

---

# Stability and Generalization of the Decentralized Stochastic Gradient Descent Ascent Algorithm

---

Miaoxi Zhu<sup>1</sup> Li Shen<sup>2\*</sup> Bo Du<sup>1\*</sup> Dacheng Tao<sup>3</sup>

<sup>1</sup> School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China

<sup>2</sup> JD Explore Academy, China <sup>3</sup> The University of Sydney, Australia

{zhumx, dubo}@whu.edu.cn, {mathshenli, dacheng.tao}@gmail.com

## Abstract

The growing size of available data has attracted increasing interest in solving minimax problems in a decentralized manner for various machine learning tasks. Previous theoretical research has primarily focused on the convergence rate and communication complexity of decentralized minimax algorithms, with little attention given to their generalization. In this paper, we investigate the primal-dual generalization bound of the decentralized stochastic gradient descent ascent (D-SGDA) algorithm using the approach of algorithmic stability under both convex-concave and nonconvex-nonconcave settings. Our theory refines the algorithmic stability in a decentralized manner and demonstrates that the decentralized structure does not destroy the stability and generalization of D-SGDA, implying that it can generalize as well as the vanilla SGDA in certain situations. Our results analyze the impact of different topologies on the generalization bound of the D-SGDA algorithm beyond trivial factors such as sample sizes, learning rates, and iterations. We also evaluate the optimization error and balance it with the generalization gap to obtain the optimal population risk of D-SGDA in the convex-concave setting. Additionally, we perform several numerical experiments which validate our theoretical findings.

## 1 Introduction

Minimax problems have shown extensive applications in machine learning, such as adversarial robustness [26, 16], GAN [11], the zero-sum game [27], multi-agent reinforcement learning [33], AUC maximization [41]. Alongside this, as the use of large-scale models has become widespread, distributed learning algorithms have emerged as a noteworthy approach for handling massive amounts of data and model parameters [5, 1]. Without a parameter server [20] aggregating all data from each local agent, decentralized algorithms that do not rely on the central structure can be advantageous when network bandwidth is low or latency is high, and they can also protect data privacy [21]. In this work, we consider the following decentralized minimax stochastic optimization problems:

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}) := \frac{1}{m} \sum_{i=1}^m F_i(\mathbf{x}, \mathbf{y}) := \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [f_i(\mathbf{x}, \mathbf{y}; \xi_i)] \quad (1)$$

where  $m$  denotes the number of agents,  $F_i$  is the local loss function,  $\xi_i$  represents local data stored on agent  $i$ , and  $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ ,  $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ . Note that the data distributions  $\mathcal{D}_i$  may differ across the agents.

The most straightforward algorithm for solving the above stochastic minimax optimization problem is to apply Stochastic Gradient Descent Ascent (SGDA) [14, 22] in a decentralized manner, named

---

\*Corresponding authors.

Table 1: Main results on different cases: SC-SC, C-C, and NC-NC represent strongly-convex-strongly-concave, convex-concave, and nonconvex-nonconcave, respectively.  $\tilde{\mathcal{O}}$  means it contains the logarithmic function.  $C_\lambda$  is a constant concerning about the spectral gap  $1 - \lambda$  of different topology which is defined in Thm. 2.  $T$  represents iterations.  $L$  is Lipschitz constant.  $\mu$  represents the strong convexity and strong concavity parameter.  $n$  denotes the sample size in each node.  $m$  denotes the number of nodes and  $0 < c \leq 1$  is configurable constant.

Cases	Measure	Bound
SC-SC	strong/weak primal-dual generalization gap	$\tilde{\mathcal{O}}\left(\frac{C_\lambda}{T^{L+\mu}} + \frac{T^{1-c}}{n}\right)$ [Thm. 2]
	strong/weak primal-dual population risk	$\mathcal{O}\left(\frac{C_\lambda}{T^{\min\{\frac{1}{2}, L+\mu\}}} + \frac{1}{n}\right)$ [Thm. 3]
C-C	weak primal-dual generalization gap	$\mathcal{O}\left(\frac{1}{(1-\lambda)^T} + \frac{1}{n}\right)$ [Thm. 4]
	weak primal-dual population risk	$\mathcal{O}\left(\frac{1}{(1-\lambda)^{T^{\frac{1}{3}}}} + \frac{T^{\frac{1}{3}}}{n}\right)$ [Thm. 5]
NC-NC	weak primal-dual generalization gap	$\mathcal{O}\left((C_\lambda T^L)^{\frac{1}{c+L}} \left(\frac{m}{n}\right)^{1-\frac{1}{c+L}}\right)$ [Thm. 6]

D-SGDA. Many algorithms [36, 7, 10, 43, 35, 38, 25, 12, 24, 34, 2] have been proposed to solve Problem (1). As for the theoretical part, they mainly focus on analyzing the convergence behavior and communication complexity of their proposed algorithms. Due to the inaccessibility of the data distribution  $\mathcal{D}_i$ , they approximate the expectation value by averaged sum on the training dataset  $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$  with local samples  $\xi_{i,l_i}$  stored in local dataset  $\mathcal{S}_i = \{\xi_{i,l_i}\}_{1 \leq l_i \leq n}$ :

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} F_S(\mathbf{x}, \mathbf{y}), \text{ with } F_S(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^m F_{\mathcal{S}_i}(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l_i=1}^n f_i(\mathbf{x}, \mathbf{y}; \xi_{i,l_i}) \quad (2)$$

However, it is insufficient to evaluate the stochastic algorithm not to consider the generalization performance, which is roughly the gap between Eq. (1) and Eq. (2). Generally speaking, saddle point of  $F_S(\mathbf{x}, \mathbf{y})$  may not be the optimal solution of  $\min_{\mathbf{x}} \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$ . As a result, the model learned by Eq. (2) may not perform well on the test dataset. In fact, the generalization gap is a crucial criterion for us to foresee the performance of the trained model on the unknown dataset. Furthermore, it is quite necessary for us to make a trade-off between the optimization error and the generalization gap to obtain models with optimal population risk (see Eq. (1)).

Concerning the stability and generalization of the minimax problem, several works [19, 30, 9, 42] have studied the generalization gap and population risk of some algorithms, including SGDA, SGDmax, PPM, and AGDA. However, these results cannot be directly extended to the decentralized case due to the additional communication step during the training process. Intuitively, the number of nodes and communication topology in decentralized training may exert a potential influence on the model's generalizability. Note that even for the decentralized minimization problem, the generalization and stability of decentralized SGD are adversely affected by an extra non-vanishing term [32], and the stability usually suffers from a constant term  $\lambda^2$  [44], compared to vanilla SGD. Building upon these findings, we argue that it is worthy to investigate the generalization and stability of D-SGDA for decentralized minimax problems, where there do exist more newly unveiled problems.

To mitigate this theoretical deficiency, we present the first comprehensive analysis of the stability and generalization of D-SGDA for the decentralized minimax problem in this paper. Specifically, we develop a refined stability analysis in a decentralized manner and derive the generalization gap and population risk for D-SGDA under different settings. The main theoretical results are summarized in Table 1. And our main contributions are summarized as follows:

- *First work on the stability and generalization of D-SGDA for decentralized minimax problem.* We extend the concepts of algorithmic stability, which includes argument stability and weak stability, to the decentralized setting. And we establish a universal connection between argument stability and different measures of generalization gap in the framework of decentralization. We propose a subtle technique to distribute the "different" samples in the neighboring datasets among agents by methods of permutation and combination.
- *New theoretical results.* Our theoretical results reveal that decentralized structure does not hurt the stability and generalizability of D-SGDA compared with SGDA and explain how topology of the communication network influences the performance in strongly-convex-strongly-concave,

convex-concave, and nonconvex-nonconcave conditions (see Table 1,2). We also evaluate the optimization error and leverage it with generalization gap to obtain the optimal population risk.

- *Experiments.* We provide several numerical experiments on AUC maximization (C-C) and adversarial learning (NC-NC) in which we vary different factors to support our theoretical findings. The preliminary experimental results align with our theoretical insights.

## 2 Related Work

**Decentralized minimax problem.** Existing works mainly focus on improving the convergence rate and communication complexity. Liu et al. [23] propose DPOSG, which is firstly applied in case of nonconvex-nonconcave, i.e., GAN training, and they prove  $\mathcal{O}(\epsilon^{-12})$  computational complexity and  $\mathcal{O}(\log(1/\epsilon))$  communication complexity on the busiest node. Xian et al. [36] propose DM-HSGD with convergence rate of  $\mathcal{O}(\kappa^3\epsilon^{-3})$  and Chen et al. [7] propose DREAM with communication rounds of  $\mathcal{O}(\kappa^2\epsilon^{-2}/\sqrt{1-\lambda_2})$  in nonconvex-strongly-concave condition. Chen et al. [6] propose SPIDER-GDA and achieve stochastic first-order oracle of  $\mathcal{O}((n + \sqrt{n}\kappa_x\kappa_y^2)\log(1/\epsilon))$  under two-sided PL condition. Rogozin et al. [31] propose a Mirror-prox based algorithm with  $\mathcal{O}(\epsilon^{-1})$  communication complexity in C-C setting. Huang [15], Luo and Ye [25] accelerates by variance reduction. Beznosikov et al. [3] considers time-varying networks with heterogeneous data, Kovalev et al. [18] provides a rigorous complexity for decentralized variational inequalities.

