]

This establishes the uniform argument stability of Algorithm 4. Furthermore, it follows the Lipschitz condition that

\[ \sup_{z,z'} \mathbb{E}_A[f(\bar{w}_t; z, z') - f(\bar{w}'_t; z, z')] \leq \frac{4G^2}{n} \sum_{j=1}^{t} \eta_j. \]

The desired result then follows from Lemma 5.1. The proof for Theorem 5.2 is completed. \( \square \)

For Algorithm 5, there is no generalization error as the data \( \{z_1, z_2, \ldots, z_T\} \) is assumed to arrive in a sequential manner with \( T \) increasing all the time which does not involve the training data.
5.2.2 Optimization Error

In this subsection, we establish the convergence rate, i.e., optimization error, of Algorithm 4 for convex, nonconvex and strongly convex problems. We consider both bounds in expectation and with high probability. Our analysis is based on the key observation

\[ f(w_{t-1}; z_t, z_{t-1}) = f(w_{t-2}; z_t, z_{t-1}) + O(\eta_{t-1}). \]

Below we only present optimization error bounds for Algorithm 4 here in the offline (finite-sum) setting where the training data of size \( n \), denoted by \( S = \{z_1, \ldots, z_n\} \), is fixed, and the optimization error is measured by \( F_S(\bar{w}_t) - \inf_{w \in W} F_S(w) \). We emphasize that all our optimization error bounds hold true for Algorithm 5 in the online learning setting with exactly the same analysis where the streaming data \( \{z_1, \ldots, z_t, \ldots\} \) is assumed to be i.i.d according to the population distribution \( D \), and the bounds for the optimization error in this case is given for the excess generalization error (excess population risk), i.e. \( F(\tilde{w}_t) - \inf_{w \in W} F(w) \).

**Theorem 5.3** Let \( w_0 = w \) and \( \{w_j : j \in [t]\} \) be produced by Algorithm 4. Let Assumption 4.1 and 4.3 hold true with \( \alpha \geq 0 \). Then, for any \( w \) independent of \( A \) we have the following convergence rates:

(a) Assume \( f \) is convex, i.e. \( \alpha = 0 \). Then, we have

\[
\mathbb{E}_A[F_S(\bar{w}_t)] - F_S(w) \leq \frac{\|w_0 - w\|_2^2 + G^2 \sum_{j=1}^{t}(2\eta_j\eta_{j-1} + \eta_j^2)}{2 \sum_{j=1}^{t} \eta_j}. \tag{5.3}
\]

(b) Let \( f \) be \( \alpha \)-strongly convex with \( \alpha > 0 \) and \( \eta_j = \frac{2}{a(j+1)} \). Then, there holds \( \mathbb{E}_A[F_S(\bar{w}_t)] - F_S(w) = O\left(G^2/(\alpha t)\right) \).

**Proof:** Consider \( j \geq 1 \). Note that \( f(\cdot; z, z') \) is \( \alpha \)-strongly convex and \( G \)-Lipschitz
continuous, we have

\[ \|w_j - w\|_2^2 \leq \|w_{j-1} - \eta_j \nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}}) - w\|_2^2 \]

\[ = \|w_{j-1} - w\|_2^2 - 2\eta_j \langle \nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}}), w_{j-1} - w \rangle + \eta_j^2 \|\nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}})\|_2^2 \]

\[ \leq (1 - \eta_j \alpha)\|w_{j-1} - w\|_2^2 - 2\eta_j [f(w_{j-1}; z_{i_j}, z_{i_{j-1}}) - f(w; z_{i_j}, z_{i_{j-1}})] + G^2 \eta_j^2 \]

\[ = (1 - \eta_j \alpha)\|w_{j-1} - w\|_2^2 - 2\eta_j [f(w_{j-2}; z_{i_j}, z_{i_{j-1}}) - f(w_{j-1}; z_{i_j}, z_{i_{j-1}})] + G^2 \eta_j^2 \]

\[ + 2\eta_j [f(w_{j-2}; z_{i_j}, z_{i_{j-1}}) - f(w_{j-1}; z_{i_j}, z_{i_{j-1}})] + G^2 \eta_j^2 \]

\[ \leq (1 - \eta_j \alpha)\|w_{j-1} - w\|_2^2 - 2\eta_j [f(w_{j-2}; z_{i_j}, z_{i_{j-1}}) - f(w; z_{i_j}, z_{i_{j-1}})] \]

\[ + 2\eta_j G \|w_{j-1} - w_{j-2}\|_2 + G^2 \eta_j^2 \]

\[ \leq (1 - \eta_j \alpha)\|w_{j-1} - w\|_2^2 - 2\eta_j [f(w_{j-2}; z_{i_j}, z_{i_{j-1}}) - f(w; z_{i_j}, z_{i_{j-1}})] \]

\[ + 2G^2 \eta_j \eta_{j-1} + G^2 \eta_j^2, \]  

(5.4)

where the last inequality used the fact that \( \|w_j - w_{j-1}\|_2 = \eta_j \|\nabla f(w_j; z_{i_j}, z_{i_{j-1}})\|_2 \leq G \eta_j \).

For the convex case, i.e. \( \alpha = 0 \), we know from (5.4) that

\[ \sum_{j=1}^{t} \eta_j [f(w_{j-2}; z_{i_j}, z_{i_{j-1}}) - f(w; z_{i_j}, z_{i_{j-1}})] \]

\[ \leq \frac{1}{2} \sum_{j=1}^{t} \|w_{j-1} - w\|_2^2 - \|w_j - w\|_2^2 + \frac{G^2}{2} \sum_{j=1}^{t} (2\eta_{j-1} \eta_j + \eta_j^2) \]

\[ \leq \frac{1}{2} \|w_0 - w\|_2^2 + \frac{G^2}{2} \sum_{j=1}^{t} (2\eta_{j-1} \eta_j + \eta_j^2). \]  

(5.5)

Taking the expectation on both sides of the above inequality and observing that \( f(\cdot; z, z') \) is convex, we get the desired estimation (5.3).

For the strongly-convex case, i.e. \( \alpha > 0 \), we obtain from (5.4) that

\[ f(w_{j-2}; z_{i_j}, z_{i_{j-1}}) - f(w; z_{i_j}, z_{i_{j-1}}) \leq \frac{\eta_{j-1}^- - \alpha}{2} \|w_{j-1} - w\|_2^2 - \frac{\eta_j^{-1}}{2} \|w_j - w\|_2^2 + G^2 \eta_{j-1} + \frac{G^2 \eta_j}{2}. \]
Now, we choose \( \eta_j = \frac{2}{\alpha(j+1)} \) for any \( j \), which implies that

\[
j[f(w_{j-2}; z_{i,j}, z_{i,j-1}) - f(w; z_{i,j}, z_{i,j-1})] \leq \frac{j(j-1)\alpha}{4}\|w_{j-1} - w\|_2^2 - \frac{j(j+1)\alpha}{4}\|w_j - w\|_2^2 + \frac{2G^2}{\alpha} + \frac{G^2j}{\alpha(j+1)} \leq \frac{\alpha}{4}[j(j-1)\|w_{j-1} - w\|_2^2 - j(j+1)\|w_j - w\|_2^2 + 3G^2].
\]

Taking the summation over \( j \) implies that

\[
\sum_{j=1}^t j[f(w_{j-2}; z_{i,j}, z_{i,j-1}) - f(w; z_{i,j}, z_{i,j-1})] \leq \frac{3G^2t}{\alpha} + \frac{\alpha}{4} \sum_{j=1}^t [j(j-1)\|w_{j-1} - w\|_2^2 - j(j+1)\|w_j - w\|_2^2] \leq \frac{3G^2t}{\alpha} + \frac{\alpha}{4}[0 - t(t+1)\|w_t - w\|_2^2] \leq \frac{3G^2t}{\alpha}.
\]

(5.6)

Dividing both sides of the above inequality by \( \sum_{j=1}^t j \) yields the desired estimation in part (b).

\[\Box\]

**Remark 5.4** The above convergence rates match those in the pointwise learning [Bottou et al., 2018]. Furthermore, if \( \eta_j = \eta \), then Eq. (5.3) becomes \( \mathbb{E}_A[F_S(w_t)] - F_S(w) = \mathcal{O}(1/(t\eta) + \eta) \) and one can choose \( \eta \approx 1/\sqrt{t} \) to get \( \mathbb{E}_A[F_S(w_t)] - F_S(w) = \mathcal{O}(1/\sqrt{t}) \). We can extend our convergence analysis to a more general update as

\[
w_t = \Pi_W(w_{t-1} - \frac{\eta}{s} \sum_{j=1}^s \nabla f(w_{t-1}; z_{it}, z_{it-1}))
\]

for \( s \in \mathbb{N} \). Indeed, one can use the observation \( f(w_{t-1}; z_{it}, z_{it-1}) = f(w_{t-s-1}; z_{it}, z_{it-s}) + \mathcal{O}(\sum_{j=1}^s \eta_{t-j}) \) to derive the convergence rate \( \mathcal{O}(\sqrt{s}/\sqrt{t}) \).

To prove high-probability bounds, we require the following lemma on concentration inequalities of martingales [Boucheron et al., 2013, Zhang, 2005].

**Lemma 5.3** Let \( \tilde{z}_1, \ldots, \tilde{z}_n \) be a sequence of random variables such that \( \tilde{z}_k \) may depend on the previous variables \( \tilde{z}_1, \ldots, \tilde{z}_{k-1} \) for all \( k = 1, \ldots, n \). Consider a sequence of functionals \( \xi_k(\tilde{z}_1, \ldots, \tilde{z}_k), k = 1, \ldots, n \). Let \( \alpha_n^2 = \sum_{k=1}^n \mathbb{E}_z \left[ (\xi_k - \mathbb{E}_z[\xi_k])^2 \right] \) be the conditional variance.
(1) Assume $|\xi_k - \mathbb{E}_{\tilde{z}_k}[\xi_k]| \leq b_k$ for each $k$. Let $\delta \in (0, 1)$. With probability at least $1 - \delta$

$$\sum_{k=1}^{n} \mathbb{E}_{\tilde{z}_k}[\xi_k] - \sum_{k=1}^{n} \xi_k \leq \left(2 \sum_{k=1}^{n} b_k^2 \log \frac{1}{\delta}\right)^{\frac{1}{2}}. \quad (5.7)$$

(2) Assume that $\xi_k - \mathbb{E}_{\tilde{z}_k}[\xi_k] \leq b$ for each $k$. Let $D \in (0, 1]$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$ we have

$$\sum_{k=1}^{n} \mathbb{E}_{\tilde{z}_k}[\xi_k] - \sum_{k=1}^{n} \xi_k \leq \frac{D \alpha_n^2}{b} + \frac{b \log \frac{1}{\delta}}{D}. \quad (5.8)$$

Below we present high-probability bounds to understand the variation of the algorithm. We need to take conditional expectation of $f(w_{2j-2}; z_{i_{2j}}, z_{i_{2j-1}})$ w.r.t. $(i_{2j}, i_{2j-1})$ to get $F_S(w_{2j-2})$. However, there is a coupling between $(i_{2j}, i_{2j-1})$ and $(i_{2j-1}, i_{2j-2})$. Therefore, one can not directly apply concentration inequalities for martingales to handle

$$\sum_{j=1}^{t} \left(f(w_{2j-2}; z_{i_{2j}}, z_{i_{2j-1}}) - F_S(w_{2j-2})\right).$$

We introduce a novel decoupling technique to handle this coupling. Note that the high-probability bounds match the bounds in expectation up to a constant factor.

**Theorem 5.4** Let $w_{-1} = w_0$ and $\{w_j : j \in [t]\}$ be produced by Algorithm 4. Let Assumption 4.1 and 4.3 hold true with $\alpha \geq 0$ and $\sup_w f(w; z, z') \leq B$ for some $B > 0$. Let $\delta \in (0, 1)$.

(a) Assume $f$ is convex, i.e. $\alpha = 0$ and let $\bar{w}_t = \sum_{j=1}^{t} \eta_j w_{j-2}/\sum_{j=1}^{t} \eta_j$. Then, for any $w \in W$, with probability at least $1 - \delta$ the following inequality holds

$$F_S(\bar{w}_t) - F_S(w) \leq \frac{1}{\sum_{j=1}^{t} \eta_j} \left(2B \left(2 \sum_{j=1}^{t} \eta_j^2 \log (2/\delta)\right)^{\frac{1}{2}} + \frac{1}{2} \|w_1 - w\|_2^2 \right.$$

$$\left. + G^2 \sum_{j=1}^{t} (\eta_{j-1} \eta_j + \frac{\eta_j^2}{2}) \right).$$

(b) Assume $f$ is $\alpha$-strongly convex with $\alpha > 0$ and $\eta_j = \frac{2}{\alpha(j+1)}$. Let $\bar{w}_t = \sum_{j=1}^{t} j w_{j-2}/\sum_{j=1}^{t} j$. 

145
Then, with probability at least $1 - \delta$, we have

$$F_S(\bar{w}_t) - F_S(w) = O\left(G^2 \log(1/\delta)/(\alpha t)\right).$$

**Proof:** For simplicity, we assume $t$ is an even number. We first consider the convex case. Let

$$\xi_j = \eta_{2j} \left( f(w_{2j-2}; z_{i_{2j}}, z_{i_{2j-1}}) - f(w; z_{i_{2j}}, z_{i_{2j-1}}) \right), \quad j \in [t/2].$$

It is obvious that $|\xi_j - \mathbb{E}_{i_{2j}, i_{2j-1}}[\xi_j]| \leq 2B\eta_{2j}$. Let $\tilde{z}_j = (i_{2j}, i_{2j-1})$. It is clear that $\tilde{z}_j, j \in [t/2]$ are i.i.d. random variables. Therefore, one can apply Part (a) of Lemma 5.3 to derive the following inequality with probability at least $1 - \delta/2$

$$\sum_{j=1}^{t/2} \mathbb{E}_{\tilde{z}_j}[\xi_j] - \sum_{j=1}^{t/2} \xi_j \leq 2B\left(2 \sum_{j=1}^{t/2} \eta_{2j}^2 \log(2/\delta)\right)^{1/2}.\ 

It is clear that $\mathbb{E}_{\tilde{z}_j}[\xi_j] = \eta_{2j} (F_S(w_{2j-2}) - F_S(w))$. Therefore, the following inequality holds with probability at least $1 - \delta/2$

$$\sum_{j=1}^{t/2} \eta_{2j} \left(F_S(w_{2j-2}) - F_S(w) - f(w_{2j-2}; z_{i_{2j}}, z_{i_{2j-1}}) + f(w; z_{i_{2j}}, z_{i_{2j-1}})\right)$$

$$\leq 2B\left(2 \sum_{j=1}^{t/2} \eta_{2j}^2 \log(2/\delta)\right)^{1/2}.$$

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

$$\sum_{j=1}^{t/2} \eta_{2j-1} \left(F_S(w_{2j-3}) - F_S(w) - f(w_{2j-3}; z_{i_{2j-1}}, z_{i_{2j-2}}) + f(w; z_{i_{2j-1}}, z_{i_{2j-2}})\right) \leq 2B\left(2 \sum_{j=1}^{t/2} \eta_{2j-1}^2 \log(2/\delta)\right)^{1/2}.$$

We can combine the above two inequalities together and derive the following inequality with...
probability $1 - \delta$

$$\sum_{j=1}^{t} \eta_j (F_S(w_{j-2}) - F_S(w) - f(w_{j-2}; z_{j-1}) + f(w; z_{j-1}, z_{j-1})) \leq 2B \left( 2 \sum_{j=1}^{t} \eta_j^2 \log(2/\delta) \right)^{\frac{1}{2}}.$$

We can combine the above inequality and Eq. (5.5) to derive the following inequality with probability at least $1 - \delta$

$$\sum_{j=1}^{t} \eta_j (F_S(w_{j-2}) - F_S(w)) \leq 2B \left( 2 \sum_{j=1}^{t} \eta_j^2 \log(2/\delta) \right)^{\frac{1}{2}} + \frac{1}{2} \|w_0 - w\|^2 + G^2 \sum_{j=1}^{t} (\eta_{j-1} \eta_j + \frac{\eta_j^2}{2}).$$

The stated bound then follows from the convexity of $F_S$.

We now turn to the strongly convex case. Let

$$\xi_j = 2j(f(w_{2j-2}; z_{i_{2j}}, z_{i_{2j-1}}) - f(w_S; z_{i_{2j}}, z_{i_{2j-1}})), \quad j \in [t/2].$$

It is clear that $|\xi_j - \mathbb{E}_{i_{2j}, i_{2j-1}}[\xi_j]| \leq 4jB \leq 2tB$ for $j \in [t/2]$. Furthermore, the conditional variance satisfies

$$\mathbb{E}_{\tilde{z}_j} \left[ (\xi_j - \mathbb{E}_{\tilde{z}_j}[\xi_j])^2 \right] \leq \mathbb{E}_{\tilde{z}_j}[\xi_j^2] \leq 4j^2G^2\|w_{2j-2} - w_S\|^2_2 \leq 8\alpha^{-1}j^2G^2(F_S(w_{2j-2}) - F_S(w_S)),$$

where the first inequality follows from $f$ is $G$-Lipschitz continuous, and the second inequality used the fact $\nabla F_S(w_S) = 0$ and $f$ is $\alpha$-strongly convex.

Note $\tilde{z}_j$ are independent random variables and

$$\mathbb{E}_{\tilde{z}_j}[\xi_j] = 2j(F_S(w_{2j-2}) - F_S(w_S)).$$

Therefore, we can apply Part (b) of Lemma 5.3 to derive the following inequality with
probability at least $1 - \frac{\delta}{2}$

$$2 \sum_{j=1}^{t/2} j \left( F_S(w_{2j-2}) - F_S(w_S) - f(w_{2j-2}; z_{i_{2j-1}}, z_{i_{2j-2}}) + f(w_S; z_{i_{2j-1}}, z_{i_{2j-2}}) \right)$$

$$\leq \frac{8G^2D \sum_{j=1}^{t/2} j^2 (F_S(w_{2j-2}) - F_S(w_S))}{2tB\alpha} + \frac{2tB \log(2/\delta)}{D}.$$

In a similar way, one can derive the following inequality with probability at least $1 - \frac{\delta}{2}$

$$\sum_{j=1}^{t/2} (2j - 1) \left( F_S(w_{2j-3}) - F_S(w_S) - f(w_{2j-3}; z_{i_{2j-2}}, z_{i_{2j-3}}) + f(w_S; z_{i_{2j-2}}, z_{i_{2j-3}}) \right)$$

$$\leq \frac{2G^2D \sum_{j=1}^{t/2} (2j - 1)^2 (F_S(w_{2j-3}) - F_S(w_S))}{2tB\alpha} + \frac{2tB \log(2/\delta)}{D}.$$

We can combine the above two inequalities together and derive the following inequality with probability at least $1 - \delta$

$$\sum_{j=1}^{t} j \left( F_S(w_{2j-2}) - F_S(w_S) - f(w_{2j-2}; z_{i_{2j-1}}, z_{i_{2j-2}}) + f(w_S; z_{i_{2j-1}}, z_{i_{2j-2}}) \right)$$

$$\leq \frac{2G^2D \sum_{j=1}^{t} j^2 (F_S(w_{2j-2}) - F_S(w_S))}{2tB\alpha} + \frac{2tB \log(2/\delta)}{D}.$$

We can combine the above inequality and Eq. (5.6) together and derive the following inequality with probability $1 - \delta$

$$\sum_{j=1}^{t} j (F_S(w_{2j-2}) - F_S(w_S)) \leq \frac{3G^2t}{\alpha} + \frac{G^2D \sum_{j=1}^{t} j (F_S(w_{2j-2}) - F_S(w_S))}{B\alpha} + \frac{2tB \log(2/\delta)}{D}.$$

Now, we take $D = \min \{1, B\alpha/(2G^2)\}$ and get the following inequality with probability at least $1 - \delta$

$$\sum_{j=1}^{t} j (F_S(w_{2j-2}) - F_S(w_S)) \leq \frac{3G^2t}{\alpha} + \frac{1}{2} \sum_{j=1}^{t} j (F_S(w_{2j-2}) - F_S(w_S))$$

$$+ 2t \log(2/\delta) \max \{B, 2G^2/\alpha\}$$

148
and therefore
\[ \sum_{j=1}^{t} j(F_S(w_{j-2}) - F_S(w_S)) \leq \frac{14G^2t \log(2/\delta)}{\alpha} + 4Bt \log(2/\delta). \]

The stated bound then follows from the convexity of $F_S$. The proof is completed. \qed

Finally, we study the convergence of Algorithm 4 associated with nonconvex functions. We first consider general smooth problems. Since we cannot find a global minimum in this setting, we measure the convergence rate in terms of gradient norms [Ghadimi and Lan, 2013]. The following theorem establishes the convergence rate $O(1/\sqrt{t})$ for the gradient norm, i.e. \( \min_{j=1,...,t} \mathbb{E}_A[\|\nabla F_S(w_j)\|^2_2] \).

**Theorem 5.5** Let \( w_{-1} = w_0 \) and \( \{w_j : j \in [t]\} \) be produced by Algorithm 4 with \( \eta_j = \eta \leq 1/(2\sqrt{L}) \). Let 4.2 hold true and \( \mathbb{E}_{i_{j+1},i_{j+2}}[\|\nabla f(w_j; z_{i_{j+1}}, z_{i_{j+2}})\|^2_2] \leq \alpha_0^2 \) for some \( \alpha_0 \) and any \( j \). Then,
\[ \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}_A[\|\nabla F_S(w_j)\|^2_2] \leq \frac{F_S(w_0)}{t\eta} + 8L\eta\alpha_0^2. \]

Furthermore, choosing \( \eta \approx 1/\sqrt{t} \) implies that \( \frac{1}{t} \sum_{j=1}^{t} \mathbb{E}_A[\|\nabla F_S(w_j)\|^2_2] = O(1/\sqrt{t}). \)

**Proof:** It is clear that $F_S$ is $L$-smooth and therefore
\[ F_S(w_j) \leq F_S(w_{j-1}) + \langle w_j - w_{j-1}, \nabla F_S(w_{j-1}) \rangle + \frac{L}{2} \|w_j - w_{j-1}\|^2 \]
\[ = F_S(w_{j-1}) - \eta_j \langle \nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}}), \nabla F_S(w_{j-1}) \rangle + \frac{L\eta_j^2}{2} \|\nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}})\|^2_2 \]

Taking expectations over both sides gives
\[ \mathbb{E}_A[F_S(w_j)] \leq \mathbb{E}_A[F_S(w_{j-1})] - \eta_j \mathbb{E}_A[\langle \nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}}), \nabla F_S(w_{j-1}) \rangle] + \frac{L\eta_j^2}{2} \mathbb{E}_A[\|\nabla f(w_{j-1}; z_{i_j}, z_{i_{j-1}})\|^2_2]. \quad (5.9) \]
According to the elementary inequality \((a+b)^2 \leq 2(a^2 + b^2)\) we know

\[
\begin{align*}
\mathbb{E}_\mathcal{A} & \left[ \| \nabla f(w_{j-1}; z_{ij}, z_{ij-1}) \|_2^2 \right] \\
& \leq 2 \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-1}; z_{ij}, z_{ij-1}) - \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2^2 \right] + 2 \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2^2 \right] \\
& \leq 2L \mathbb{E}_\mathcal{A} \left[ \| w_{j-1} - w_{j-2} \|_2^2 \right] + 2\alpha_0^2 = 2L\eta_j^2 \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-2}; z_{ij-1}, z_{ij-2}) \|_2^2 \right] + 2\alpha_0^2 \\
& \leq \frac{1}{2} \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-2}; z_{ij-1}, z_{ij-2}) \|_2^2 \right] + 2\alpha_0^2,
\end{align*}
\]

where we have used the \(L\)-smoothness, the assumption \(\mathbb{E}_{ij, ij-1} \left[ \| \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2^2 \right] \leq \alpha_0^2\) and \(4L\eta_j^2 \leq 1\). It is clear that \(\mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_0; z_{i1}, z_{i0}) \|_2^2 \right] \leq \alpha_0^2\). It is easy to use an induction and the above inequality to show that

\[
\mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-1}; z_{ij}, z_{ij-1}) \|_2^2 \right] \leq 4\alpha_0^2, \quad \forall j. \tag{5.10}
\]

Furthermore, the smoothness assumption implies that

\[
\begin{align*}
\langle \nabla f(w_{j-1}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-1}) \rangle \\
= \langle \nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-1}) \rangle \\
+ \langle \nabla f(w_{j-1}; z_{ij}, z_{ij-1}) - \nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-1}) \rangle \\
= \langle \nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-2}) \rangle + \langle \nabla f(w_{j-1}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-1}) - \nabla F_S(w_{j-2}) \rangle + \\
+ \langle \nabla f(w_{j-1}; z_{ij}, z_{ij-1}) - \nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-1}) \rangle \\
\geq \langle \nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-2}) \rangle \\
- L \| w_{j-1} - w_{j-2} \|_2 \left( \| \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2 + \| \nabla F_S(w_{j-1}) \|_2 \right).
\end{align*}
\]

According to Schwartz inequality, the variance assumption and Eq. (5.10), we know

\[
\begin{align*}
\mathbb{E}_\mathcal{A} \left[ \| w_{j-1} - w_{j-2} \|_2 \left( \| \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2 + \| \nabla F_S(w_{j-1}) \|_2 \right) \right] \\
\leq \frac{1}{2\eta_j} \mathbb{E}_\mathcal{A} \left[ \| w_{j-1} - w_{j-2} \|_2^2 \right] + \eta_j \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2^2 \right] + \| \nabla F_S(w_{j-1}) \|_2^2 \\
\leq \frac{\eta_j}{2} \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-2}; z_{ij-1}, z_{ij-2}) \|_2^2 \right] + \eta_j \mathbb{E}_\mathcal{A} \left[ \| \nabla f(w_{j-2}; z_{ij}, z_{ij-1}) \|_2^2 \right] + \| \nabla F_S(w_{j-1}) \|_2^2 \\
\leq 2\alpha_0^2 \eta_j + 2\alpha_0^2 \eta_j - 1.
\end{align*}
\]
We can combine the above two inequalities together and get

\[ E_A\left[ \nabla f(w_{j-1}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-1}) \right] \geq E_A\left[ \nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-2}) \right] - 4L\alpha_0^2\eta_{j-1}. \]

We can combine (5.9), (5.10) and the above inequality, and get

\[
E_A[F_S(w_j)] \leq E_A[F_S(w_{j-1})] - \eta_j E_A[\nabla f(w_{j-2}; z_{ij}, z_{ij-1}), \nabla F_S(w_{j-2})] + 4L\alpha_0^2\eta_j\eta_{j-1} \\
+ 2L\eta_j^2\alpha_0^2
\]

\[
\leq E_A[F_S(w_{j-1})] - \eta_j E_A[\|\nabla F_S(w_{j-2})\|_2^2] + 4L\alpha_0^2(\eta_j\eta_{j-1} + \eta_j^2),
\]

where the last inequality holds since \(w_{j-2}\) is independent of \(i_j\) and \(i_{j-1}\). The above inequality can be reformulated as

\[
\eta_j E_A[\|\nabla F_S(w_{j-2})\|_2^2] \leq E_A[F_S(w_{j-1})] - E_A[F_S(w_j)] + 4L\alpha_0^2(\eta_j\eta_{j-1} + \eta_j^2). \tag{5.11}
\]

We can take a summation of the above inequality and get

\[
\sum_{j=1}^{t} \eta_j E_A[\|\nabla F_S(w_{j-2})\|_2^2] \leq F_S(w_0) + 4L\alpha_0^2 \sum_{j=1}^{t} (\eta_j\eta_{j-1} + \eta_j^2).
\]

Since \(\eta_j = \eta\), we further get

\[
\sum_{j=1}^{t} E_A[\|\nabla F_S(w_{j-2})\|_2^2] \leq \eta^{-1} F_S(w_0) + 8L\alpha_0^2 t\eta.
\]

The proof is completed. \(\square\)

Our analysis requires the following additional assumption.