**Stability and generalization.** There are mainly two approaches to investigating the generalization: algorithm-independent generalization, which is also called uniform convergence generalization, and algorithm-dependent generalization respectively. Where the former may degrade to a vacuous conclusion in [28] and we adopt the latter method in our paper which can better explain the generalization behavior of a detailed algorithm. Bousquet and Elisseeff [4] come up with algorithmic stability, Elisseeff et al. [8] extend the concept to randomized algorithms. Hardt et al. [13] further develop the framework by connecting algorithmic stability with the generalization gap. Sun et al. [32] and Zhu et al. [44] extend the generalization and stability analysis to D-SGD. In the minimax problem, Zhang et al. [42] focus on argument stability and prove  $\mathcal{O}(1/n)$  weak and strong generalization bounds for the SC-SC condition; Farnia and Ozdaglar [9] analyze the uniform stability and generalization gap of GDA, GDmax, and PPM (proximal point method) in the case of NC-NC and Lei et al. [19] summarize the connection between different measures of stability and generalization gap and further develop the corresponding high-probability results. Xing et al. [37] specify the generalization gap for adversarial training and Yang et al. [39] investigate the stability-based generalization of SGDA with differential privacy constraints. Ozdaglar et al. [30] propose a new metric to better evaluate the generalization performance even in the case when the existing metric fails.

## 3 Problem Formulation

In this section, we provide the necessary assumptions, notations, terminologies of population risk, generalization gap, and algorithmic stability in decentralized minimax problems.

### 3.1 Basic Assumptions

**Notations.** We use bold lower case to denote vectors and bold upper case to denote matrices.  $\|\cdot\|_2$  means  $\ell_2$  norm for vectors and  $\|\cdot\|_F$  means Frobinius norm for matrices, and we will omit the subscript when the type of norm is clear from the context.  $\mathbf{1}_n \in \mathbb{R}^n$  denotes the all-one vector and  $\lambda_i(\cdot)$  represents the  $i$ -th largest eigenvalue of a matrix.  $[n] := \{1, 2, \dots, n\}$ .

**Assumption 1 (Lipschitz continuous).** Each local function  $f_i$  is differentiable and there exists  $G > 0$  that  $f_i$  is  $G$ -Lipschitz continuous with respect to both  $\mathbf{x}$  and  $\mathbf{y}$  on any given sample  $\xi_i$ , i.e.,

$$|f_i(\mathbf{x}, \mathbf{y}; \xi_i) - f_i(\mathbf{x}', \mathbf{y}'; \xi_i)| \leq G \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\|_2.$$

**Assumption 2 (Lipschitz smooth).** Each local function  $f_i$  is differentiable and there exists  $L > 0$  that  $f_i$  is  $L$ -Lipschitz smooth with respect to both  $\mathbf{x}$  and  $\mathbf{y}$  on any given sample  $\xi_i$ , i.e.,

$$\left\| \begin{pmatrix} \nabla_{\mathbf{x}} f_i(\mathbf{x}, \mathbf{y}; \xi_i) - \nabla_{\mathbf{x}} f_i(\mathbf{x}', \mathbf{y}'; \xi_i) \\ \nabla_{\mathbf{y}} f_i(\mathbf{x}, \mathbf{y}; \xi_i) - \nabla_{\mathbf{y}} f_i(\mathbf{x}', \mathbf{y}'; \xi_i) \end{pmatrix} \right\| \leq L \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\|_2.$$

**Definition 1 (Convexity-Concavity).** For each local loss function  $f_i(\mathbf{x}, \mathbf{y}; \xi_i)$ , we say that  $f_i$  is  $\mu_x$ -strongly convex on  $\mathbf{x}$  if for any given  $\mathbf{y}$  and on any given sample  $\xi_i$ , there holds:

$$f_i(\mathbf{x}', \mathbf{y}; \xi_i) \geq f_i(\mathbf{x}, \mathbf{y}; \xi_i) + \nabla_{\mathbf{x}} f_i(\mathbf{x}, \mathbf{y}; \xi_i)^T (\mathbf{x}' - \mathbf{x}) + \frac{\mu_x}{2} \|\mathbf{x}' - \mathbf{x}\|^2, \mu_x \geq 0, \forall \mathbf{x}, \mathbf{x}'.$$

we say that  $f_i$  is  $\mu_y$ -strongly concave on  $\mathbf{y}$  if for any given  $\mathbf{x}$  and on any given sample  $\xi_i$ , there holds:

$$f_i(\mathbf{x}, \mathbf{y}'; \xi_i) \leq f_i(\mathbf{x}, \mathbf{y}; \xi_i) + \nabla_{\mathbf{y}} f_i(\mathbf{x}, \mathbf{y}; \xi_i)^T (\mathbf{y}' - \mathbf{y}) - \frac{\mu_y}{2} \|\mathbf{y}' - \mathbf{y}\|^2, \mu_y \geq 0, \forall \mathbf{y}, \mathbf{y}'.$$

We can call it is convex w.r.t.  $\mathbf{x}$  when  $\mu_x = 0$  and concave w.r.t.  $\mathbf{y}$  when  $\mu_y = 0$ .

**Remark 1.** Assumptions about the Lipschitz continuity and smoothness are commonly used in the context of decentralized minimax optimization problems [42, 9, 32].

### 3.2 Decentralized Stochastic Gradient Descent Ascent (D-SGDA)

In decentralized setting, each node will exchange information alternatively and we represent the communication network between nodes as  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , which is a connected graph with node set  $\mathcal{V} = \{1, 2, \dots, m\}$  and edge set  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ . Specifically,  $(i, l) \in \mathcal{E}$  indicates that agent  $l$  can receive information from agent  $i$  and therefore we symbolize the *in* and *out* neighbors as  $\mathcal{N}^{in}(i) \triangleq \{l \in \mathcal{V}, (l, i) \in \mathcal{E}\}$  and  $\mathcal{N}^{out}(i) \triangleq \{l \in \mathcal{V}, (i, l) \in \mathcal{E}\}$  respectively. In an undirected graph, there is no consideration about the order, thus  $(i, l) \in \mathcal{E}$  implies  $(l, i) \in \mathcal{E}$  and the *in* and *out* neighbors are identical which we will abbreviate as  $\mathcal{N}$  for brief. In our work, we focus on undirected graphs. The communication graph is associated with an adjacency matrix, which is also called a mixing matrix,  $\mathbf{W} = [\omega_{ij}] \in \mathbb{R}^{m \times m}$ . It implies the connection between  $m$  agents that  $\omega_{ij} > 0$  if and only if  $(j, i) \in \mathcal{E}$ , otherwise  $\omega_{ij} = 0$ . And there are some basic assumptions about the mixing matrix which is commonly used in decentralized settings [17, 21, 23, 32].

**Assumption 3 (Mixing matrix).** We assume the mixing matrix  $\mathbf{W} = [\omega_{ik}] \in [0, 1]^{m \times m}$  defined on the graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is a symmetric doubly stochastic matrix, which holds the property that  $\mathbf{W}^T = \mathbf{W}$  and  $\mathbf{W} \mathbf{1}_m = \mathbf{1}_m, \mathbf{1}_m^T \mathbf{W} = \mathbf{1}_m^T$ . Besides, we assume  $\lambda := \max\{|\lambda_2|, |\lambda_m(\mathbf{W})|\} \in (0, 1)$ .

For a symmetric doubly stochastic matrix,  $\mathbf{W}$  holds the property that:  $\lambda_1 = 1$ . For different topologies,  $\lambda \rightarrow 1$  implies the sparsity while  $\lambda \rightarrow 0$  implies the complete connection. Nedić et al. [29] and Ying et al. [40] list upper bounds for the spectral gap  $1 - \lambda$  over the commonly communication network. More knowledge on decentralized optimization is placed in **Appendix A**.

In this paper, we study the decentralized minimax problem solved via D-SGDA (see Algorithm 1). We use the superscript to denote the  $t$ -th iteration and the subscript to denote the  $i$ -th local agent. During iteration, each client first computes its local gradient approximation by  $\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)})$  and  $\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)})$  respectively where  $j_t(i)$  is randomly chosen from  $[n]$ . Then each client communicates with its neighbor  $\mathcal{N}(i)$  and updates by SGDA.

---

#### Algorithm 1 D-SGDA

---

**Initialize:**  $\mathbf{x}_i^0 = 0; \mathbf{y}_i^0 = 0, i = 1, \dots, m$

**for**  $t = 1, 2, \dots, T$  **do**

$$\mathbf{x}_i^{t+1} = P_{\mathcal{X}} \left( \sum_{k \in \mathcal{N}(i)} \omega_{ik} \mathbf{x}_k^t - \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) \right)$$

$$\mathbf{y}_i^{t+1} = P_{\mathcal{Y}} \left( \sum_{k \in \mathcal{N}(i)} \omega_{ik} \mathbf{y}_k^t + \eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) \right)$$

**end for**

**Output:**  $\mathbf{x}^t = \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i^t; \mathbf{y}^t = \frac{1}{m} \sum_{i=1}^m \mathbf{y}_i^t$

---

### 3.3 Generalization Gap

In a sense, we obtain the result by minimizing the empirical one  $F_S$  in Eq. (2), which differs from the population one  $F$  in Eq. (1). So we can not guarantee the same performance on the unknown distribution as on the training dataset. And therefore the gap between the empirical one and the population one reflects the ability of generalization. Unlike the standard learning theory which only contains a single variable that can directly define the population risk and empirical risk by the objective function[4]. Owing to the structure of minimax, there are different methods to define the population and empirical risk as concluded in [19], where primal-dual measure starts from the idea of duality gap in optimization. And we first introduce two types of population risks as follows.

**Definition 2 (Population risk).** For a randomized model  $(\mathbf{x}, \mathbf{y})$ , we define the population risk as:

1. *Weak primal-dual population risk:*  $\Delta^w(\mathbf{x}, \mathbf{y}) = \sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}[F(\mathbf{x}, \mathbf{y}')] - \inf_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}[F(\mathbf{x}', \mathbf{y})]$ .

2. *Strong primal-dual population risk*:  $\Delta^s(\mathbf{x}, \mathbf{y}) = \mathbb{E}[\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathbf{y})]$ .

Here the expectation is taken over the randomness of the model. By replacing function  $F$  with function  $F_S$  (see Eq. (2)) in Def. 2 when considering the empirical risk, we obtain the corresponding *weak primal-dual empirical risk*  $\Delta_S^w(\mathbf{x}, \mathbf{y}) = \sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}[F_S(\mathbf{x}, \mathbf{y}')] - \inf_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}[F_S(\mathbf{x}', \mathbf{y})]$  and *strong primal-dual empirical risk*  $\Delta_S^s(\mathbf{x}, \mathbf{y}) = \mathbb{E}[\sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathbf{y})]$  respectively. Subtracting the empirical risk from population risk, we can define the generalization gap as follows.

**Definition 3 (Generalization gap).** For a randomized model  $(\mathbf{x}, \mathbf{y})$ , we define the corresponding generalization gap as:

1. *Weak primal-dual generalization gap*:  $\epsilon_{gen}^w(\mathbf{x}, \mathbf{y}) = \Delta^w(\mathbf{x}, \mathbf{y}) - \Delta_S^w(\mathbf{x}, \mathbf{y})$ .
2. *Strong primal-dual generalization gap*:  $\epsilon_{gen}^s(\mathbf{x}, \mathbf{y}) = \Delta^s(\mathbf{x}, \mathbf{y}) - \Delta_S^s(\mathbf{x}, \mathbf{y})$ .

**Remark 2.** Notice that we revise the name as generalization gap to avoid misunderstanding since "generalization error" usually refers to the empirical risk in learning theory (see [4]). Our primal target (see Problem (1)) is to obtain small population risk (Def. 2) which can be considered as a summation like  $\Delta^w(\mathbf{x}, \mathbf{y}) = \epsilon_{gen}^w(\mathbf{x}, \mathbf{y}) + \Delta_S^w(\mathbf{x}, \mathbf{y})$ , where generalization gap reflects how well the model generalizes and empirical risk reflects the optimization performance. The strong primal-dual risk is stronger than the weak one due to  $\Delta^w(\mathbf{x}, \mathbf{y}) \leq \Delta^s(\mathbf{x}, \mathbf{y})$  according to Jensen's inequality. But in some cases, it is sufficient to bound weak primal-dual population risk such as the MDP [42].

### 3.4 Algorithmic Stability

Inspired by [13] that the generalization gap of an  $\epsilon$ -stable algorithm can be bounded by  $\epsilon$ . And this connection between stability and generalization for minimax problem is furthermore established in [19]. For a randomized algorithm  $\mathcal{A}$  solving the problem (2), we use  $\mathcal{A}(\mathcal{S}) = (\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S}))$  to denote the output of applying algorithm  $\mathcal{A}$  on dataset  $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ . Let  $\mathbb{E}_{\mathcal{S}}$  and  $\mathbb{E}_{\mathcal{A}}$  denote taking expectation on the randomness of the algorithm  $\mathcal{A}$  and the dataset  $\mathcal{S}$  respectively. Sometimes we omit the subscript as  $\mathbb{E}[\cdot]$  when it is clear from the context. Next, we first refine the definitions of algorithmic stability in a decentralized manner and then provide a connection between algorithmic stability and generalization gap in the framework of decentralization.

**Definition 4 (Decentralized neighboring dataset).** We call  $\mathcal{S}, \mathcal{S}'$  the decentralized neighboring datasets when there are at most one different sample in each local dataset, where  $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_m\}$ ,  $\mathcal{S}' = \{\mathcal{S}'_1, \mathcal{S}'_2, \dots, \mathcal{S}'_m\}$  and each  $\mathcal{S}_i$  and  $\mathcal{S}'_i$  differs by at most one sample.

**Definition 5 (Decentralized algorithmic stability).** For a randomized algorithm  $\mathcal{A}$ , we say:

1.  $\mathcal{A}$  is  $\epsilon$ -**argument stable** if there holds for any neighboring datasets  $\mathcal{S}, \mathcal{S}'$ :

$$\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathcal{A}_x(\mathcal{S}) - \mathcal{A}_x(\mathcal{S}') \\ \mathcal{A}_y(\mathcal{S}) - \mathcal{A}_y(\mathcal{S}') \end{pmatrix} \right\|_2 \leq \epsilon.$$

2.  $\mathcal{A}$  is  $\epsilon$ -**weakly stable** if there holds for any neighboring datasets  $\mathcal{S}, \mathcal{S}'$ :

$$\sup_{\xi} \left[ \sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}_{\mathcal{A}} [f(\mathcal{A}_x(\mathcal{S}), \mathbf{y}'; \xi) - f(\mathcal{A}_x(\mathcal{S}'), \mathbf{y}'; \xi)] + \sup_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}_{\mathcal{A}} [f(\mathbf{x}', \mathcal{A}_y(\mathcal{S}); \xi) - f(\mathbf{x}', \mathcal{A}_y(\mathcal{S}'); \xi)] \right] \leq \epsilon.$$

where  $\xi \triangleq \{\xi_1, \dots, \xi_m\}$  denotes sample index with  $\xi_i \in \mathcal{D}_i$  and  $f(\mathbf{x}, \mathbf{y}; \xi) \triangleq \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}, \mathbf{y}; \xi_i)$ .

And we further specify the stability error as  $\epsilon_{sta}^{arg}(\mathcal{A})$  and  $\epsilon_{sta}^w(\mathcal{A})$  respectively.

**Remark 3.** The definition of neighboring datasets in the decentralized setting can degenerate to the traditional neighboring datasets where there is at most a single different sample between  $\mathcal{S}$  and  $\mathcal{S}'$ . And the refined concepts of algorithmic stability are also fit for classic stability without decentralization in [19]. These facts validate that our definitions above are well-defined. Notice that the argument stability can imply weak stability because of the property of Lipschitz continuity (see Assumption 1). Specifically speaking, when algorithm  $\mathcal{A}$  is  $\epsilon$ -argument stable, then it is  $\sqrt{2}G\epsilon$ -weakly stable. So we will mainly focus on the argument stability in the rest part.

**Theorem 1 (Connection).** For an  $\epsilon$ -argument stable decentralized algorithm  $\mathcal{A}$ , under Assumption 1, we have the following different measures of generalization gap:

- a. *Weak primal-dual generalization gap*:  $\epsilon_{gen}^w(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \leq \sqrt{2}G\epsilon$ .
- b. *Strong primal-dual generalization gap holds under extra Assumption 2* when  $f_i$  is  $\mu_x$ SC- $\mu_y$ SC:  $\epsilon_{gen}^s(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \leq G\sqrt{2 + \frac{2L^2}{\mu^2}}\epsilon$ , where  $\mu \triangleq \min\{\mu_x, \mu_y\}$ .

**Remark 4.** The complete proof is provided in **Appendix D**. We establish the connection between argument stability and a different measure of generalization gap under different constraints, where weak primal-dual generalization gap does not require any convexity or concavity in part **a**. and strong primal-dual generalization gap requires both strong convexity and strong concavity in part **b**. And this connection is not limited to the single decentralized algorithm D-SGDA, but a universal connection for decentralized minimax algorithms. Actually weak stability is sufficient to prove the weak primal-dual generalization gap in part **a**. that when algorithm  $\mathcal{A}$  is  $\epsilon$ -weakly stable, we have  $\mathbb{E}_{\mathcal{A}, \mathcal{S}}[\epsilon_{gen}^w(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S}))] \leq \epsilon$ . And the theorem implies that once we have access to the stability error, we can derive the generalization gap as an accompanying result.