**Assumption 5.1** Assume \(F_S\) satisfies the Polyak-Łojasiewicz (PL) condition with parameter \(\mu > 0\), i.e., for \(w_S \in \arg\min_{w \in \mathcal{W}} F_S(w)\), there holds

\[
2\mu(F_S(w) - F_S(w_S)) \leq \|\nabla F_S(w)\|_2^2
\]

for all \(w \in \mathcal{W}\).

The PL condition means that the suboptimality in terms of function values can be bounded by gradients [Karimi et al., 2016]. Functions under the PL condition have found various
applications including neural networks, matrix factorization, generalized linear models and robust regression (see, e.g., [Karimi et al., 2016]). In particular, AUC maximization problem with the classifier given by a one hidden layer network satisfies the PL condition as shown in [Liu et al., 2019].

We now turn to nonconvex problems under a PL condition. Theorem 5.6 gives convergence rates of the order $O(1/t)$, which match the existing results for standard SGD in pointwise learning [Karimi et al., 2016].

**Theorem 5.6** Assume Assumption 4.2 and 5.1 hold true. Assume there is some $\alpha_0 \geq 0$ such that $\mathbb{E}_{i_{j+1},i_{j+2}}[\|\nabla f(w_j; z_{i_{j+1}}, z_{i_{j+2}})\|^2] \leq \alpha_0^2$ for any $j$. Let $w_{-1} = w_0$ and $\{w_j : j \in [t]\}$ be produced by Algorithm 4 with $\eta_j = 2/(\mu(j + 1))$. Then

$$
\mathbb{E}_A[F_S(w_t) - F_S(w_s)] \leq \frac{32L\alpha_0^2}{\mu^2} \left( \frac{1}{t + 1} + \frac{\log(et)}{\mu t (t + 1)} \right).
$$

**Proof:** According to the elementary inequality $\frac{1}{2}(a + b)^2 \leq a^2 + b^2$ we know

$$
\mathbb{E}_A[\|\nabla F_S(w_{j-2})\|^2] \geq -\mathbb{E}_A[\|\nabla F_S(w_{j-2}) - \nabla F_S(w_{j-1})\|^2] + 2^{-1} \mathbb{E}_A[\|\nabla F_S(w_{j-1})\|^2]
$$

$$
\geq -L \mathbb{E}_A[\|w_{j-2} - w_{j-1}\|^2] + 2^{-1} \mathbb{E}_A[\|\nabla F_S(w_{j-1})\|^2]
$$

$$
= -L \eta_{j-1}^2 \mathbb{E}_A[\|\nabla f(w_{j-2}; z_{i_{j-1}}, z_{i_{j-2}})\|^2] + 2^{-1} \mathbb{E}_A[\|\nabla F_S(w_{j-1})\|^2]
$$

$$
\geq -4L \eta_{j-1}^2 \alpha_0^2 + 2^{-1} \mathbb{E}_A[\|\nabla F_S(w_{j-1})\|^2],
$$

where we have used (5.10). This together with (5.11) gives

$$
2^{-1} \eta_j \mathbb{E}_A[\|\nabla F_S(w_{j-1})\|^2] \leq 4L \eta_j \eta_{j-1}^2 \alpha_0^2 + \mathbb{E}_A[F_S(w_{j-1})] - \mathbb{E}_A[F_S(w_j)] + 4L \alpha_0^2 (\eta_j \eta_{j-1} + \eta_j^2).
$$

It then follows from the PL condition that

$$
\mu \eta_j \mathbb{E}_A[F_S(w_{j-1}) - F_S(w_s)] \leq \mathbb{E}_A[F_S(w_{j-1})] - \mathbb{E}_A[F_S(w_j)] + 4L \alpha_0^2 (\eta_j \eta_{j-1} + \eta_j^2 + \eta_j \eta_j^2).
$$

We can reformulate the above inequality as

$$
\mathbb{E}_A[F_S(w_j) - F_S(w_s)] \leq (1 - \mu \eta_j) \mathbb{E}_A[F_S(w_{j-1}) - F_S(w_s)] + 4L \alpha_0^2 (\eta_j \eta_{j-1} + \eta_j^2 + \eta_j \eta_j^2-1).
$$
Now, taking \( \eta_j = 2/ (\mu (j+1)) \), we get
\[
\mathbb{E}_A[F_S (w_j) - F_S (w_S)] \leq \frac{j-1}{j+1} \mathbb{E}_A[F_S (w_{j-1}) - F_S (w_S)] + \frac{4L\alpha_0^2}{\mu^2} \left( \frac{8}{j(j+1)} + \frac{8}{j^2(j+1)\mu} \right).
\]

We can multiple both sides by \( j(j+1) \) and get
\[
j(j+1)\mathbb{E}_A[F_S (w_j) - F_S (w_S)] \leq (j-1)j\mathbb{E}_A[F_S (w_{j-1}) - F_S (w_S)] + \frac{4L\alpha_0^2}{\mu^2} (8 + 8j^{-1}\mu^{-1}).
\]

Taking a summation of the above inequality gives
\[
t(t+1)\mathbb{E}_A[F_S (w_t) - F_S (w_S)] \leq \frac{32L\alpha_0^2}{\mu^2} \sum_{j=1}^{t} (1 + j^{-1}\mu^{-1}).
\]

The stated bound then follows. The proof is completed.

\[\square\]

### 5.2.3 Excess Generalization Error

In this subsection, we present excess generalization error bounds of Algorithm 4 in terms of the sample size, iteration number and step size, which shows how to tune these parameters to get a model with good generalization.

We first study smooth and non-smooth problems for the convex case, and derive the excess generalization bounds of the order \( \mathcal{O}(1/\sqrt{n}) \) in both cases. We use the notation \( B \asymp \tilde{B} \) if there exist constants \( c_1, c_2 > 0 \) such that \( c_1 \tilde{B} \leq B \leq c_2 \tilde{B} \).

**Theorem 5.7 (Nonsmooth Problems)** Let \( w_{-1} = w_0 \) and \( \{w_t : t \in [T]\} \) be produced by Algorithm 4 with \( \eta_t = \eta > 0 \). Let \( \bar{w}_T = \sum_{t=1}^{T} \eta_t w_{t-2}/ \sum_{t=1}^{T} \eta_t \). Let Assumption 4.1 and 4.3 hold true with \( \alpha = 0 \). Then, we have

\[
\mathbb{E}_{S,A}[F(\bar{w}_T)] - F(w^*) = \mathcal{O}\left( \sqrt{T \eta} + \frac{T \eta}{n} + \frac{1 + T \eta^2}{T \eta} \right).
\]  \hfill (5.12)

Furthermore, selecting \( T \asymp n^2 \) and \( \eta \asymp T^{-\frac{3}{4}} \) yields that \( \mathbb{E}_{S,A}[F(\bar{w}_T)] - F(w^*) = \mathcal{O}(1/\sqrt{n}) \).

**Proof:** According to Theorem 5.1, we know
\[
\mathbb{E}_{S,A}[F(\bar{w}_T) - F_S(\bar{w}_T)] = \mathcal{O}\left( \sqrt{T \eta} + \frac{T \eta}{n} \right).
\]
Furthermore, according to Part (a) of Theorem 5.3 with $w^* = w$, we know
\[
\mathbb{E}_{\mathcal{S},\mathcal{A}}[F_S(\bar{w}_T) - F_S(w^*)] = \mathcal{O}\left(\frac{1 + T\eta^2}{T\eta}\right). \tag{5.13}
\]

We can plug the above generalization error bound and optimization error bound back into the error decomposition (5.2), and get (5.12). Taking $T \asymp n^2$ and $\eta \asymp T^{-\frac{3}{4}}$ in Eq. (5.12), we immediately get $\mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\bar{w}_T)] - F(w^*) = \mathcal{O}(1/\sqrt{n})$. The desired result is proved.

**Theorem 5.8 (Smooth Problems)** Let Assumption 4.1, 4.2 and 4.3 hold true with $\alpha = 0$.

Let $\bar{w}_T = \sum_{t=1}^{T} \eta_t w_{t-2}/\sum_{t=1}^{T} \eta_t$. Then, there holds
\[
\mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\bar{w}_T)] - F(w^*) = \mathcal{O}\left(\frac{T\eta}{n} + \frac{1 + T\eta^2}{T\eta}\right). \tag{5.14}
\]

Furthermore, choosing $T \asymp n$ and $\eta \asymp T^{-\frac{3}{4}}$ implies that $\mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\bar{w}_T)] - F(w^*) = \mathcal{O}(1/\sqrt{n})$.

**Proof:** According to Theorem 5.2, we know
\[
\mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\bar{w}_T) - F(w^*)] = \mathcal{O}\left(\frac{T\eta}{n}\right).
\]

We can plug the above generalization error bound and the optimization error bound (5.13) back into the error decomposition (5.2), and get (5.14). Taking $T \asymp n^2$ and $\eta \asymp T^{-\frac{3}{4}}$ in Eq. (5.14), we immediately get $\mathbb{E}_{\mathcal{S},\mathcal{A}}[F(\bar{w}_T)] - F(w^*) = \mathcal{O}(1/\sqrt{n})$. The proof is completed.

**Remark 5.5** Notice that the gradient complexity (i.e. the number of computing gradients) of Algorithm 4 is identical to the number of iterations $T$. The above results show, to get excess generalization bounds $\mathcal{O}(1/\sqrt{n})$, that Algorithm 4 requires a gradient complexity $\mathcal{O}(n^2)$ for nonsmooth problems, and $\mathcal{O}(n)$ for smooth problems. This matches the existing generalization analysis for pointwise learning [Bassily et al., 2020, Hardt et al., 2016, Lei and Ying, 2020]. In Appendix 5.4, additional results are provided where we propose Algorithm 7 based on the iterative localization technique [Feldman et al., 2020] in order to reduce the gradient complexity $\mathcal{O}(n^2)$ required in Theorem 5.7 to $\mathcal{O}(n)$ for nonsmooth problems.

**Remark 5.6** As stated in the introduction, Algorithm 4 can be considered as a specific case of the classic pairwise learning algorithm [Kar et al., 2013] with a FIFO buffering set $B_{t-1}$ of
A key difficulty in the generalization analysis is that $w_{t-1}$ depends on $B_{t-1}$, which renders the standard martingale analysis not applicable. [Kar et al. 2013] proposed to remove this coupling effect by considering $\sup_{w} \left[ f(w; B_{t-1}) - F_{S}(w) \right]$, which is why they only derived the excess generalization error bound $O(1/\sqrt{s})$. We introduce novel techniques to handle the coupling in both generalization analysis and optimization error analysis. For the generalization analysis, our strategy is to write the stability as a deterministic function of several indicator functions on whether we select the different point in neighboring datasets, and then finally consider the randomness of these indicator functions. This delay of considering expectation successfully decouples the coupling between $w_{t-1}$ and $B_{t-1}$. For the optimization error analysis, our novelty is to observe that $f(w_{t-1}; z_{i_t}, z_{i_{t-1}}) = f(w_{t-2}; z_{i_t}, z_{i_{t-1}}) + O(\eta_{t-1})$, which removes the decoupling since $w_{t-2}$ is now independent of both $i_t$ and $i_{t-1}$. Since the additional term $O(\eta_{t-1})$ here is a term of smaller magnitude, the coupling effect is removed without incurring any additional cost.

Finally, we consider the generalization analysis for nonconvex problems under the PL condition. To prove Theorem 5.9, we first introduce a lemma motivated by the arguments in [Hardt et al., 2016].

**Lemma 5.4** Let $S = \{z_i\}_{i \in [n]}$ and $S' = \{z'_i\}_{i \in [n]}$ be neighboring datasets differing by a single example. Let $\{w_t\}_t$ and $\{w'_t\}_t$ be produced by Algorithm 4 w.r.t. $S$ and $S'$, respectively. Let Assumption 4.1 hold and $\sup_{z, z'} f(w, z, z') \leq B$. Let $\Delta_t = \|w_t - w'_t\|_2$. Then for every $z, z' \in Z$ and every $t_0 \in [n]$, there holds

$$\mathbb{E} \left[ |f(w_T; z, z') - f(w'_T; z, z')| \right] \leq G \mathbb{E} [\Delta_T | \Delta_{t_0} = 0] + \frac{Bt_0}{n}.$$  

**Proof:** Without loss of generality, we assume that $S$ and $S'$ differ by the last example. Let $\mathcal{E}$ denote the event $\Delta_{t_0} = 0$. Then we have

$$\mathbb{E} \left[ |f(w_T; z, z') - f(w'_T; z, z')| \right] = \mathbb{E} \left[ |f(w_T; z, z') - f(w'_T; z, z')| \right] \mathbb{P}\{\mathcal{E}\}$$

$$+ \mathbb{E} \left[ |f(w_T; z, z') - f(w'_T; z, z')| \right] \mathbb{P}\{\mathcal{E}^c\},$$
where $\mathcal{E}^c$ denotes the complement of $\mathcal{E}$. Furthermore, we know
\[
\mathbb{P}\{\mathcal{E}^c\} \leq \sum_{t=1}^{t_0} \mathbb{P}\{i_t = n\} = \frac{t_0}{n}.
\]
We can combine the above two inequalities and the Lipschitz continuity of $f$ to derive the stated bound, which completes the proof.

Finally, we study nonconvex pairwise learning under the PL condition.

**Theorem 5.9**  Let Assumption 4.1, 4.2 and 5.1 hold true. Let $\alpha_0, B > 0$. Let $w_{-1} = w_0$ and $\{w_j : j \in [t]\}$ be produced by Algorithm 4 with $\eta_j = 2/(\mu(j + 1))$. If $\sup_{z, z'} f(w_j; z, z') \leq B$ and $\mathbb{E}_{i_{j+2}, i_{j+2}} \|\nabla f(w_j; z_{i_{j+1}}, z_{i_{j+2}})\|^2 \leq \alpha_0^2$ for any $j$, then
\[
\mathbb{E}[F(w_T)] - F(w^*) = O\left(\frac{T^{2L+\mu}}{n}\right) + O\left(1/(T\mu^2)\right).
\]
Furthermore, choosing $T \asymp n^{-\frac{2L+\mu}{4L+2\mu}} \mu^{-\frac{4L+2\mu}{4L+2\mu}}$ yields that $\mathbb{E}[F(w_T)] - F(w^*) = n^{-\frac{2L+\mu}{4L+2\mu}} \mu^{-\frac{4L+2\mu}{4L+2\mu}}$.