## 4 Theoretical Results on D-SGDA

In this section, we study algorithmic stability and generalization bound in SC-SC, C-C, and NC-NC settings in Sec 4.1, Sec 4.2 and Sec 4.3, respectively. Due to the space limitation, the proofs are placed in the **Appendix E, F, G** respectively.

### 4.1 Results on Strongly-Convex-Strongly-Concave Case

Below, we first characterize the argument stability with fixed and decaying learning rates, respectively.

**Theorem 2 (Argument Stability).** *Under Assumption 1,2,3 when each  $f_i$  is  $\mu_x$ -strongly convex and  $\mu_y$ -strongly concave, we have the argument stability bound for D-SGDA (denoted as  $\mathcal{A}$ ):*

$$\epsilon_{sta}^{arg}(\mathcal{A}) \leq \frac{2G}{n} \sum_{k=0}^{T-1} \eta_k^{max} \prod_{s=k+1}^{T-1} (1-\eta_s^{min}) \frac{L\mu}{L+\mu} + 4GL \sum_{k=1}^{T-1} \left( \eta_k^{max} \sum_{s=0}^{k-1} \eta_s^{max} \lambda^{k-1-s} \right) \prod_{j=k+1}^{T-1} (1-\eta_j^{min}) \frac{L\mu}{L+\mu}.$$

where  $\eta_t^{max} \triangleq \max\{\eta_{x,t}, \eta_{y,t}\}$ ,  $\eta_t^{min} \triangleq \min\{\eta_{x,t}, \eta_{y,t}\}$ ,  $\mu = \min\{\mu_x, \mu_y\}$ . Furthermore,

- for fixed learning rates,  $\epsilon_{sta}^{arg}(\mathcal{A}) \leq 2G \frac{L+\mu}{\eta^{min} L \mu} \left( \frac{2(\eta^{max})^2 L}{1-\lambda} + \frac{\eta^{max}}{n} \right)$ .
- for decaying learning rates with  $\eta_t^{min} = \frac{1}{\mu(t+1)}$  and  $\eta_t^{max} = \frac{1}{\mu(t+1)^c}$ ,  $c \leq 1$ , we have:

$$\epsilon_{sta}^{arg}(\mathcal{A}) \leq \frac{2G}{\mu n T^{\frac{L}{L+\mu}}} \sum_{k=0}^{T-1} \frac{1}{(k+1)^{c-\frac{L}{L+\mu}}} + \frac{4GL}{\mu^2 T^{\frac{L}{L+\mu}}} \sum_{k=1}^{T-1} \frac{1}{(k+1)^{c-\frac{L}{L+\mu}}} \frac{C_\lambda}{k^c}.$$

$$\text{where } C_\lambda \triangleq \frac{(k/e)^c}{\lambda(\ln \frac{1}{\lambda})^c} + \frac{2e^{-1}}{\lambda \ln \frac{1}{\lambda}} + \frac{2^c}{\lambda \ln \frac{1}{\lambda}}.$$

**Remark 5. (i) Bound analysis and comparison.** For case **a**. with fixed learning rates, the argument stability is bounded by  $\mathcal{O}(\frac{\eta}{1-\lambda} + \frac{1}{n})$ , which can reach  $\mathcal{O}(\frac{1}{(1-\lambda)^T} + \frac{1}{n})$  when  $\eta \sim \frac{1}{T}$ . For case **b**. with decaying learning rates, we should require  $2c \geq \frac{L}{L+\mu} + 1$  otherwise the bound can tend to infinity, then we have  $\epsilon_{sta}^{arg}(\mathcal{A}) \leq \frac{2G}{\mu(1-c+\frac{L}{L+\mu})} \frac{T^{1-c}}{n} + \frac{4GLC_\lambda}{\mu^2 T^{\frac{L}{L+\mu}}} \left( \frac{1_{2c > L/(L+\mu)+1}}{2^{c-\frac{L}{L+\mu}-1}} + \ln T \cdot 1_{2c = L/(L+\mu)+1} \right)$ . Both can be bounded by  $\tilde{\mathcal{O}}(\frac{T^{1-c}}{n} + \frac{C_\lambda}{T^{\frac{L}{L+\mu}}})$  with  $\tilde{\mathcal{O}}$  containing the logarithmic function into consideration, which matches the corresponding results for SGDA (see Thm.2.(e) in [19]) except an extra multiplication factor  $C_\lambda$ :  $\tilde{\mathcal{O}}(\frac{1}{\sqrt{T}} + \frac{1}{N})$  with decaying learning rates and  $N$  the total sample size. **(ii) Factor influence.** It is apparent that a smaller stability bound is achieved with larger sample size  $n$ , and smaller learning rates  $\eta$  (which is associated with larger  $T$ ). And notice that the decaying learning rates may slightly underperform than the fixed one, it is easy to explain that at the beginning decaying learning rates are too large and therefore result in weaker stability. The influence of these factors is consistent with vanilla SGDA. **(iii) Effect of topology.** So the major difference lies in  $C_\lambda$  and the number of nodes  $m$ . For  $C_\lambda = \frac{(k/e)^c}{\lambda(\ln \frac{1}{\lambda})^c} + \frac{2e^{-1}}{\lambda \ln \frac{1}{\lambda}} + \frac{2^c}{\lambda \ln \frac{1}{\lambda}}$ , when  $\lambda \rightarrow 1$ , it is bounded by  $\mathcal{O}(\frac{1}{\lambda \ln \frac{1}{\lambda}})$  and when  $\lambda \rightarrow 0$ ,  $C_\lambda$  is bounded by  $\mathcal{O}(\frac{1}{\lambda(\ln \frac{1}{\lambda})^c})$ . We list some common topology with estimated value of  $\lambda$  along with the upper bound of  $C_\lambda$  in Table 2. We can conclude that topology with a denser connection (larger spectral gap  $1 - \lambda$ ) will have a smaller stability error, i.e., be more stable. In extreme conditions, the fully connected network will behave as well as vanilla SGDA, and the stability error of the disconnected network will diverge. Furthermore, under the same topology, fewer nodes will result in better stability.

**Generalization gap.** According to Thm. 1, we can directly derive the weak and strong primal-dual generalization gap of D-SGDA as  $\sqrt{2G}\epsilon_{sta}^{arg}$  and  $G\sqrt{2 + \frac{2L^2}{\mu^2}\epsilon_{sta}^{arg}}$  respectively. Therefore, we hold the same analysis as stability in above Remark 5.

Next, we will first derive the optimization error and then provide the population risk by decomposition  $\Delta^s(\mathbf{x}, \mathbf{y}) = \epsilon_{gen}^s(\mathbf{x}, \mathbf{y}) + \Delta_{\mathcal{S}}^s(\mathbf{x}, \mathbf{y})$ . Population risk is an important evaluation for the performance of a stochastic learning algorithm, which will evaluate how our model obtained by training dataset behave over the whole distribution. Notice that we will use the average output instead of the last iterate in analyzing the optimization errors. We denote that:

$$\mathbf{x}_{ave}^T \triangleq \frac{\sum_{t=0}^{T-1} \eta_{\mathbf{x},t} \mathbf{x}^t}{\sum_{t=0}^{T-1} \eta_{\mathbf{x},t}}, \quad \mathbf{y}_{ave}^T \triangleq \frac{\sum_{t=0}^{T-1} \eta_{\mathbf{y},t} \mathbf{y}^t}{\sum_{t=0}^{T-1} \eta_{\mathbf{y},t}}. \quad (3)$$

**Theorem 3 (Strong primal-dual population risk).** *Under Assumption 1,2,3, when each  $f_i$  is  $\mu_{\mathbf{x}}SC - \mu_{\mathbf{y}}SC$ , we have the strong primal-dual population risk as follows, where  $\eta_t^{max} \triangleq \max\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ ,  $\eta_t^{min} \triangleq \min\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ ,  $\mu = \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ , and  $(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$  is defined in Eq. (3):*

a. for fixed learning rates,

$$\begin{aligned} \Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) &\leq G\sqrt{2 + \frac{2L^2}{\mu^2}} (2G\frac{L + \mu}{\eta^{min}L\mu} (\frac{2(\eta^{max})^2L}{1 - \lambda} + \frac{\eta^{max}}{n})) + \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min}T} \\ &\quad + \eta^{max}G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1 - \lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}}. \end{aligned}$$

b. for decaying learning rates that  $\eta_t^{min} = \frac{1}{\mu(t+1)}$  and  $\eta_t^{max} = \frac{1}{\mu(t+1)^c}$  with  $c \leq 1$  and  $2c \geq \frac{L}{L+\mu} + 1$ ,

$$\begin{aligned} &\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \\ &\leq G\sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G}{\mu(1-c + \frac{L}{L+\mu})} \frac{T^{1-c}}{n} + \frac{4GLC_{\lambda}}{\mu^2 T^{\frac{L}{L+\mu}}} (\frac{\mathbf{1}_{2c \neq L/(L+\mu)+1}}{2c - \frac{L}{L+\mu} - 1} + \ln T \cdot \mathbf{1}_{2c=L/(L+\mu)+1}) \right) \\ &\quad + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2(1 + \ln T)}{2\mu T} + \frac{\mathbf{1}_{c \neq 1}}{(1-c)T^c} + \frac{(1 + \ln T)\mathbf{1}_{c=1}}{T} + \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu T^c} (\frac{\mathbf{1}_{c \neq 1}}{1-c} + \ln T \cdot \mathbf{1}_{c=1}). \end{aligned}$$