**Proof:** Let $S' = \{z_1, \ldots, z_{n-1}, z_n'\}$, where $z'_n$ is independently drawn from $\mathcal{D}$. Let $\{w'_k\}$ be produced by Algorithm 4 w.r.t. $S'$. If $i_t \neq n$ and $i_{t-1} \neq n$, then
\[
\|w_t - w'_t\|_2 \leq \|w_{t-1} - \eta_t \nabla f(w_{t-1}; z_{i_{t-1}}, z_{i_{t-1}}) - w'_{t-1} - \eta_t \nabla f(w'_{t-1}; z_{i_{t-1}}, z_{i_{t-1}})\|_2 \\
\leq \|w_{t-1} - w'_{t-1}\|_2 + \eta_t \|\nabla f(w_{t-1}; z_{i_{t-1}}, z_{i_{t-1}}) - \nabla f(w'_{t-1}; z_{i_{t-1}}, z_{i_{t-1}})\|_2 \\
\leq (1 + L\eta_t)\|w_{t-1} - w'_{t-1}\|_2,
\]
where in the last inequality we used the smoothness of $f$.

Otherwise, it follows from the Lipschitz condition that $\|w_t - w'_t\|_2 \leq \|w_{t-1} - w'_{t-1}\|_2 + 2G\eta_t$. Consequently, it follows that
\[
\|w_t - w'_t\|_2 \leq (1 + L\eta_t)\|w_{t-1} - w'_{t-1}\|_2 + 2G\eta_t \mathbb{I}_{[i_t \neq n \text{ and } i_{t-1} \neq n]} \\
\quad + (\|w_{t-1} - w'_{t-1}\|_2 + 2G\eta_t) \mathbb{I}_{[i_t = n \text{ or } i_{t-1} = n]} \\
\quad \leq (1 + L\eta_t)\|w_{t-1} - w'_{t-1}\|_2 + 2G\eta_t \mathbb{I}_{[i_t = n \text{ or } i_{t-1} = n]}.
\]
We can apply the above inequality recursively and get

\[ \Delta_t \leq 2G \sum_{k=t_0+1}^{t} \eta_k \mathbb{I}_{[i_k=n \text{ or } i_{k-1}=n]} \prod_{k'=k+1}^{t} (1 + L\eta_{k'}) + \Delta_{t_0} \prod_{k=t_0+1}^{t} (1 + L\eta_k). \]

Since \( \Delta_{t_0} = 0 \) implies \( i_{t_0} \neq n \), we have

\[ \mathbb{E}[\Delta_t | \Delta_{t_0} = 0] \leq 2G \sum_{k=t_0+1}^{t} \eta_k \mathbb{E}[\mathbb{I}_{[i_k=n \text{ or } i_{k-1}=n]} | \Delta_{t_0} = 0] \prod_{k'=k+1}^{t} (1 + L\eta_{k'}) \]

\[ = 2G \sum_{k=t_0+2}^{t} \eta_k \mathbb{E}[\mathbb{I}_{[i_k=n \text{ or } i_{k-1}=n]} | \Delta_{t_0} = 0] \prod_{k'=k+1}^{t} (1 + L\eta_{k'}) \]

\[ = 2G \sum_{k=t_0+2}^{t} \eta_k \mathbb{E}[\mathbb{I}_{[i_k=n \text{ or } i_{k-1}=n]}] \prod_{k'=k+1}^{t} (1 + L\eta_{k'}), \]

where we have used the independency between \( \Delta_{t_0} \) and \( i_t \) for \( t > t_0 \). It then follows that

\[ \mathbb{E}[\Delta_t | \Delta_{t_0} = 0] \leq 2G \sum_{k=t_0+2}^{t} \eta_k \mathbb{E}[\mathbb{I}_{[i_k=n]} + \mathbb{I}_{[i_{k-1}=n]}] \prod_{k'=k+1}^{t} (1 + L\eta_{k'}) \]

\[ \leq \frac{4G}{n} \sum_{k=t_0+2}^{t} \eta_k \prod_{k'=k+1}^{t} \exp(L\eta_{k'}) \leq \frac{8G}{\mu n} \sum_{k=t_0+2}^{t} \frac{1}{k+1} \exp \left( 2L\mu^{-1} \sum_{k'=k+1}^{t} \frac{1}{k'} \right) \]

\[ \leq \frac{8G}{\mu n} \sum_{k=t_0+2}^{t} \frac{1}{k+1} \exp \left( 2L\mu^{-1} \log(t/k) \right) = \frac{8G}{\mu n} \sum_{k=t_0+2}^{t} \frac{1}{k+1} \left( \frac{t}{k} \right)^{2L\mu^{-1}} \]

\[ \leq \frac{8G}{\mu n} \frac{2L\mu^{-1}}{t_0} \sum_{k=t_0+2}^{t} k^{-1-2L\mu^{-1}} \leq \frac{8G}{\mu n(2L\mu^{-1})} \left( \frac{t}{t_0} \right)^{2L\mu^{-1}}. \]

Here we use \( 1 + x \leq \exp(x) \) and \( \eta_j = \frac{2}{\mu(j+1)} \). We can plug the above inequality back into Lemma 5.4 and derive

\[ \mathbb{E}[f(w_T; z, z') - f(w'_T; z, z')] \leq \frac{4G^2}{nL} \left( \frac{T}{t_0} \right)^{2L\mu^{-1}} + Bt_0. \]

We can choose \( t_0 \approx T^{\frac{2L}{2L+\mu}} \) and get the following generalization error bounds

\[ \mathbb{E}[f(w_T; z, z') - f(w'_T; z, z')] = \mathcal{O} \left( \frac{T^{\frac{2L}{2L+\mu}}}{n} \right). \]
Lemma 5.1 then implies $E[F(w_T) - F_S(w_T)] = O\left(\frac{T^2 + \mu}{n}\right)$. Furthermore, according to Theorem 5.6 we have the following optimization error bounds

$$E_A[F_S(w_T)] - \inf_w[F_S(w)] = O\left(\frac{1}{T\mu^2}\right).$$

The desired result follows by combining the above two inequalities together and using the fact $E[\inf_w[F_S(w)]] \leq E[F_S(w^*)] = F(w^*)$.

**Remark 5.7** As pointed out before, there is no generalization error for the OGD algorithm, i.e. Algorithm 3 as the i.i.d. data is given in a streaming manner and the iteration number equals the number of the available data (i.e. $t = n$). In this setting, the optimization error is identical to the excess generalization error, which will be estimated in Subsection 5.2.2.

### 5.3 Application: Differentially Private SGD for Pairwise Learning

We now use Algorithm 4 and our stability analysis (i.e. Theorem 5.2) to develop a differentially private algorithm for pairwise learning.

**Algorithm 6** Differentially Private Localized SGD for Pairwise Learning

1: **Inputs:** Dataset $S = \{z_i : i \in [n]\}$, parameters $\epsilon, \delta > 0$, and learning rate $\eta$, initial point $w_0$
2: Set $K = \lceil \log_2 n \rceil$ and divide $S$ into $K$ disjoint subsets $\{S_1, \cdots, S_K\}$ where $|S_k| = n_k = 2^{-k}n$.
3: for $k = 1$ to $K$ do
4:   Set $\eta_k = 4^{-k}\eta$
5:   Compute $\bar{w}_k$ by Algorithm 4 based on $S_k$ and initiated at $w_{k-1}$ for $\lceil n_k \log(4/\delta) \rceil$ steps.
6: Set $w_k = \bar{w}_k + u_k$ where $u_k \sim \mathcal{N}(0, \sigma_k^2 I_d)$ with $\sigma_k = 12G\eta_k \log(4/\delta)\sqrt{2\log(2.5/\delta)/\epsilon}$.
7: end for
8: **Outputs:** $w_K$

Our proposed DP algorithm for pairwise learning is described in Algorithm 6 which is inspired by the iterative localization technique [Feldman et al., 2020] for pointwise learning. The privacy and utility guarantees are given by the following theorem. Here $D$ denotes the diameter of $W$. 

158
Lemma 5.5 Let $A: Z^n \rightarrow \mathbb{R}^d$ be a randomized algorithm with $\ell_2$-sensitivity of $\Delta$ with probability at least $1 - \frac{\delta}{2}$. Then the Gaussian mechanism $M(S) = A(S) + u$ where $u \sim \mathcal{N}(0, (2\Delta^2 \log(2.5/\delta)/\epsilon^2)I_d)$ satisfies $(\epsilon, \delta)$-DP.

**Proof:** Let $S$ and $S'$ be two neighboring datasets. Denote $E$ as the set when $A$ satisfies $\ell_2$-sensitivity of $\Delta$, i.e. $E = \{\|A(S) - A(S')\|_2 \leq \Delta\}$. Then we know $\mathbb{P}[E] \geq 1 - \frac{\delta}{2}$. In favor of $E$, by classical results for Gaussian mechanism, we know $M$ satisfies $(\epsilon, \delta/2)$-DP with $\sigma = \Delta \sqrt{2 \log(2.5/\delta)/\epsilon}$. Therefore, for any $\epsilon > 0$ and any event $O$ in the output space of $M$, we have

$$\mathbb{P}[M(S) \in O] = \mathbb{P}[M(S) \in O | E] \mathbb{P}[E] + \mathbb{P}[M(S) \in O | E^c] \mathbb{P}[E^c]$$

$$\leq (e^\epsilon \mathbb{P}[M(S') \in O | E] + \frac{\delta}{2}) \mathbb{P}[E] + \frac{\delta}{2}$$

$$\leq e^\epsilon \mathbb{P}[M(S') \in O \cap E] + \frac{\delta}{2} + \frac{\delta}{2}$$

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

where the first inequality is because $M$ satisfies $(\epsilon, \delta/2)$-DP when $E$ occurs and $\mathbb{P}[E^c] \leq \delta/2$, the second and third inequalities are by basic properties of probability. The proof is completed. 

\[\square\]

Lemma 5.6 Let $X_1, \ldots, X_t$ be independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{j=1}^t X_j$ and $\mu = \mathbb{E}[X]$. Then for any $\tilde{\gamma} > 0$, with probability at least $1 - \exp\left(-\mu\tilde{\gamma}^2/(2 + \tilde{\gamma})\right)$ we have $X \leq (1 + \tilde{\gamma})\mu$.

In order to prove the privacy guarantee and excess generalization bound for Algorithm 6, we also need the following high probability $\ell_2$-sensitivity of the output of Algorithm 4.

 Lemma 5.7 Let $\{\bar{w}_t\}$ and $\{\bar{w}_t'\}$ be the updates by Algorithm 4 based on the neighboring datasets $S$ and $S'$, respectively. If $f$ is convex and $L$-smooth and $\eta_t = \eta \leq 2/L$, then with probability at least $1 - \gamma$ we have

$$\|\bar{w}_t - \bar{w}_t'\|_2 \leq 4G\eta\left(\frac{t}{n} + \log(2/\gamma) + \sqrt{\frac{t\log(2/\gamma)}{n}}\right).$$
Proof: Without loss of generality, we assume the different example between $S$ and $S'$ is the $n$-th item. By the proof of Theorem 5.2, we know

$$\|w_t - w'_t\|_2^2 \leq 2G \sum_{j=1}^{t} \eta_j \left( \mathbb{1}_{[i_j = n]} + \mathbb{1}_{[i_j - 1 = n]} \right).$$

Applying Lemma 5.6, with probability at least $1 - \gamma$ there holds

$$\sum_{j=1}^{t} \left( \mathbb{1}_{[i_j = n]} + \mathbb{1}_{[i_j - 1 = n]} \right) \leq \frac{2t}{n} + 2 \log(2/\gamma) + 2 \sqrt{\frac{t \log(2/\gamma)}{n}}.$$

It then follows from the convexity of $\| \cdot \|_2$ that

$$\|\bar{w}_t - \bar{w}'_t\|_2 \leq 4G \eta \left( \frac{t}{n} + \log(2/\gamma) + \sqrt{\frac{t \log(2/\gamma)}{n}} \right),$$

which implies the desired result.

Theorem 5.10 Let Assumption 4.1, 4.2 and 4.3 hold true with $\alpha = 0$. Let $\{w_k : k \in [K]\}$ be produced by Algorithm 6 with $\eta = \frac{D}{G} \min\left\{ \frac{\log(4/\delta)}{\sqrt{n}}, \frac{\epsilon}{12 \log(4/\delta) \sqrt{2d \log(2.5/\delta)}} \right\} \leq \frac{2}{L}$. Then, Algorithm 6 satisfies $(\epsilon, \delta)$-DP and, with gradient complexity $O(n \log(1/\delta))$, we have the utility bound that

$$E[F(w_K) - F(w^*)] = O(GD\left( \frac{1}{\sqrt{n}} + \frac{\sqrt{d \log(4/\delta)}}{en} \right)).$$

Proof: We first consider the privacy guarantee of Algorithm 6. Since we run Algorithm 4 for $\lceil n_k \log(4/\delta) \rceil$ steps for each $k$, by Lemma 5.7 we know with probability $1 - \delta/2$

$$\|\bar{w}_k - \bar{w}'_k\|_2 \leq 12G \eta_k \log(4/\delta).$$

Therefore, by Lemma 5.5, each iteration $k$ of Algorithm 6 is $(\epsilon, \delta)$-DP. Since the partition of the dataset $S$ is disjoint, and each iteration $k$ of Algorithm 6 we only use one subset, thus the whole process satisfies $(\epsilon, \delta)$-DP.

Next we investigate the utility bound of Algorithm 6. Let $\bar{w}_0 = w^*$ and $u_0 = w_0 - w^*$,
\[
E[F(\mathbf{w}_K) - F(\mathbf{w}^*)] = \sum_{k=1}^{K} E[F(\mathbf{w}_k) - F(\mathbf{w}_{k-1})] + E[F(\mathbf{w}_K) - F(\mathbf{w}_K)]
\] (5.15)

Denote \( F_{S_k} \) be the empirical objective based on sample \( S_k \). For the first term on the RHS of (5.15), we have

\[
E[F(\mathbf{w}_k) - F(\mathbf{w}_{k-1})]
= E[F(\mathbf{w}_k) - F_{S_k}(\mathbf{w}_k)] + E[F_{S_k}(\mathbf{w}_k) - F_{S_k}(\mathbf{w}_{k-1})] + E[F_{S_k}(\mathbf{w}_{k-1}) - F(\mathbf{w}_{k-1})]
\leq 8G^2 \log(4/\delta) \eta_k + \left( \frac{E[\|u_{k-1}\|^2]}{2\eta_k n_k} + \frac{3G^2 \eta_k}{2} \right) \leq \frac{E[\|u_{k-1}\|^2]}{2\eta_k n_k} + 18 \log(4/\delta) G^2 \eta_k,
\]

where the second identity is because \( \mathbf{w}_{k-1} \) is independent of \( S_k \) and the inequality follows from Theorem 5.3 Part (a) and Theorem 5.2. Recall that by definition \( \eta \leq \frac{D \epsilon}{12G \log(4/\delta) \sqrt{2d \log(2.5/\delta)}} \), so that for all \( k \geq 0 \),

\[
E[\|u_k\|^2] = d\sigma^2_k = d\left(\frac{4^{-k} G \eta}{\epsilon}\right)^2 \leq 16^{-k} D^2.
\]

Plugging the above estimate into (5.15) it follows

\[
E[F(\mathbf{w}_K) - F(\mathbf{w}^*)] \leq \sum_{k=1}^{K} \frac{8 \cdot 16^{-k} D^2}{4^{-k} 2^{-k} \eta_m} + 18 \log(4/\delta) 4^{-k} G^2 \eta + 4^{-K} G D
\leq \sum_{k=1}^{K} 2^{-k} \left( \frac{8D^2}{\eta m} + 18 \log(4/\delta) G^2 \eta \right) + \frac{GD}{n^2}
= \left( GD \left( \frac{1}{\sqrt{n}} + \frac{\sqrt{d \log^3(1/\delta)}}{\epsilon n} \right) \right),
\]

where in the second inequality used \( K = \lceil \log_{2} n \rceil \), and the last inequality is due to \( \eta = \frac{D \epsilon}{G \min\left\{ \frac{\log(4/\delta)}{\sqrt{n}}, \frac{1}{12G \log(4/\delta) \sqrt{2d \log(2.5/\delta)}} \right\}} \). The desired excess generalization error bound is proved.

Finally, we investigate the gradient complexity argument. Since we run Algorithm for \( n_k \) at iteration \( k \). Therefore, the total gradient complexity is \( O\left( \sum_{k=1}^{K} n_k \right) = O(n \log(1/\delta)) \). The proof is completed. \( \square \)
The main difference from the pointwise setting in [Feldman et al., 2020] is that Algorithm 6 involves the coupling dependency between \( \{i_t, i_{t-1}\} \) at time \( t \) in Algorithm 4 and \( \{i_{t-1}, i_{t-2}\} \) at time \( t - 1 \), which renders the direct application of the standard concentration inequalities infeasible. We propose a novel decomposition to circumvent this hurdle.

**Remark 5.8** The above bound matches the lower bound given in [Bassily et al., 2014] for \((\epsilon, \delta)\)-differentially private pointwise learning up to a \( \log(1/\delta) \) term. Our utility bound improves over the previous bound \( O(\sqrt{d \log(1/\delta)} \log(n/\delta)/(\sqrt{n} \epsilon)) \) in the work of Huai et al. [2020]. During the preparation of this work, we notice a very recent paper [Xue et al., 2021] also studied the private version of pairwise algorithm by using the localization technique. Their algorithm establishes the optimal rate \( O(1/\sqrt{n} + \sqrt{d \log(1/\delta)}/(en)) \) which, however, needs an expensive gradient complexity \( O(n^3 \log(1/\delta)) \). As a comparison, we achieve nearly optimal utility bound with linear gradient complexity \( O(n \log(1/\delta)) \).

**Remark 5.9** In the next section, we further remove the Assumption 4.2 required in Theorem 5.10 and propose a private algorithm (stated as Algorithm 8 there) that achieves the optimal rate \( O(1/\sqrt{n} + \sqrt{d \log(1/\delta)}/(en)) \) with gradient complexity \( O(n^2 \log(1/\delta)) \). Such bound improves over the previous known results with nonsmooth losses [Yang et al., 2021a] where the utility bound was \( O(\sqrt{d \log(1/\delta)} \log(n/\delta)/(\sqrt{n} \epsilon)) \).

### 5.4 Localization-Based Algorithm on Non-smooth problems

In this section, we provide additional results on how to reduce the gradient complexity \( O(n^2) \) required in Theorem 5.7 to \( O(n) \) for nonsmooth problems. This improvement is attained by Algorithm 7 which is motivated by the iterative localization technique [Feldman et al., 2020].

We provide two technical lemmas before we present the proof of Theorem 5.11.

**Lemma 5.8** Assume Assumption 4.1 and 4.3 hold true with \( \alpha = 0 \) and denote \( \hat{w}_k = \arg \min_w F_k(w; S_k) \), then

\[
\mathbb{E}[\|\hat{w}_k - \bar{w}_k\|_2^2] = O(G^2 \zeta_k^2 n_k).
\]

**Proof:** Note that \( F_k \) is \( \frac{2}{\zeta_k n_k} \)-strongly convex, by the convergence of Algorithm 4.
Algorithm 7 Localized SGD for Pairwise Learning

1: **Inputs:** Dataset \( S = \{ z_i : i = 1, \ldots, n \} \), parameter \( \zeta \), initial point \( w_0 \)
2: Set \( K = \lceil \log_2 n \rceil \) and divide \( S \) into \( K \) disjoint subsets \( \{ S_1, \ldots, S_K \} \) such that \( |S_k| = n_k = 2^{-k}n \)
3: for \( k = 1 \) to \( K \) do
4: \( \zeta_k = 2^{-k} \zeta \)
5: Compute \( \bar{w}_k \in \mathcal{W} \) by Algorithm 4 with step sizes \( \eta_j = \frac{\zeta n_k}{j+1}, j \in [T_k] \) and \( T_k \asymp n_k \) iterations based on the objective \( F_k \) where
   \[
   F_k(w; S_k) = \frac{1}{n_k(n_k - 1)} \sum_{z, z' \in S_k: z \neq z'} f(w; z, z') + \frac{1}{\zeta n_k} \| w - \bar{w}_{k-1} \|_2^2
   \]
6: end for
7: **Outputs:** \( \bar{w}_K \)

in Theorem 5.3 Part (b), we know that

\[
\frac{\alpha_k}{2} \mathbb{E}[\| \bar{w}_k - \hat{w}_k \|_2^2] \leq \mathbb{E}[F_k(\bar{w}_k; S_k) - F_k(\hat{w}_k; S_k)] = O\left( \frac{G^2}{\alpha_k n_k} \right)
\]

which implies

\[
\mathbb{E}[\| \bar{w}_k - \hat{w}_k \|_2^2] = O\left( G^2 \zeta_k^2 n_k \right).
\]

The proof is completed. \( \square \)

**Lemma 5.9** Let Assumption 4.1 and 4.3 hold true with \( \alpha = 0 \). For any \( w \in \mathcal{W} \), we know that

\[
\mathbb{E}[F(\hat{w}_k) - F(w)] \leq \frac{\mathbb{E}[\| \bar{w}_{k-1} - w \|_2^2]}{\zeta_k n_k} + 2G^2 \zeta_k.
\]

**Proof:** Let \( r(w; z, z') = f(w, z, z') + \frac{1}{\zeta n_k} \| w - \bar{w}_{k-1} \|_2^2, R(w) = \mathbb{E}_{z, z'}[r(w; z, z')] \) and \( w_R^* = \arg \min_{w \in \mathcal{W}} R(w) \). By the proof of Theorem 6 in [Shalev-Shwartz et al. 2009], one has that

\[
\mathbb{E}[F(\hat{w}_k) + \frac{1}{\zeta_k n_k} \| \bar{w}_k - \bar{w}_{k-1} \|_2^2 - F(w) - \frac{1}{\zeta_k n_k} \| w - \bar{w}_{k-1} \|_2^2] = \mathbb{E}[R(\hat{w}_k) - R(w)] \leq \mathbb{E}[R(\hat{w}_k) - R(w_R^*)] \leq 2G^2 \zeta_k,
\]
which implies that
\[
\mathbb{E}[F(\hat{w}_k) - F(w)] \leq 2G^2 \zeta_k - \frac{1}{\zeta_k n_k} \mathbb{E}[\|\hat{w}_k - \hat{w}_{k-1}\|^2] + \frac{1}{\zeta_k n_k} \mathbb{E}[\|w - w_{k-1}\|^2]
\]
\[
\leq 2G^2 \zeta_k + \frac{1}{\zeta_k n_k} \mathbb{E}[\|w - w_{k-1}\|^2].
\]

The proof is completed. \(\square\)

The next theorem shows that the empirical risk minimization can imply models with good excess generalization error by Algorithm 7.

**Theorem 5.11** Let Assumption 4.1 and 4.3 hold true with \(\alpha = 0\) and let \(D\) be the diameter of \(\mathcal{W}\). Let \(\{\bar{w}_k : k \in [K]\}\) be produced by Algorithm 7 with \(\zeta = \frac{D}{G\sqrt{n}}\). Then we have the following excess generalization error bounds
\[
\mathbb{E}[F(\bar{w}_K) - F(w^*)] = \mathcal{O}\left(\frac{GD}{\sqrt{n}}\right)
\]
with gradient complexity \(\mathcal{O}(n)\).

**Proof:** Let \(\bar{w}_0 = w^*\), we have
\[
\mathbb{E}[F(\bar{w}_K)] - F(w^*) = \sum_{k=1}^{K} \mathbb{E}[F(\bar{w}_k)] - F(\bar{w}_{k-1})] + \mathbb{E}[F(\bar{w}_K) - F(\bar{w}_K)].
\]

For the first term we have
\[
\sum_{k=1}^{K} \mathbb{E}[F(\bar{w}_k) - F(\bar{w}_{k-1})] \leq \sum_{k=1}^{K} \left(\frac{\mathbb{E}[\|\bar{w}_{k-1} - \hat{w}_{k-1}\|^2]}{\zeta_k n_k} + 2G^2 \zeta_k\right)
\]
\[
= \mathcal{O}\left(\frac{D^2}{\zeta n} + \sum_{k=2}^{K} G^2 \zeta_k + \sum_{k=1}^{K} 2^{-k} G^2 \zeta\right)
\]
\[
= \mathcal{O}\left(\frac{D^2}{\zeta n} + G^2 \zeta\right)
\]
(5.17)

where the first inequality is by Lemma 5.9, the second inequality is by Lemma 5.8 and
ζ = \frac{D}{\sqrt{n}}. For the second term we have

\begin{align*}
\mathbb{E}[F(\bar{w}_K) - F(\hat{w}_K)] & \leq G \mathbb{E}[\|\bar{w}_K - \hat{w}_K\|_2] \leq G \sqrt{\mathbb{E}[\|\bar{w}_K - \hat{w}_K\|^2]} = O(G^2 \zeta_K \sqrt{n_K}) \\
& = O\left(2^{-2k} G^2 \zeta \sqrt{n} \right) = O(G^2 \zeta) \quad (5.18)
\end{align*}

where the first inequality is by G-Lipschitz continuity of F, the second inequality is by Jensen’s inequality, the first identity is by Lemma 5.8 and the second identity is by \( n_k = 2^{-k}n \).

Now putting (5.17) and (5.18) back to (5.16) and using \( \zeta = \frac{D}{\sqrt{n}} \), we derive

\[ \mathbb{E}[F(\bar{w}_K)] - F(w^*) = O\left(\frac{GD}{\sqrt{n}} \right). \]

Finally we investigate the gradient complexity. Since \( F_k \) is \( \frac{2}{\zeta_k n_k} \)-strongly convex, by Theorem 5.3 Part (b), we need to choose \( T_k \approx n_k \) so that Lemma 5.8 holds. Therefore, in total, we require \( O\left( \sum_{k=1}^{K} n_k \right) = O(n) \) gradient complexity, which yields the desired result.

Furthermore, we propose a differentially private algorithm based on iterative localization [Feldman et al., 2020] for nonsmooth pairwise learning problems. The algorithm is presented as follows.

**Algorithm 8** Differentially Private Localized SGD for Pairwise Learning

1: **Inputs:** Dataset \( S = \{z_i : i \in [n]\} \), parameters \( \epsilon, \delta \), and \( \zeta \), initial points \( w_0 \)
2: Set \( K = \lceil \log_2 n \rceil \) and divide \( S \) into \( K \) disjoint subsets \( \{S_1, \ldots, S_K\} \) where \( |S_k| = n_k = 2^{-k}n \).
3: for \( k = 1 \) to \( K \) do
4: \hspace{1em} Set \( \zeta_k = 4^{-k} \zeta \)
5: \hspace{1em} Compute \( \bar{w}_k \in W \) by Algorithm 4 with step sizes \( \eta_j = \frac{\zeta_k n_k}{j+1} \) on objective \( F_k \) such that with prob \( 1 - \delta \),
6: \hspace{2em} \( F_k(\bar{w}_k; S_k) - \min_{w \in W} F_k(w; S_k) \leq G^2 \zeta_k / n_k \)
7: \hspace{2em} where \( F_k(w; S_k) = \frac{1}{n_k(n_k-1)} \sum_{z,z' \in S_k: z \neq z'} f(w; z, z') + \frac{1}{\zeta_k n_k} \|w - w_{k-1}\|^2 \)
8: \hspace{1em} Set \( w_k = \bar{w}_k + u_k \) where \( u_k \sim \mathcal{N}(0, \sigma_k^2 I_d) \) with \( \sigma_k = 4G \zeta_k \sqrt{\log(2.5/\delta)}/\epsilon. \)
9: end for
10: **Outputs:** \( w_K \)

We are now ready to present the privacy guarantee and utility bound of Algorithm 8 in the following theorem. The proof differs from the iterative localization algorithm in
pointwise learning [Feldman et al., 2020] since we employ our high probability convergence results for non-smooth losses in pairwise learning.