**Remark 6.** By Jensen's inequality (see Remark 2), we can conclude that weak primal-dual population risk also satisfies the conclusions above. **(i) Bound analysis and comparison.** For fixed learning rates, it is interesting to see that we should choose  $\eta \sim 1/\sqrt{T}$  to get the optimal population risk of  $\mathcal{O}(\frac{1}{n} + \frac{1}{(1-\lambda)\sqrt{T}})$ , although when  $\eta \sim 1/T$  we can get optimal generalization performance but the optimization error will not converge. While for the decaying learning rates, the population risk bound is  $\tilde{\mathcal{O}}(T^{1-c}/n + C_{\lambda}/T^{\min\{\frac{1}{2}, \frac{L}{L+\mu}\}})$ , which matches the corresponding results for SGDA (see Thm.3.(c) in [19]) of  $\mathcal{O}(\ln N/N\mu)$  when  $n \sim T^{\min\{\frac{1}{2}, \frac{L}{L+\mu}\}}$ . **(ii) Topology influence.** We omit the trivial factor influence analysis here (or see Remark 5). And the effect of topology is captured quantitatively by  $C_{\lambda}$  and  $\frac{1}{1-\lambda}$  which have been discussed in Remark 5 and Table 2.

## 4.2 Results on Convex-Concave Case

In this section, we provide the argument stability and weak primal-dual population risk of D-SGDA algorithm for the NC-NC condition in the following theorems with proof in Appendix F.

**Theorem 4 (Argument Stability).** *Under Assumption 1,2,3, when each  $f_i$  is convex-concave, we have the argument stability bound for D-SGDA (denoted as  $\mathcal{A}$ ):*

$$\epsilon_{sta}^{arg}(\mathcal{A}) \leq \frac{2G}{n} \sum_{k=0}^{T-1} \eta_k^{max} + 4GL \sum_{k=1}^{T-1} \left( \eta_k^{max} \sum_{s=0}^{k-1} \eta_s^{max} \lambda^{k-1-s} \right).$$

Table 2:  $\lambda$  value of different topology. Here 0 means the extra term will disappear and N/A means the term will diverge.

Topology	$\lambda$ ([40])	$C_{\lambda}$	$\frac{1}{1-\lambda}$
fully connected	0	0	0
exponential	$1 - \frac{2}{1+\ln m}$	$\mathcal{O}(\ln m)$	$\mathcal{O}(\ln m)$
grid	$1 - \frac{1}{m \ln m}$	$\mathcal{O}(m \ln m)$	$\mathcal{O}(m \ln m)$
ring	$1 - \frac{16\pi^2}{3m^2}$ ([21])	$\mathcal{O}(m^2)$	$\mathcal{O}(m^2)$
star	$1 - \frac{1}{m^2}$	$\mathcal{O}(m^2)$	$\mathcal{O}(m^2)$
disconnected	1	N/A	N/A

**Remark 7. (i) Bound analysis and comparison.** When there is no strong convexity or strong concavity, we can no longer choose the decaying learning rates, otherwise the argument stability error may not converge. For fixed learning rates,  $\epsilon_{sta}^{arg}$  is upper bounded by  $\mathcal{O}(\frac{\eta T}{n} + \frac{\eta^2 T}{1-\lambda})$ , which is slightly looser than SC-SC condition. While we can still choose  $\eta \sim 1/T$  to obtain optimal result  $\mathcal{O}(\frac{1}{n} + \frac{1}{(1-\lambda)T})$ . Besides, compared with the corresponding result for SGDA (see Thm.2.(b) in [19]) of  $\mathcal{O}(\frac{\sqrt{T}}{N} + \frac{1}{\sqrt{N}})$ , we can approach it when  $\eta \sim 1/T^{3/4}$  and  $n \sim T^{3/4}$ . **(ii) Topology influence.** In C-C condition, the effect of topology on the stability is quantified by  $\frac{1}{1-\lambda}$  which has been discussed in Table 2. And we can conclude that denser topology is more stable and fewer nodes will increase stability under the same topology.

**Generalization gap.** Thm. 1 implies  $\epsilon_{gen}^w(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \leq \sqrt{2}G\epsilon_{sta}^{arg}(\mathcal{A})$ . So the generalization gap holds with the same quantitative analysis as stability above. Analogously we can present the weak primal-dual population risk in the following theorem.

**Theorem 5 (Weak primal-dual population risk).** Under Assumption 1,2,3, when each  $f_i$  C-C, we have the weak primal-dual population risk as follows, where  $\eta_t^{max} \triangleq \max\{\eta_{x,t}, \eta_{y,t}\}$ ,  $\eta_t^{min} \triangleq \min\{\eta_{x,t}, \eta_{y,t}\}$ ,  $\mu = \min\{\mu_x, \mu_y\}$ , and  $(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$  is defined in Eq. (3):

$$\begin{aligned} \Delta^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) &\leq \sqrt{2}G\left(\frac{2G\eta^{max}T}{n} + \frac{4GL(\eta^{max})^2T}{1-\lambda}\right) + \frac{C_x^2 + C_y^2}{2\eta^{min}T} \\ &\quad + \eta^{max}G^2 + \frac{4(C_x + C_y)GL\eta^{max}}{1-\lambda} + \frac{2(C_x + C_y)G}{\sqrt{T}}. \end{aligned}$$

**Remark 8.** The weak primal-dual population risk attains optimal of  $\mathcal{O}(\frac{T^{1/3}}{n} + \frac{1}{(1-\lambda)T^{1/3}})$  when we choose  $\eta^{max} = \eta^{min} \sim 1/T^{\frac{2}{3}}$ . Note that we select  $\eta \sim 1/T$  to obtain optimal generalization performance (see Remark 7), but the optimization error will diverge in that case. Compared with the result of SGDA:  $\mathcal{O}(N^{-1/2})$  (see Thm.3.(b) in [19]), our result can approach it by  $n^{1/2} \sim T^{1/3}$ . Then the effect of topology and number of nodes on the population risk is reflected by  $\frac{1}{1-\lambda}$ , which has been discussed in Remark 5 and Table 2.

### 4.3 Results on Nonconvex-Nonconcave Case

In this section, we present the weak stability and weak primal-dual generalization gap of D-SGDA algorithm for the NC-NC problem in the following theorem with proof in Appendix G.

**Theorem 6 (Weak stability).** Under Assumption 1,2,3, denoting D-SGDA algorithm as  $\mathcal{A}$ , we have the following weak stability bound when  $\eta_t^{max} \triangleq \max\{\eta_{x,t}, \eta_{y,t}\}$ , and  $\eta_t^{min} \triangleq \min\{\eta_{x,t}, \eta_{y,t}\}$ :

a. for fixed learning rates that  $\eta_t^{max} = \eta^{max}$ , and  $\eta_t^{min} = \eta^{min}$ ,

$$\epsilon_{sta}^w(\mathcal{A}) \leq 2\sqrt{2}G^2\left(\frac{\eta^{max}T}{n} + \frac{2L\eta^{max}T}{1-\lambda}\right).$$

b. for decaying learning rates that  $\eta_t^{min} = \frac{1}{t+1}$ , and  $\eta_t^{max} = \frac{1}{(t+1)^c}$ ,  $c \leq 1$ ,

$$\epsilon_{sta}^w(\mathcal{A}) \leq (c+L)(c+L-1)^{\frac{1}{c+L}}\left(\frac{2\sqrt{2}G^2T^L}{(c+L-1)n} + \frac{4\sqrt{2}G^2LC_\lambda T^L}{2c+L-1}\right)^{\frac{1}{c+L}}\left(\frac{Bm}{n}\right)^{1-\frac{1}{c+L}}.$$

**Weak primal-dual generalization gap.** According to Thm. 1, we can derive the weak primal-dual generalization gap as  $\epsilon_{gen}^w(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \leq \sqrt{2}G\epsilon_{sta}^w(\mathcal{A})$  for D-SGDA in NC-NC condition.

**Remark 9.** For case a. with fixed learning rates, the weak stability and weak primal-dual generalization gap is bounded by  $\mathcal{O}(\frac{\eta T}{n} + \frac{\eta^2 T}{1-\lambda})$ , which can reach  $\mathcal{O}(\frac{1}{n} + \frac{1}{(1-\lambda)T})$  when  $\eta \sim \frac{1}{T}$ . For case b. with decaying learning rates, the stability and generalization gap is bounded by  $\mathcal{O}((C_\lambda)^{\frac{1}{c+L}} T^{\frac{L}{c+L}} (\frac{m}{n})^{1-\frac{1}{c+L}})$ . It is evident to analyze the influence of factors that, larger sample size and fewer nodes will result in a smaller stability error and generalization gap, which coincides with results in (S)C-(S)C conditions (see Remark 5). Approaching to the weak primal-dual generalization bound of  $\mathcal{O}(n^{-\frac{2c\rho+1}{2c\rho+3}} T^{\frac{2c\rho}{2c\rho+3}})$  provided  $\rho$ -weakly-convex-weakly-concave (see Thm.5 in [19]).  $C_\lambda$  reflects the effect of topology on the stability and generalization gap and its value has been discussed in Remark 5 and Table 2.

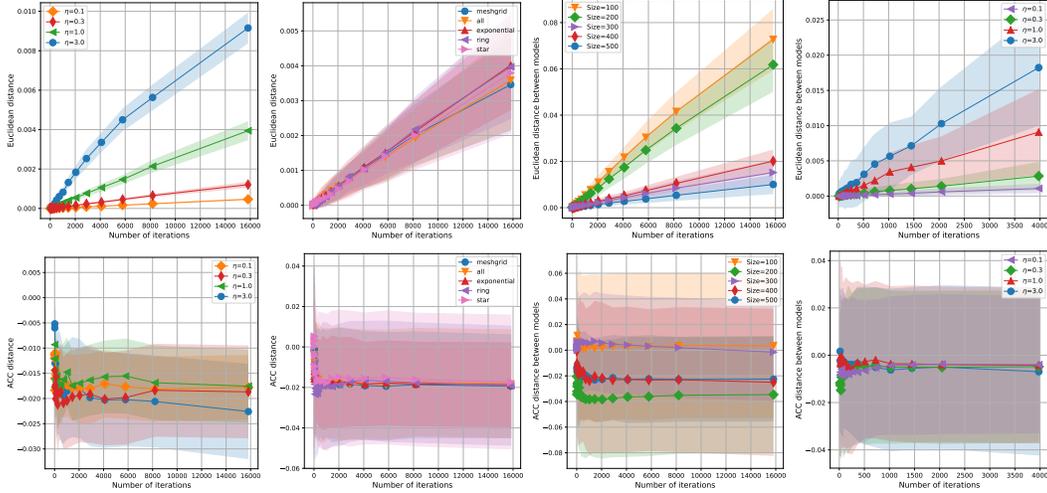


Figure 1:  $\Delta$  against the number of iterations, with the first row showing different settings. From left to right, the settings include varying learning rates, communication typologies, sample sizes on the `w5a` dataset, and learning rates(`svmguid3`). The generalization error is displayed at the bottom accordingly.

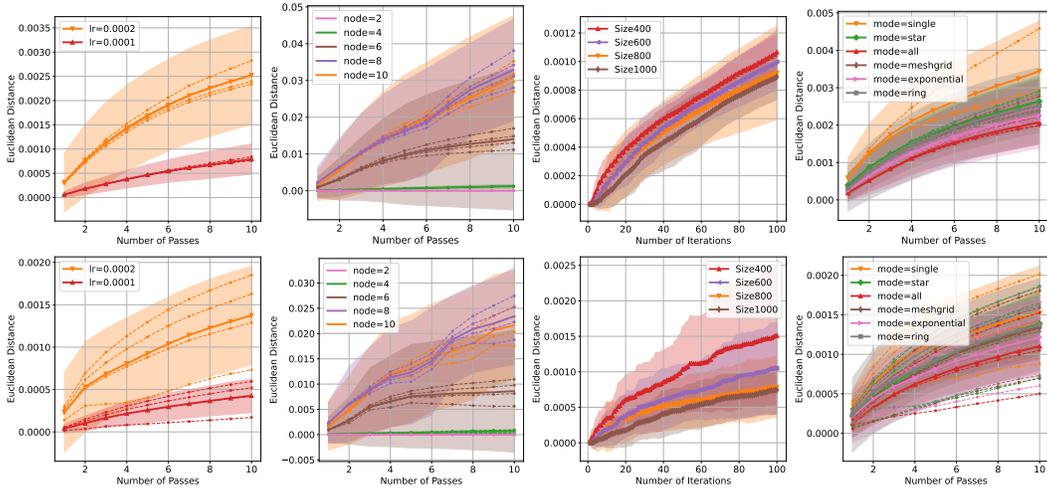


Figure 2: The first row shows the performance of the generator. From left to right, the settings include varying learning rates, the number of nodes, sample sizes, and communication typologies on the `MNIST` dataset. The performance of the discriminator is displayed at the bottom accordingly. The dashes denote different layers.

## 5 Experiments

**Experiments Setup.** We evaluate our theoretical results of the C-C case by adopting the SOLAM method [41] to solve the AUC problem on two datasets `svmguid3` and `w5a`, and the NC-NC case by solving the generative adversarial network on `MNIST`. We extend the methods for both cases to a decentralized implementation. Our experimental setting follows the way conducted in [13, 19] to study how the stability and generalizability of D-SGDA would behave along the learning process with different factors, including learning rates, typologies, nodes, and sample sizes. We employ the same randomized method to generate two model sequences, one for the original data and another for a one-observation perturbing data, and subsequently calculate the Euclidean distance  $\Delta$  between their respective parameter sets. Additional implementation details can be found in the **Appendix**.

**Results analysis:** From Fig. 1 and Fig. 2, we can observe that: (i) faster learning rates, fewer number of nodes and smaller sample size can result in a smaller Euclidean distance between weights and a smaller difference between training dataset and validation dataset; (ii) the performance of different topology on stability: fully connected(all) > exponential > grid  $\approx$  ring  $\geq$  star > disconnected(single).

These validate our theoretical results of the algorithmic stability for D-SGDA (see Remark 5 below Thm. 2). And the impact of different topologies also coincides with our discussion in Table 2.

## 6 Conclusion

In this paper, we provide the first comprehensive analysis for the stability and generalization of D-SGDA for decentralized minimax problems. Our theoretical results show that a decentralized structure does not destroy the stability and generalization of D-SGDA, instead we can leverage between the iterations and the number of nodes, as well as sample size to achieve better population performance. Numerical experiments also validate our theory. Our analysis technique has the potential used for studying the ability and generalization of other decentralized minimax algorithms.

**Limitation&Broader Impacts.** In our analysis, we require the Lipschitz smoothness, which may be further relaxed in future work. In addition, D-SGDA can converge with the heterogeneous data distribution, whose stability and generalization are still unexplored in this work. Since our work focuses on the theoretical understanding of D-SGDA, it does not suffer from negative impacts.

## Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62225113.

The authors appreciate Erdun Gao for providing valuable suggestions on the experimental implementation. The authors also thank Tongtian Zhu for pointing out typos as well as useful discussions on the experimental results.

## References

- [1] Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.
- [2] Aleksandr Beznosikov, Gesualdo Scutari, Alexander Rogozin, and Alexander Gasnikov. Distributed saddle-point problems under data similarity. *Advances in Neural Information Processing Systems*, 34:8172–8184, 2021.
- [3] Aleksandr Beznosikov, Pavel Dvurechenskii, Anastasiia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Advances in Neural Information Processing Systems*, 35:38116–38133, 2022.
- [4] Olivier Bousquet and Andr e Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- [5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [6] Lesi Chen, Boyuan Yao, and Luo Luo. Faster stochastic algorithms for minimax optimization under polyak- $\{L\}$ ojasiewicz condition. *Advances in Neural Information Processing Systems*, 35:13921–13932, 2022.
- [7] Lesi Chen, Haishan Ye, and Luo Luo. A simple and efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2212.02387*, 2022.
- [8] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbling. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.
- [9] Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pages 3174–3185. PMLR, 2021.

- [10] Hongchang Gao. Decentralized stochastic gradient descent ascent for finite-sum minimax problems. *arXiv preprint arXiv:2212.02724*, 2022.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [12] Hongyi Guo, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Decentralized single-timescale actor-critic on zero-sum two-player stochastic games. In *International Conference on Machine Learning*, pages 3899–3909. PMLR, 2021.
- [13] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [15] Feihu Huang. Adaptive federated minimax optimization with lower complexities. *arXiv preprint arXiv:2211.07303*, 2022.
- [16] Zhuo Huang, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, and Tongliang Liu. Robust generalization against photon-limited corruptions via worst-case sharpness minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16175–16185, 2023.
- [17] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*.
- [18] Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, and Alexander Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. *arXiv preprint arXiv:2202.02771*, 2022.
- [19] Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.
- [20] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 583–598, 2014.
- [21] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.
- [22] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.
- [23] Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, and Payel Das. A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33:11056–11070, 2020.
- [24] Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8767–8775, 2021.
- [25] Luo Luo and Haishan Ye. Decentralized stochastic variance reduced extragradient method. *arXiv preprint arXiv:2202.00509*, 2022.