**Theorem 5.12** Let Assumption 4.1 and 4.3 hold true with $\alpha = 0$ and let $D$ be the diameter of $W$. Let $\{w_k : k \in [K]\}$ be produced by Algorithm 8 with $\zeta = D \min \left\{ \frac{G}{\sqrt{n}}, \frac{\epsilon}{4\sqrt{d\log(1/\delta)}} \right\}$. Then Algorithm 8 satisfies $(\epsilon, \delta)$-DP. Furthermore we have the following excess generalization error bounds

$$
\mathbb{E}[F(w_K) - F(w^*)] = O\left( GD \left( \frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n} \right) \right)
$$

with no more than $O(n^2 \log(1/\delta))$ stochastic gradient computations.

**Proof:** We first consider the privacy guarantee of Algorithm 8. For any neighboring datasets $S = \{S_1, \ldots, S_K\}$ and $S' = \{S'_1, \ldots, S'_K\}$ differing by one example, where $S'$ follows the same partition as $S$, and $S_i \cap S_j = \emptyset$ if $i \neq j$. Let $\hat{w}_k = \arg \min_w F_k(w; S_k)$ and $\hat{w}'_k = \arg \min_w F_k(w; S'_k)$. We first investigate the $\ell_2$-sensitivity of $\hat{w}_k$. Since $F_k$ is $\alpha_k = \frac{2}{\zeta_k n_k}$-strongly convex, by Theorem 6 in [Shalev-Shwartz et al., 2009] we have

$$
\|\hat{w}_k - \hat{w}'_k\|_2 \leq \frac{4G}{\alpha_k n_k} = 2G\zeta_k,
$$

where $\hat{w}'_k$ is the return from Line 5 in Algorithm 8 based on $F_k(w; S'_k)$. By the strong convexity of $F_k$ again, we have with probability at least $1 - \delta$

$$
\frac{\alpha_k}{2} \|\hat{w}_k - \hat{w}'_k\|_2^2 \leq F_k(\hat{w}_k; S_k) - F_k(\hat{w}'_k; S_k) \leq \frac{G^2 \zeta_k}{n_k}
$$

which implies $\|\hat{w}_k - \hat{w}'_k\|_2 \leq G\zeta_k$. This further implies $\hat{w}_k$ has $\ell_2$-sensitivity of $4G\zeta_k$ with probability $1 - \delta$. Therefore, by Lemma 5.5, each iteration $k$ of Algorithm 8 is $(\epsilon, \delta)$-DP. Since the partition of the dataset $S$ is disjoint, and each iteration $k$ of Algorithm 8 we only use one subset, thus the whole process will still be $(\epsilon, \delta)$-DP.

Next we investigate the utility bound of Algorithm 8. Firstly, for any fixed $w$,

$$
\mathbb{E}[F(\hat{w}_k) - F(w)] = \mathbb{E}[F(\hat{w}_k) - F(\bar{w})] + \mathbb{E}[F(\bar{w}_k) - F(\bar{w})] \\
\leq \frac{\mathbb{E}[\|w_{k-1} - w\|_2^2]}{\zeta_k n_k} + 3G^2 \zeta_k
$$

166
where we used Lemma 5.9 and \( \| \bar{w}_k - \hat{w}_k \|_2 \leq G \zeta_k \). Denote \( \bar{w}_0 = w^* \) and \( u_0 = w_0 - w^* \), we have

\[
\mathbb{E}[F(w_K) - F(w^*)] = \sum_{k=1}^{K} \mathbb{E}[F(\bar{w}_k) - F(\bar{w}_{k-1})] + \mathbb{E}[F(w_K) - F(\bar{w}_K)] \\
\leq \sum_{k=1}^{K} \left( \frac{\mathbb{E}[\|u_{k-1}\|_2^2]}{\zeta_k n_k} + 3G^2 \zeta_k \right) + G \mathbb{E}[\|u_K\|_2]. \tag{5.19}
\]

Recall that by definition \( \zeta \leq \frac{D_\epsilon}{4G \sqrt{d \log(2.5/\delta)}} \), so that for all \( k \geq 0 \), there holds

\[
\mathbb{E}[\|u_k\|_2^2] = d \sigma_k^2 = d \left( \frac{4^{-k} G \zeta}{\epsilon} \right)^2 \leq 16^{-k} D^2.
\]

Plugging the above estimate into (5.19) it follows

\[
\mathbb{E}[F(w_K) - F(w^*)] \leq \sum_{k=1}^{K} 2^{-k} \left( \frac{8D^2}{\zeta n} + 3G^2 \zeta \right) + 4^{-K} GD \\
\leq \sum_{k=1}^{K} 2^{-k} GD \left( \frac{8}{n} \max \left\{ \sqrt{n}, \frac{\sqrt{d \log(1/\delta)}}{\epsilon} \right\} + \frac{1}{2 \sqrt{n}} \right) + \frac{GD}{n^2} \\
\leq 9GD \left( \frac{1}{\sqrt{n}} + \frac{\sqrt{d \log(1/\delta)}}{n \epsilon} \right) + \frac{GD}{n^2}.
\]

This yields the desired utility bound.

Finally, we investigate the gradient complexity argument. Since \( F_k \) is \( \frac{2}{\zeta_k n_k} \)-strongly convex. We know from Theorem 5.4 Part (b), after \( T_k \asymp n_k^2 \log(1/\delta) \) iterations, we have with probability \( 1 - \delta \)

\[
F_k(\bar{w}_k; S_k) - \min_w F_k(w; S_k) = O \left( \frac{G^2 \zeta_k n_k \log(1/\delta)}{n_k^2 \log(1/\delta)} \right) = O \left( \frac{G^2 \zeta_k}{n_k} \right)
\]

which satisfies the requirement at Line 5 of Algorithm 8. Therefore, in total the gradient complexity is of the form \( O \left( \sum_{k=1}^{K} n_k^2 \log(1/\delta) \right) = O(n^2 \log(1/\delta)) \). The proof is completed.

\[\square\]
CHAPTER 6

Conclusion

We investigated the utility of stochastic gradient methods for the minimax problem and the pairwise learning problem via the lens of algorithmic stability. And we extended our analysis to the closely related field of differential privacy. Our contributions can be summarized as follows.

We presented a comprehensive stability and generalization analysis of stochastic algorithms for minimax objective functions. We introduced various generalization measures and stability measures, and present a systematic study on their quantitative relationship. In particular, we obtained the first minimax optimal risk bounds for SGDA in a general convex-concave case, covering both smooth and nonsmooth setting. We also derived the first non-trivial risk bounds for nonconvex-nonconcave problems. Our bounds show how to early-stop the algorithm in practice to train a model with better generalization. Our main results versus related work for this part are listed in Table A.1. Furthermore, 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 in terms of the weak primal-dual population risk while providing $(\epsilon, \delta)$-DP for both smooth and nonsmooth cases. For the nonconvex-strongly-concave case, assuming that the empirical risk satisfies the PL condition we proved DP-SGDA can achieve a utility bound in terms of the excess primal population risk. Our main results versus related work for this part are listed in Table A.2.

There are some interesting problems for further investigation. Our primal generalization bounds require a strong concavity assumption. It is interesting to remove this assumption. On the other front, it remains an open question to us on understanding how the (strong) concavity of dual variables can help generalization in a nonconvex setting. 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.
On the other front, we provide the first-ever-known stability analysis of SGD for pairwise learning with non-smooth losses and obtain optimal excess risk bounds $\tilde{O}(1/\sqrt{n})$. We extend our analysis to unbounded parameter space and achieve a rate of $\tilde{O}(n^{-1/3})$. We apply our stability results to study differentially private SGD algorithms in pairwise learning. Our output perturbation method achieves utility bound $\tilde{O}(\sqrt{d}/(\sqrt{n}\epsilon))$, which matches the previous results in [Huai et al., 2020] for smooth losses. We provide two examples to illustrate our stability and differential privacy results. In particular, the analysis for the example of metric learning shows the advantage of nuclear norm constraint over Frobenius norm constraint which solved an open question raised in [Cao et al., 2016]. Furthermore, we propose simple stochastic and online gradient descent algorithms for pairwise learning. The key idea is to build a gradient estimator by pairing the current instance with the previous instance, which enjoys favorable computation and storage complexity. We leverage the lens of algorithmic stability to study its generalization and apply tools in optimization theory to study its convergence rates for various problems including convex/nonconvex and smooth/nonsmooth settings. We also use our algorithms and stability analysis to develop a new DP algorithm for pairwise learning with differential privacy constraints which significantly improves the existing results. The main difference from pointwise learning in the analysis is the coupling between models and previous instances, which is handled by introducing novel decoupling techniques.

Here we only considered SGD with replacement. It would be interesting to extend our analysis to SGD without replacement which is drawing increasing interests. For future work, it would be interesting to see whether the analysis and results still hold true if the current example $z_i$ in Algorithm 4 is paired with one arbitrary previous example (e.g., $z_{i_1}$). Other future work would be a systematic extension of our algorithms using other acceleration schemes such as momentum and variance reduction techniques.
Table A.1 summarizes our results for generalization bounds of minimax problems. Bounds are stated in expectation or with high probability (H.P.). For risk bounds, the optimal $T$ (number of iterations) is chosen to trade-off generalization and optimization. Here, C-C means convex-concave, C-$\rho$-SC means convex-$\rho$-strongly-concave, $\rho$-SC means nonconvex-$\rho$-strongly-concave, Lip means Lipschitz continuity, S means the smoothness, D means a decay of weak-convexity-weak-concavity parameter along the optimization process as Eq. (2.50) and PL means the two-sided condition as Assumption 2.3. AGDA means Alternating Gradient Descent Ascent and (R)-ESP means the (regularized)-empirical risk saddle point. $c$ is a parameter in the step size and $L$ is given in Assumption 2.2.

Table A.2 summarizes results for DP-SGDA, which 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. $\Delta_w(A_w(S), A_v(S))$ is the weak PD population risk and $R(A_w(S)) - \min_w R(w)$ is the excess primal population risk.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
<th>Assumption</th>
<th>Measure</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESP</td>
<td>Zhang et al. [2021]</td>
<td>$\rho$-SC-SC, Lip</td>
<td>Weak PD Risk</td>
<td>$O(1/(n\rho))$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\rho$-SC-SC, Lip, S</td>
<td>Strong PD Risk</td>
<td>$O(1/(n\rho^2))$</td>
</tr>
<tr>
<td>R-ESP</td>
<td></td>
<td>C-C, Lip</td>
<td>Weak PD Risk</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
<tr>
<td>SGDA</td>
<td>Farnia and Ozdaglar [2021]</td>
<td>$\rho$-SC-SC, Lip, S</td>
<td>Weak PD Generalization</td>
<td>$O(\log(n)/(n\rho))$</td>
</tr>
<tr>
<td>SGDmax</td>
<td></td>
<td>C-C, Lip, S</td>
<td>Weak PD Risk</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
<tr>
<td>PPM</td>
<td></td>
<td>C-C, Lip, S</td>
<td>Weak PD Risk</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
<tr>
<td>SGDA</td>
<td>This work</td>
<td>C-C, Lip (S)</td>
<td>(H.P.) Primal Risk</td>
<td>$O(1/(\sqrt{n}\rho))$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C-C, Lip</td>
<td>H.P. Plain Risk</td>
<td>$O(\log(n)/\sqrt{n})$</td>
</tr>
<tr>
<td>SGDA</td>
<td>Farnia and Ozdaglar [2021]</td>
<td>$\rho$-SC-SC, Lip</td>
<td>Weak PD Risk</td>
<td>$O(\sqrt{\log n}/\sqrt{n})$</td>
</tr>
<tr>
<td>SGDA</td>
<td>This work</td>
<td>Lip, S</td>
<td>Weak PD Generalization</td>
<td>$O(T^{1/2} + 1/n)$</td>
</tr>
<tr>
<td>SGDA</td>
<td>This work</td>
<td>$\rho$-WC-WC, Lip</td>
<td>Weak PD Generalization</td>
<td>$O(T^{2\rho+3}/n^{2\rho+3})$</td>
</tr>
<tr>
<td>AGDA</td>
<td></td>
<td>$\rho$-SC, PL, Lip, S</td>
<td>Primal Risk</td>
<td>$O(n^{1/3} + \sqrt{T/n})$</td>
</tr>
</tbody>
</table>

Table A.1: Summary of Results in Chapter 2

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Assumption</th>
<th>Measure</th>
<th>Rate</th>
<th>Complexity</th>
<th>Simplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSEG</td>
<td>C-C, Lip, S</td>
<td>$\triangle^w(A_w(S), A_v(S))$</td>
<td>$O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)$</td>
<td>$O(n^2)$</td>
<td>Single-loop</td>
</tr>
<tr>
<td>NISPP</td>
<td>C-C, Lip, S</td>
<td>$\triangle^w(A_w(S), A_v(S))$</td>
<td>$O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n^2}\right)$</td>
<td>$O(n^{3/2} \log(n))$</td>
<td>Double-loop</td>
</tr>
<tr>
<td>DP-SGDA</td>
<td>C-C, Lip, S</td>
<td>$\triangle^w(A_w(S), A_v(S))$</td>
<td>$O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)$</td>
<td>$O(n^{3/2})$</td>
<td>Single-loop</td>
</tr>
<tr>
<td></td>
<td>C-C, Lip</td>
<td>$\triangle^w(A_w(S), A_v(S))$</td>
<td>$O\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)$</td>
<td>$O(n^{3/2})$</td>
<td>Single-loop</td>
</tr>
<tr>
<td></td>
<td>PL-SC, Lip, S</td>
<td>$R(A_w(S)) - \min_w R(w)$</td>
<td>$O\left(\frac{1}{n^{1/3}} + \frac{\sqrt{d\log(1/\delta)}}{n^{2/3}\epsilon}\right)$</td>
<td>$O(n^{3/2})$</td>
<td>Single-loop</td>
</tr>
</tbody>
</table>

Table A.2: Summary of Results in Chapter 3
APPENDIX B
Empirical Evaluations

B.1 Experiments for Chapter 2

In this subsection, we report preliminary experimental results to validate our theoretical results. We consider two datasets available at the LIBSVM website: svmguide3 and w5a [Chang and Lin, 2011]. We follow the experimental setup in Hardt et al. [2016] to study how the stability of SGDA would behave along the learning process. To this end, we build a neighboring dataset $S'$ by changing the last example of the training set $S$. We apply the same randomized algorithm to $S$ and $S'$ and get two model sequences $\{(w_t, v_t)\}$ and $\{(w'_t, v'_t)\}$. Then we evaluate the Euclidean distance $\Delta_t = (\|w_t - w'_t\|^2 + \|v_t - v'_t\|^2)^{\frac{1}{2}}$. We consider the SOLAM [Ying et al., 2016] algorithm, which is the SGDA for the solving the problem (1.2) (a minimax reformulation of the AUC maximization problem). We consider step sizes $\eta_t = \eta/\sqrt{T}$ with $\eta \in \{0.1, 0.3, 1, 3\}$. We repeat the experiments 25 times and report the average of the experimental results as well as the standard deviation. In Figure 3.1, we report $\Delta_t$ as a function of the number of passes (the iteration number divided by $n$). It is clear that the Euclidean distance continues to increase during the learning process. Furthermore, the Euclidean distances increase if we consider larger step sizes. This phenomenon is consistent with our stability bounds in Theorem 2.2.