- [26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.
- [27] Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.
- [28] Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [29] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.
- [30] Asuman E Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. What is a good metric to study generalization of minimax learners? In *Advances in Neural Information Processing Systems*.
- [31] Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, Dmitry Kovalev, Pavel Dvurechensky, and Alexander Gasnikov. Decentralized distributed optimization for saddle point problems. *arXiv preprint arXiv:2102.07758*, 2021.
- [32] Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9756–9764, 2021.
- [33] Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. *Advances in Neural Information Processing Systems*, 31, 2018.
- [34] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.
- [35] Xidong Wu, Zhengmian Hu, and Heng Huang. Decentralized riemannian algorithm for nonconvex minimax problems. *arXiv preprint arXiv:2302.03825*, 2023.
- [36] Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34: 25865–25877, 2021.
- [37] Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021.
- [38] Yangyang Xu. Decentralized gradient descent maximization method for composite nonconvex strongly-concave minimax problems. *arXiv preprint arXiv:2304.02441*, 2023.
- [39] Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Vashney, Siwei Lyu, and Yiming Ying. Differentially private sgd for minimax problems. In *Uncertainty in Artificial Intelligence*, pages 2192–2202. PMLR, 2022.
- [40] Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34:13975–13987, 2021.
- [41] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016.
- [42] Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR, 2021.

- [43] Xin Zhang, Zhuqing Liu, Jia Liu, Zhengyuan Zhu, and Songtao Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:18825–18838, 2021.
- [44] Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized sgd. In *International Conference on Machine Learning*, pages 27479–27503. PMLR, 2022.

# Appendix

## Table of Contents

---

<b>A Decentralized Optimization</b>	<b>14</b>
<b>B Notation</b>	<b>14</b>
<b>C Important Lemmas</b>	<b>15</b>
<b>D Proof of the Connection</b>	<b>17</b>
<b>E Proof in the Strongly-Convex-Strongly-Concave Case</b>	<b>19</b>
E.1 Proof of Stability . . . . .	19
E.2 Proof of the Empirical Risk . . . . .	23
E.3 Proof of Strong/Weak Primal-Dual Population Risk . . . . .	28
<b>F Proof in the Convex-Concave Case</b>	<b>30</b>
F.1 Proof of Stability . . . . .	30
F.2 Proof of the Empirical Risk . . . . .	30
F.3 Proof of Weak Primal-Dual Population Risk . . . . .	30
<b>G Proof in Nonconvex-Nonconcave Case</b>	<b>31</b>
G.1 Important Lemmas . . . . .	31
G.2 Proof of Stability . . . . .	32
<b>H Primal Metric</b>	<b>33</b>
<b>I Additional Experiments</b>	<b>35</b>
I.1 General setup. . . . .	35
I.2 Detailed implementations. . . . .	35

---

## A Decentralized Optimization

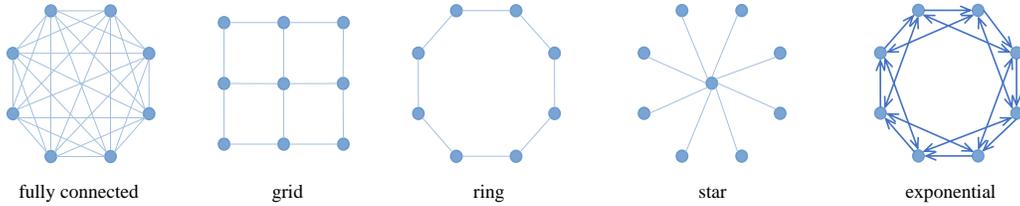


Figure 3: Visualization of different typologies. Note that the lines without arrows mean undirected, such as the fully connected, the grid, the ring, and the star graph. While lines with arrows denote the direction, which shows how information flows from one node to another. Our study objective is undirected graph in this paper, with the exponential graph an illustrative example of directed graph.

## B Notation

1.  $\mathbf{x} \in \mathbb{R}^{d_x}$  and  $\mathbf{y} \in \mathbb{R}^{d_y}$ , we use  $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \in \mathbb{R}^{d_x+d_y}$  to denote the concatenation.

2.  $\mathbf{X}^t \triangleq (\mathbf{x}_1^t, \dots, \mathbf{x}_m^t)^T \in \mathbb{R}^{m \times d_x}$ ;  $\mathbf{Y}^t \triangleq (\mathbf{y}_1^t, \dots, \mathbf{y}_m^t)^T \in \mathbb{R}^{m \times d_y}$ ; and we use  $[\mathbf{X}, \mathbf{Y}]^t \in \mathbb{R}^{m \times (d_x + d_y)}$  to denote the concatenation  $\left[ \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{y}_1^t \end{pmatrix}, \begin{pmatrix} \mathbf{x}_2^t \\ \mathbf{y}_2^t \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{x}_m^t \\ \mathbf{y}_m^t \end{pmatrix} \right]^T$ .
3. We use  $\boldsymbol{\xi}^t$  to denote the samples collected in the  $t$ -th iteration, i.e.,  $\{\xi_{1,j_t(1)}, \dots, \xi_{m,j_t(m)}\}$ .
4. We denote the gradients in the  $t$ -th iteration as follows:

$$\begin{aligned} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) &\triangleq (\nabla_{\mathbf{x}} f_1(\mathbf{x}_1^t, \mathbf{y}_1^t; \xi_{1,j_t(1)}), \dots, \nabla_{\mathbf{x}} f_m(\mathbf{x}_m^t, \mathbf{y}_m^t; \xi_{m,j_t(m)}))^T \in \mathbb{R}^{m \times d_x} \\ \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) &\triangleq (\nabla_{\mathbf{y}} f_1(\mathbf{x}_1^t, \mathbf{y}_1^t; \xi_{1,j_t(1)}), \dots, \nabla_{\mathbf{y}} f_m(\mathbf{x}_m^t, \mathbf{y}_m^t; \xi_{m,j_t(m)}))^T \in \mathbb{R}^{m \times d_y} \\ \begin{pmatrix} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \\ \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \end{pmatrix} &\triangleq \left[ \begin{pmatrix} \nabla_{\mathbf{x}} f_1(\mathbf{x}_1^t, \mathbf{y}_1^t; \xi_{1,j_t(1)}) \\ \nabla_{\mathbf{y}} f_1(\mathbf{x}_1^t, \mathbf{y}_1^t; \xi_{1,j_t(1)}) \end{pmatrix}, \dots, \begin{pmatrix} \nabla_{\mathbf{x}} f_m(\mathbf{x}_m^t, \mathbf{y}_m^t; \xi_{m,j_t(m)}) \\ \nabla_{\mathbf{y}} f_m(\mathbf{x}_m^t, \mathbf{y}_m^t; \xi_{m,j_t(m)}) \end{pmatrix} \right]^T \in \mathbb{R}^{m \times (d_x + d_y)} \end{aligned}$$

5. We use  $\mathbf{P} \in \mathbb{R}^{m \times m}$  to denote the matrix whose elements are all  $\frac{1}{m}$ ,  $\mathbf{I}$  to denote the identity matrix, and  $\mathbf{1}_m \in \mathbb{R}^m$  to denote the vector with all elements equal to 1.

## C Important Lemmas

In this section, we provide some important lemmas as fundamentals of the following proof.

**Lemma 1.** We define  $G_{g,\eta}$  as

$$G_{g,\eta} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} P_{\mathcal{X}}(\mathbf{x} - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y})) \\ P_{\mathcal{Y}}(\mathbf{y} + \eta_{\mathbf{y}} \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})) \end{pmatrix}$$

with  $\eta^{\max} \triangleq \max\{\eta_{\mathbf{x}}, \eta_{\mathbf{y}}\}$  and  $\eta^{\min} \triangleq \min\{\eta_{\mathbf{x}}, \eta_{\mathbf{y}}\}$ . Under the assumption that  $g$  is  $L$ -Lipschitz smooth, then we have:

- a.  $G_{g,\eta}$  is  $(1 + \eta^{\max} L)$ -expansive.
- b. When  $g$  is  $\mu_{\mathbf{x}}$ -strongly-convex w.r.t.  $\mathbf{x}$  and  $\mu_{\mathbf{y}}$ -strongly-concave w.r.t.  $\mathbf{y}$ , letting  $\frac{L+\mu}{2}(\eta^{\max})^2 \leq \eta^{\min} \leq \frac{L+\mu}{2} \frac{1}{L\mu}$  where  $\mu \triangleq \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ , then  $G_{g,\eta}$  is  $(1 - \eta^{\min} \frac{L\mu}{L+\mu})$ -expansive.