In this section, we investigate the stability of SGDA on a nonconvex-nonconcave problem. We consider the vanilla GAN structure proposed in Goodfellow et al. [2014]. The generator and the discriminator consist of 4 fully connected layers, and use the leaky rectified linear activation before the output layer. The generator uses the hyperbolic tangent activation at the output layer. The discriminator uses the sigmoid activation at the output layer. In order to make experiments more interpretable in terms of stability, we remove all forms of regularization such as the weight decay or dropout in the original paper. In order to truly implement SGDA, we generate only one noise for updating both the discriminator and the generator at each iteration. This differs from the common GAN training strategy.

1The source codes are available at https://github.com/zhenhuan-yang/minimax-stability.
which uses different noises for updating the discriminator and the generator. We employ the \texttt{mnist} dataset \cite{LeCun:98} and build neighboring datasets \( S \) and \( S' \) by removing a randomly chosen datum indexed by \( i \) from \( S \) and \( i + 1 \) from \( S' \). The algorithm is run based on the same trajectory for \( S \) and \( S' \) by fixing the random seed. We randomly pick 5 different \( i \)'s and 5 different random seeds (total 25 runs). The step sizes for the discriminator and the generator are chosen as constants, i.e., \( \eta = 0.0002 \). We compute the Euclidean distance, i.e., Frobenius norm, between the parameters trained on the neighboring datasets. Note that we do not target at optimizing the test accuracy, but give an interpretable visualization to validate our theoretical findings. The results are given in Figure B.2.

It is clear that the parameter distances for both the generator and the discriminator continue to increase during the training process of SGDA, which is consistent with our analysis in Section 2.4.

B.2 Experiments for Chapter 3

In this section, we evaluate the performance of DP-SGDA by taking AUC maximization as an example. 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.
Figure B.2: The parameter distance versus the number of passes. Left: generator, right: discriminator. 'total' is the mean normalized Euclidean distance across all layers and the shaded area is the standard deviation.

B.2.1 Description of Datasets

In experiments, we use three benchmark datasets. Specifically, ijcnn1 dataset from LIBSVM repository, 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 B.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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Classes</th>
<th>#Training Samples</th>
<th>#Testing Samples</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ijcnn1</td>
<td>2</td>
<td>39,992</td>
<td>9,998</td>
<td>22</td>
</tr>
<tr>
<td>MNIST</td>
<td>10</td>
<td>60,000</td>
<td>10,000</td>
<td>784</td>
</tr>
<tr>
<td>Fashion-MNIST</td>
<td>10</td>
<td>60,000</td>
<td>10,000</td>
<td>784</td>
</tr>
</tbody>
</table>

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

B.2.2 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)).
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. [2019], 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. The training settings for NSEG and DP-SGDA on all datasets are shown in Table B.2.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Datasets</th>
<th>Batch Size</th>
<th>Learning Rate</th>
<th>Epochs</th>
<th>Projection Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ori</td>
<td>DP</td>
<td>Ori</td>
</tr>
<tr>
<td>NSEG</td>
<td>ijcnn1</td>
<td>64</td>
<td>300</td>
<td>300</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>MNIST</td>
<td>64</td>
<td>11</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Fashion-MNIST</td>
<td>64</td>
<td>11</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>DP-SGDA (Linear)</td>
<td>ijcnn1</td>
<td>64</td>
<td>300</td>
<td>300</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>MNIST</td>
<td>64</td>
<td>11</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Fashion-MNIST</td>
<td>64</td>
<td>11</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>DP-SGDA (MLP)</td>
<td>ijcnn1</td>
<td>64</td>
<td>3000</td>
<td>3001</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>MNIST</td>
<td>64</td>
<td>900</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Fashion-MNIST</td>
<td>64</td>
<td>900</td>
<td>1000</td>
<td>100</td>
</tr>
</tbody>
</table>

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

B.2.3 DP-SGDA for AUC Maximization

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

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

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

$$F_S(\theta, a, b, v) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{n_+}(h(\theta; x_i) - a)^2\mathbb{I}[y_i = 1] + \frac{1}{n_-}(h(\theta; x_i) - b)^2\mathbb{I}[y_i = -1] + 2(1 + v)\left(\frac{1}{n_-}h(\theta; x_i)\mathbb{I}[y_i = -1] - \frac{1}{n_+}h(\theta; x_i)\mathbb{I}[y_i = 1]\right) - \frac{1}{n}v^2 \right\}$$
Algorithm 9 DP-SGDA for AUC Maximization

1: **Inputs:** Private dataset $S = \{z_i : i \in [n]\}$, privacy budget $\epsilon, \delta$, number of iterations $T$, learning rates $\{\gamma_t, \lambda_t\}_{t=1}^T$, initial points $(\theta_0, a_0, b_0, v_0)$
2: Compute $n_+ = \sum_{i=1}^n [y_i = 1]$ and $n_- = \sum_{i=1}^n [y_i = -1]$
3: Compute noise parameters $\sigma_1$ and $\sigma_2$ based on Eq. (3.2)
4: for $t = 1$ to $T$ do
5: Randomly select a batch $S_t$
6: For each $j \in I_t$, compute gradients based on Eq. (B.1)
7: Sample independent noises $\xi_t \sim \mathcal{N}(0, \sigma^2 I_{d+2})$ and $\zeta_t \sim \mathcal{N}(0, \sigma^2)$
8: Update

\[
\begin{pmatrix}
\theta_{t+1} \\
a_{t+1} \\
b_{t+1}
\end{pmatrix} = \begin{pmatrix}
\theta_t \\
a_t \\
b_t
\end{pmatrix} - \gamma_t \left( \frac{1}{m} \sum_{j \in I_t} \begin{pmatrix}
\nabla_\theta f(\theta_t, a_t, b_t, v_t; z_j) \\
\nabla_a f(\theta_t, a_t, b_t, v_t; z_j) \\
\nabla_b f(\theta_t, a_t, b_t, v_t; z_j)
\end{pmatrix} + \xi_t \right)
\]

\[v_{t+1} = v_t + \lambda_t \left( \frac{1}{m} \sum_{j \in I_t} \nabla_v f(\theta_t, a_t, b_t, v_t; z_j) + \zeta_t \right)\]

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

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

\[
\nabla_\theta f(\theta, a, b, v; z_j) = \frac{2}{n_+} (h(\theta; x_j) - a) \nabla h(\theta; x_j) [y_j = 1]
+ \frac{2}{n_-} (h(\theta; x_j) - b) \nabla h(\theta; x_j) [y_j = -1]
+ 2(1 + v) \left( \frac{1}{n_-} \nabla h(\theta; x_j) [y_j = -1] - \frac{1}{n_+} \nabla h(\theta; x_j) [y_j = 1] \right)
\]

\[
\nabla_a f(\theta, a, b, v; z_j) = \frac{2}{n_+} (a - h(\theta; x_j)) [y_j = 1],
\]

\[
\nabla_b f(\theta, a, b, v; z_j) = \frac{2}{n_-} (b - h(\theta; x_j)) [y_j = -1]
\]

\[
\nabla_v f(\theta, a, b, v; z_j) = 2 \left( \frac{1}{n_-} h(\theta; x_j) [y_j = -1] - \frac{1}{n_+} h(\theta; x_j) [y_j = 1] \right) - \frac{2}{n} v
\] (B.1)

The pseudo-code can be found in Algorithm 9.

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).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ijcnn1</th>
<th>MNIST</th>
<th>Fashion-MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Linear</td>
<td>MLP</td>
<td>Linear</td>
</tr>
<tr>
<td></td>
<td>NSEG</td>
<td>DP-SGDA</td>
<td>NSEG</td>
</tr>
<tr>
<td>Original</td>
<td>92.191</td>
<td>92.448</td>
<td>96.609</td>
</tr>
<tr>
<td>$\epsilon=0.1$</td>
<td>90.106</td>
<td>91.110</td>
<td>92.763</td>
</tr>
<tr>
<td>$\epsilon=0.5$</td>
<td>90.346</td>
<td>91.357</td>
<td>95.840</td>
</tr>
<tr>
<td>$\epsilon=1$</td>
<td>90.355</td>
<td>91.371</td>
<td>96.167</td>
</tr>
<tr>
<td>$\epsilon=5$</td>
<td>90.363</td>
<td>91.383</td>
<td>96.294</td>
</tr>
<tr>
<td>$\epsilon=10$</td>
<td>90.363</td>
<td>91.386</td>
<td>96.297</td>
</tr>
</tbody>
</table>

Table B.3: Comparison of AUC performance in NSEG and DP-SGDA (Linear and MLP settings) on three datasets with different $\epsilon$ and $\delta=1e-6$. The “Original” means no noise ($\epsilon = \infty$) is added in the algorithms.

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 $\delta$ from $10^{-4}, 10^{-5}, 10^{-6}$. Due to space limitation, we only report the performance when $\delta = 1e-6$. More results can be found in Appendix B.3. To estimate the Lipschitz constants $G_{\mathbf{w}}$ and $G_{\mathbf{v}}$ (in Theorem 3.1),
we first run the algorithms without adding noise. Then we calculate the maximum gradient norms of AUC loss w.r.t \( w \) and \( v \) and assign them as \( G_w \) and \( G_v \), respectively. According to these parameters, we calculate the noise parameter \( \sigma \) by applying autodp\(^2\), which is widely used in the existing works [Wang et al. 2019].

**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.

### B.2.4 Results

**General AUC Performance vs Privacy.** The general performance of all algorithms under linear and MLP settings of AUC optimization is shown in Table B.3. Since the standard deviation of the AUC performance is around \([0, 0.1\%]\) and the difference between different algorithms 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 \( \epsilon \), the AUC performance is increased. The reason is that \( \epsilon \) and \( \sigma \) have opposite trends according to equation (3.2). The relation between \( \epsilon \) and AUC score also verifies our Theorem 3.2 and Theorem 3.5. Third, to verify our statement in Remark 3.2, we compare the \( \sigma \) values from NSEG and DP-SGDA on all datasets in Figure B.3(a). From the figure, it is clear that the \( \sigma \) from NSEG is larger than ours in all \( \epsilon \) settings since it is calibrated based on the gradients’ sensitivity from both \( w \) and \( v \). In fact, the sensitivity w.r.t. \( v \) is small as it is a one-dimensional variable for AUC maximization. Therefore, NSEG leads to overestimate on the noise addition towards \( v \). From Table B.3 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

\(^2\)https://github.com/yuxiangw/autodp
of the most important factors affecting the model performance. Therefore, we compare the
AUC performance with respect to the different hidden units in Figure B.3(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 B.3(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 3.1 and Theorem 3.3 we find mini-batch
size can influence the Gaussian noise variances $\sigma_w^2$ and $\sigma_v^2$ 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 B.3(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.

### B.3 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 B.4. From Table B.4, we can find that the performance will be decreased when
decrease the value of $\delta$ in the same $\epsilon$ settings. The reason is that the small $\delta$ is corresponding
to a large value of $\sigma$ based on Theorem 3.1. A large $\sigma$ means a large noise will be added to the
gradients during the training updates. Therefore, the AUC performance will be decreased
as $\delta$ 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 B.2.4.

We also compare the $\sigma$ values from NSEG and DP-SGDA methods on all datasets in
Figure B.4 (a) with setting $\delta=1e^{-5}$ and (b) $\delta=1e^{-4}$. From the figure, it is clear that the $\sigma$ from NSEG is larger than ours in all $\epsilon$ settings. This implies the noise generated from NSEG is also larger than ours.

### Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ijcnn1</th>
<th>MNIST</th>
<th>Fashion-MNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>MLP</td>
<td>Linear</td>
</tr>
<tr>
<td>NSEG</td>
<td>Ours</td>
<td>Ours</td>
<td>NSEG</td>
</tr>
<tr>
<td>Original</td>
<td>92.191</td>
<td>92.448</td>
<td>96.609</td>
</tr>
<tr>
<td>$\epsilon=0.1$</td>
<td>90.231</td>
<td>91.229</td>
<td>94.020</td>
</tr>
<tr>
<td>$\epsilon=0.5$</td>
<td>90.352</td>
<td>91.366</td>
<td>96.108</td>
</tr>
<tr>
<td>$\epsilon=1$</td>
<td>90.358</td>
<td>91.376</td>
<td>96.316</td>
</tr>
<tr>
<td>$\epsilon=5$</td>
<td>90.363</td>
<td>91.385</td>
<td>96.326</td>
</tr>
<tr>
<td>$\epsilon=10$</td>
<td>90.363</td>
<td>91.387</td>
<td>96.329</td>
</tr>
<tr>
<td>$\epsilon=0.1$</td>
<td>90.168</td>
<td>91.169</td>
<td>93.274</td>
</tr>
<tr>
<td>$\epsilon=0.5$</td>
<td>90.349</td>
<td>91.362</td>
<td>96.029</td>
</tr>
<tr>
<td>$\epsilon=1$</td>
<td>90.357</td>
<td>91.373</td>
<td>96.209</td>
</tr>
<tr>
<td>$\epsilon=5$</td>
<td>90.363</td>
<td>91.384</td>
<td>96.300</td>
</tr>
<tr>
<td>$\epsilon=10$</td>
<td>90.363</td>
<td>91.386</td>
<td>96.301</td>
</tr>
<tr>
<td>$\epsilon=0.1$</td>
<td>90.106</td>
<td>91.110</td>
<td>92.763</td>
</tr>
<tr>
<td>$\epsilon=0.5$</td>
<td>90.346</td>
<td>91.357</td>
<td>95.840</td>
</tr>
<tr>
<td>$\epsilon=1$</td>
<td>90.355</td>
<td>91.371</td>
<td>96.167</td>
</tr>
<tr>
<td>$\epsilon=5$</td>
<td>90.363</td>
<td>91.383</td>
<td>96.294</td>
</tr>
<tr>
<td>$\epsilon=10$</td>
<td>90.363</td>
<td>91.386</td>
<td>96.297</td>
</tr>
</tbody>
</table>

Table B.4: Comparison of AUC performance in NSEG and DP-SGDA (Linear and MLP settings) on three datasets with different $\epsilon$ and different $\delta$. The “Original” means no noise ($\epsilon = \infty$) is added in the algorithms.

### B.4 Experiments for Chapter 5

We now report some preliminary experiments on AUC maximization with

$$f(w; (x, y), (x', y')) = \ell(w^\top (x - x'))I_{[y = 1 \land y' = -1]}$$

where $\ell$ is a surrogate loss function, e.g., the hinge loss $\ell(t) = (1 - t)_+$.  

In this section, we provide the experimental details and additional experiments to support our theoretical findings. The datasets we used are from LIBSVM website [Chang and Lin, 2011]. The statistics of the data is included in Table B.5. For data with multiple

---

3 The source codes are available at [https://github.com/zhenhuan-yang/simple-pairwise](https://github.com/zhenhuan-yang/simple-pairwise)
classes, we convert the first half of class numbers to be the positive class and the second half of class numbers to be the negative class.

<table>
<thead>
<tr>
<th></th>
<th>diabtes</th>
<th>german</th>
<th>ijcnn1</th>
<th>letter</th>
<th>mnist</th>
<th>usps</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>768</td>
<td>1,000</td>
<td>49,990</td>
<td>15,000</td>
<td>60,000</td>
<td>7,291</td>
</tr>
<tr>
<td>d</td>
<td>8</td>
<td>24</td>
<td>22</td>
<td>161</td>
<td>780</td>
<td>256</td>
</tr>
</tbody>
</table>

Table B.5: Data Statistics. \( n \) is the number of samples and \( d \) is the number of features.

The purpose of our first experiment is to compare our algorithm, i.e. Algorithm 4, against four existing algorithms for pairwise learning in terms of generalization and CPU running time on several datasets available from the LIBSVM website [Chang and Lin, 2011]. These algorithms are: 1) OLP [Kar et al., 2013] uses a buffer \( B_t \) updated by a variant of Reservoir sampling with replacement where the buffer size is chosen to be 200 in order to guarantee the maximum AUC score as indicated in Kar et al. [2013]; 2) OAMgra [Zhao et al., 2011] is tailored for AUC maximization with the hinge loss which uses buffers by Reservoir sampling. The buffer size is set to be 100 for both positive and negative buffers as suggested in that paper; 3) SGDpair [Lei et al., 2020] randomly pick a pair from \( \binom{n}{2} \) pairs by uniform distribution; 4) SPAUC [Lei and Ying, 2021a], where AUC maximization problem with the least square loss was reformulated as stochastic saddle point (min-max) problem. Note that SPAUC and OAMgra can only apply to AUC maximization problem with the least square loss and hinge loss, respectively.
For each dataset, we have used 80% of the data for training and the remaining 20% for testing. The results are based on 25 runs of random shuffling. The generalization performance is reported using the average AUC score and standard deviation on the test data. To determine proper hyper parameters, we conduct 5-fold cross validation on the training sets: 1) for Algorithm 4 and SGD\textsubscript{pair}, we select step sizes $\eta_t = \eta \in 10^{-3:3}$ and $W$ diameter $D \in 10^{-3:3}$; 2) for OLP we select step sizes $\eta_t = \eta/\sqrt{t}$ where $\eta \in 10^{-3:3}$ and $W$ diameter $D \in 10^{-3:3}$; 3) for OAM\textsubscript{gra} we select learning rate parameter $C \in 10^{-3:3}$; 4) for SPAUC we select step sizes $\eta_t = \eta/\sqrt{t}$ where $\eta \in 10^{-3:3}$.

To validate the generalization ability, the surrogate loss for Algorithm 4, SGD\textsubscript{pair} and OLP is chosen to be the hinge loss. Average AUC scores of different algorithms are listed in Table B.6 where we can see that our algorithm yields competitive generalization performance with OAM and OLP using a large buffering set. In particular, two additional results are added for comparison, i.e. OLP-RS1 and OAM-RS1 which denote the OLP [Kar et al., 2013] and OAM\textsubscript{gra} [Zhao et al., 2011] with Reservoir sampling and the buffering set size $s = 1$, respectively. We can see that OLP-RS1 and OAM-RS1 are inferior to other algorithms. This inferior performance for OLP and OAM with a small buffering set was also observed in the experiments of Kar et al. [2013], Zhao et al. [2011].

Furthermore, we report more plots on CPU running time against the AUC score on different datasets. Figure B.6 contains more convergence plots for the hinge loss. For a fair comparison of Algorithm 4 with SPAUC, the loss function is chosen as the least square loss for Algorithm 4. SGD\textsubscript{pair} and OLP. The results are shown in Figure B.7. We can see there that SPAUC performs very well among most of the datasets. However, this algorithm was designed very specifically for the AUC maximization problem with the least square loss.
while our algorithm can handle any loss functions and any pairwise learning problems. We can also observe that our algorithm and SGD\textsubscript{pair} converge in a similar CPU running time. In fact, Algorithm 4 is slightly faster than SGD\textsubscript{pair} when the number of samples gets larger. This is partly due to different sampling schemes in Algorithm 4 and SGD\textsubscript{pair}. Indeed, at each iteration SGD\textsubscript{pair} picks a random pair of examples from \( \binom{n}{2} \) pairs, while Algorithm 4 only needs to randomly pick one example from \( n \) individual ones. Figure B.8 depicts the CPU times of these two sampling schemes versus the the number of examples \( n \). We can see that, when the sample size \( n \) increases, the sampling scheme used in SGD\textsubscript{pair} needs significantly more time than our algorithm.

To fairly compare the CPU running time, we apply the following uniform setting across all algorithms: 1) \( \mathcal{I} \) is an \( \ell_2 \) ball with the same diameter; 2) the step sizes \( \eta_t = \eta \) which is tuned by cross validation. We report the results in Figure B.5 for the hinge loss. We can see that CPU running time for our algorithm and SGD\textsubscript{pair} are similar while OLP and OAM needs more time to converge. The possible reason behind this is that they have a high gradient.
Next, we investigate our Algorithm 4 in the non-convex setting. To this end, we use the logistic link function $\logit(t) = \left(1 + \exp(-t)\right)^{-1}$ and then the square loss surrogate function $\ell(t) = (1 - t)^2$. That is, the loss function for the problem of AUC maximization becomes $f(w; (x, y), (x', y')) = (1 - \logit(w^\top(x - x')))^2I[y = 1 \land y' = -1]$. Although $f$ is non-convex, it was shown that it satisfies the PL condition [Foster et al., 2018]. The results are reported in Figure B.9 which shows that Algorithm 4 also converges very quickly in this non-convex setting.

Finally, we compare our differentially private algorithm for AUC maximization (i.e. Algorithm 6) with the logistic loss $\ell(t) = \log(1 + \exp(-t))$ against the state-of-art algorithm DPEGD [Xue et al., 2021]. DPEGD used gradient descent and the localization technique to guarantee privacy. Algorithm 4 was used as non-private baseline, i.e. $\epsilon = 0$. Here, $\delta = \frac{1}{n}$ as suggested in the previous work [Xue et al., 2021]. We consider the effect of different privacy budget $\epsilon$'s against the generalization ability. The implementation across all algorithms is based on fixed training size 256. Average AUC scores over 25 times repeated experiments are listed in Table B.7 and B.8 for the datasets of diabetes and german, respectively. These
Figure B.8: CPU running time of different sampling schemes against the sample size $n$

results demonstrate Algorithm 6 achieves competitive performance with DPEGD using full gradient descent.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$\epsilon = 0.2$</th>
<th>$\epsilon = 0.5$</th>
<th>$\epsilon = 0.8$</th>
<th>$\epsilon = 1.0$</th>
<th>$\epsilon = 1.5$</th>
<th>$\epsilon = 2.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our</td>
<td>.690 ± .094</td>
<td>.751 ± .028</td>
<td>.771 ± .016</td>
<td>.783 ± .024</td>
<td>.784 ± .018</td>
<td>.789 ± .018</td>
</tr>
<tr>
<td>DPEGD</td>
<td>.624 ± .109</td>
<td>.727 ± .055</td>
<td>.768 ± .027</td>
<td>.796 ± .011</td>
<td>.797 ± .017</td>
<td>.792 ± .016</td>
</tr>
</tbody>
</table>

Table B.7: Average AUC ± standard deviation on diabetes. Non-Private result is .813 ± .016.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$\epsilon = 0.2$</th>
<th>$\epsilon = 0.5$</th>
<th>$\epsilon = 0.8$</th>
<th>$\epsilon = 1.0$</th>
<th>$\epsilon = 1.5$</th>
<th>$\epsilon = 2.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our</td>
<td>.614 ± .035</td>
<td>.672 ± .064</td>
<td>.721 ± .024</td>
<td>.725 ± .032</td>
<td>.747 ± .019</td>
<td>.749 ± .021</td>
</tr>
<tr>
<td>DPEGD</td>
<td>.598 ± .018</td>
<td>.703 ± .039</td>
<td>.723 ± .029</td>
<td>.742 ± .028</td>
<td>.753 ± .017</td>
<td>.757 ± .018</td>
</tr>
</tbody>
</table>

Table B.8: Average AUC ± standard deviation on german. Non-Private result is .763 ± .016.

We also report the CPU running times of Algorithm 6 and DPEGD. In this setting, we fix the privacy budget $\epsilon = 1$ and vary the training size $n$. The results are reported in Table B.9. These results shows that Algorithm 6 can arrive competitive performance with DPEGD with significantly less CPU running time.
Figure B.9: Convergence of Algorithm 4 for the generalized linear model

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>diabetes</th>
<th>german</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>$n = 100$</td>
<td>$n = 200$</td>
</tr>
<tr>
<td>Our</td>
<td>.709 ± .051</td>
<td>.788 ± .019</td>
</tr>
<tr>
<td></td>
<td>.046 ± .010</td>
<td>.096 ± .019</td>
</tr>
<tr>
<td>DPEGD</td>
<td>.705 ± .070</td>
<td>.772 ± .017</td>
</tr>
<tr>
<td></td>
<td>.421 ± .067</td>
<td>.885 ± .158</td>
</tr>
</tbody>
</table>

Table B.9: Average AUC score and average CPU running time ± standard deviation.
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{A} )</td>
<td>Optimization/Learning algorithm</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Strong convexity parameter of objective</td>
</tr>
<tr>
<td>( D )</td>
<td>Sample distribution</td>
</tr>
<tr>
<td>( \delta )</td>
<td>Privacy violation tolerance</td>
</tr>
<tr>
<td>( \mathbb{E} )</td>
<td>Expectation</td>
</tr>
<tr>
<td>( \ell )</td>
<td>Loss function</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>Privacy budget</td>
</tr>
<tr>
<td>( \eta )</td>
<td>Stepsize parameter</td>
</tr>
<tr>
<td>( \mathbb{I} )</td>
<td>Indicator function</td>
</tr>
<tr>
<td>( \kappa^2 )</td>
<td>Condition number</td>
</tr>
<tr>
<td>( \langle \cdot , \cdot \rangle )</td>
<td>Inner product</td>
</tr>
<tr>
<td>( \mu )</td>
<td>PL condition parameter of objective</td>
</tr>
<tr>
<td>( \nabla )</td>
<td>Derivative</td>
</tr>
<tr>
<td>( \mathcal{N} )</td>
<td>Gaussian distribution</td>
</tr>
<tr>
<td>( \mathcal{O} )</td>
<td>Upper bound</td>
</tr>
<tr>
<td>( \mathbb{P} )</td>
<td>Probability</td>
</tr>
<tr>
<td>( \Pi )</td>
<td>Projection</td>
</tr>
<tr>
<td>( \mathbb{R} )</td>
<td>Set of real numbers</td>
</tr>
<tr>
<td>( \mathcal{R} )</td>
<td>Rademacher complexity</td>
</tr>
<tr>
<td>( \rho )</td>
<td>Strong concavity parameter of objective</td>
</tr>
</tbody>
</table>
\( \sigma^2 \) Gaussian noise variance
\( \Delta^s \) Strong primal-dual risk
\( \Delta^w \) Weak primal-dual risk
\( \varepsilon \) Stability measure
\( \mathcal{V} \) Dual parameter space
\( \mathcal{W} \) Primal parameter space
\( x \) Feature of example
\( \mathcal{X} \) Input space
\( \mathcal{Y} \) Output space
\( z \) Example from sample space
\( Z \) Sample space
\( B \) Variance of empirical gradients
\( D \) Diameter of parameter space
\( d \) Feature dimension
\( e \) Base of natural logarithm
\( F \) Population risk
\( f \) Objective/Loss function
\( F_S \) Empirical risk
\( G \) Lipschitz continuity parameter of objective
\( L \) Strong smoothness parameter of objective
\( m \) Mini-batch size
\( n \) Number of sample
$R$  Primal population risk

$R_S$  Primal empirical risk

$S$  Training dataset

$T$  Number of iterations

$y$  Label of example
Acronyms

DP  Differential Privacy. iii

DP-SGDA  Differentially-Private Stochastic Gradient Descent Ascent. iii

ERM  Empirical Risk Minimization. iii

MLP  Multilayer perception. 174

PL  Polyak-Łojasiewicz. iii

SGD  Stochastic Gradient Descent. iv

SGDA  Stochastic Gradient Descent Ascent. iii

SGM  Stochastic Gradient Method. iii
BIBLIOGRAPHY


Junhong Lin, Yunwen Lei, Bo Zhang, and Ding-Xuan Zhou. Online pairwise learning algorithms with convex loss functions. *Information Sciences*, 406:57–70, 2017.