*Proof.* For case **a.** where we do not require strong convexity or strong concavity, we have:

$$\begin{aligned} \left\| G_{g,\eta} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} - G_{g,\eta} \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} \right\| &\leq \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' - \eta_{\mathbf{x}} (\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}')) \\ \mathbf{y} - \mathbf{y}' - \eta_{\mathbf{y}} (\nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})) \end{pmatrix} \right\| \\ &= \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\| + \left\| \begin{pmatrix} \eta_{\mathbf{x}} (\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}')) \\ \eta_{\mathbf{y}} (\nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})) \end{pmatrix} \right\| \\ &\leq (1 + \eta^{\max} L) \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\| \end{aligned}$$

When function  $g$  is further  $\mu_{\mathbf{x}}$ -strongly convex and  $\mu_{\mathbf{y}}$ -strongly concave in case **b.**, thus we have:

$$\begin{aligned} &\left\| G_{g,\eta} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} - G_{g,\eta} \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} \right\|^2 \\ &= \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' - \eta_{\mathbf{x}} (\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}')) \\ \mathbf{y} - \mathbf{y}' - \eta_{\mathbf{y}} (\nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})) \end{pmatrix} \right\|^2 \\ &= \|\mathbf{x} - \mathbf{x}' - \eta_{\mathbf{x}} (\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}'))\|^2 + \|\mathbf{y} - \mathbf{y}' - \eta_{\mathbf{y}} (\nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}))\|^2 \\ &= \|\mathbf{x} - \mathbf{x}'\|^2 + \eta_{\mathbf{x}}^2 \|\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}'))\|^2 - 2\eta_{\mathbf{x}} \langle \mathbf{x} - \mathbf{x}', \nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g(\mathbf{x}', \mathbf{y}') \rangle \\ &\quad + \|\mathbf{y} - \mathbf{y}'\|^2 + \eta_{\mathbf{y}}^2 \|\nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})\|^2 - 2\eta_{\mathbf{y}} \langle \mathbf{y} - \mathbf{y}', \nabla_{\mathbf{y}} g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y}) \rangle \end{aligned} \tag{4}$$

Recalling the co-coercivity of a  $L$ -Lipschitz smooth and  $\mu$ -strongly convex function  $f(\mathbf{x})$  that [13]:

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L + \mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 + \frac{L\mu}{L + \mu} \|\mathbf{x} - \mathbf{y}\|^2$$

So we can get:

$$\begin{aligned}
& \langle \mathbf{x} - \mathbf{x}', \nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}}g(\mathbf{x}', \mathbf{y}') \rangle + \langle \mathbf{y} - \mathbf{y}', \nabla_{\mathbf{y}}g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}}g(\mathbf{x}, \mathbf{y}) \rangle \\
&= \left\langle \begin{pmatrix} \nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}}g(\mathbf{x}', \mathbf{y}') \\ \nabla_{\mathbf{y}}g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}}g(\mathbf{x}, \mathbf{y}) \end{pmatrix}, \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\rangle \\
&\geq \frac{1}{L + \mu} \left\| \begin{pmatrix} \nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}}g(\mathbf{x}', \mathbf{y}') \\ \nabla_{\mathbf{y}}g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}}g(\mathbf{x}, \mathbf{y}) \end{pmatrix} \right\|^2 + \frac{L\mu}{L + \mu} \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\|^2 \\
&= \frac{1}{L + \mu} \|\nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}}g(\mathbf{x}', \mathbf{y}')\|^2 + \frac{1}{L + \mu} \|\nabla_{\mathbf{y}}g(\mathbf{x}', \mathbf{y}') - \nabla_{\mathbf{y}}g(\mathbf{x}, \mathbf{y})\|^2 \\
&\quad + \frac{L\mu}{L + \mu} \|\mathbf{x} - \mathbf{x}'\|^2 + \frac{L\mu}{L + \mu} \|\mathbf{y} - \mathbf{y}'\|^2
\end{aligned} \tag{5}$$

Combining above inequalities (4) and (5), we can get:

$$\left\| G_{g,\eta} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} - G_{g,\eta} \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} \right\| \leq \left(1 - \eta^{\min} \frac{L\mu}{L + \mu}\right) \left\| \begin{pmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{pmatrix} \right\|$$

when  $\frac{L+\mu}{2}(\eta^{\max})^2 \leq \eta^{\min} \leq \frac{L+\mu}{2} \frac{1}{L\mu}$  satisfies.  $\square$

**Lemma 2.** Letting  $\mathbf{P} \in \mathbb{R}^{m \times m}$  denote the matrix whose elements are all  $\frac{1}{m}$ , following the update rule of D-SGDA (see Algorithm 1), we have:

$$\left\| (\mathbb{I} - \mathbf{P}) [\mathbf{X}, \mathbf{Y}]^t \right\| \leq 2\sqrt{m}G \sum_{s=0}^{t-1} \eta_s^{\max} \lambda^{t-1-s}$$

where  $\eta_s^{\max} = \max\{\eta_{\mathbf{x},s}, \eta_{\mathbf{y},s}\}$ .

*Proof.* The Lipschitz continuity (see Assumption 1) implies that for any  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{y} \in \mathcal{Y}$  on a given sample  $\xi$ , the gradient value is bounded by  $G$ , i.e.,  $\left\| \begin{pmatrix} \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \xi^t) \\ -\eta_{\mathbf{y},t} \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \xi^t) \end{pmatrix} \right\| \leq \max\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\} \sqrt{m}G$ . Following the proof of Lemma 8 in [32], we can get the result, which can be expressed as  $[\sum_{i=1}^m \|\mathbf{x}_i^t - \mathbf{x}^t\|^2 + \|\mathbf{y}^t - \mathbf{y}^t\|^2]^{1/2} \leq 2\sqrt{m}G \sum_{s=0}^{t-1} \eta_s^{\max} \lambda^{t-1-s}$   $\square$

**Lemma 3.** When  $0 < \lambda < 1$ ,

$$\sum_{j=0}^{t-1} \frac{\lambda^{t-1-j}}{(j+1)^k} \leq \frac{C_\lambda}{t^k}, k > 0; \quad C_\lambda \triangleq \frac{(k/e)^k}{\lambda(\ln \frac{1}{\lambda})^k} + \frac{2e^{-1}}{\lambda \ln \frac{1}{\lambda}} + \frac{2^k}{\lambda \ln \frac{1}{\lambda}}$$

*Proof.*

$$\sum_{j=0}^{t-1} \frac{\lambda^{t-1-j}}{(j+1)^k} = \lambda^{t-1} + \sum_{j=1}^{t-1} \frac{\lambda^{t-1-j}}{(j+1)^k} \leq \lambda^{t-1} + \int_1^t \frac{\lambda^{t-1-x}}{x^k} dx = \lambda^{t-1} + \lambda^{t-1} \int_1^t \frac{\lambda^{-x}}{x^k} dx$$

whereas for the integral we have:

$$\begin{aligned}
\int_1^t \frac{\lambda^{-x}}{x^k} dx &= \int_1^{\frac{t}{2}} \frac{\lambda^{-x}}{x^k} dx + \int_{\frac{t}{2}}^t \frac{\lambda^{-x}}{x^k} dx \leq \lambda^{-\frac{t}{2}} \int_1^{\frac{t}{2}} \frac{1}{x^k} dx + \frac{1}{(\frac{t}{2})^k} \int_{\frac{t}{2}}^t \lambda^{-x} dx \\
&= \int_1^{\frac{t}{2}} \frac{1}{x^k} dx \leq \begin{cases} \ln \frac{t}{2} & k = 1 \\ \frac{(\frac{t}{2})^{1-k}}{1-k} & 0 < k < 1 \end{cases}
\end{aligned}$$

Therefore we have:

$$\sum_{j=0}^{t-1} \frac{\lambda^{t-1-j}}{(j+1)^k} \leq \begin{cases} \lambda^{t-1} + \lambda^{\frac{t}{2}-1} \ln \frac{t}{2} + \frac{1}{t^k} \frac{2^k}{\lambda \ln \frac{1}{\lambda}} & k = 1 \\ \lambda^{t-1} + \lambda^{\frac{t}{2}-1} t^{1-k} + \frac{1}{t^k} \frac{2^k}{\lambda \ln \frac{1}{\lambda}} & 0 < k < 1 \end{cases}$$

When  $0 < k < 1$ ,  $\sum_{j=0}^{t-1} \frac{\lambda^{t-1-j}}{(j+1)^k} \leq \frac{1}{t^k} \left( \lambda^{t-1} t^k + \lambda^{\frac{k}{2}-1} t + \frac{2^k}{\lambda \ln \frac{1}{\lambda}} \right)$ , where  $\lambda^{t-1} t^k \leq \frac{(k/e)^k}{\lambda (\ln \frac{1}{\lambda})^k}$ ,  $\lambda^{\frac{k}{2}-1} t \leq \frac{2e^{-1}}{\lambda \ln \frac{1}{\lambda}}$ . So we can define  $C_\lambda = \frac{(k/e)^k}{\lambda (\ln \frac{1}{\lambda})^k} + \frac{2e^{-1}}{\lambda \ln \frac{1}{\lambda}} + \frac{2^k}{\lambda \ln \frac{1}{\lambda}}$  that  $\sum_{j=0}^{t-1} \frac{\lambda^{t-1-j}}{(j+1)^k} \leq \frac{C_\lambda}{t^k}$ . And for the case  $k = 1$  we can roughly consider the logarithm as constants so our discussion can be put together into the above category as  $k = 1$ .  $\square$

**Remark 10.** Considering function  $h(\lambda) = \frac{1}{\lambda \ln \frac{1}{\lambda}}$ , it monotonically decrease in interval  $(0, \frac{1}{e})$  and monotonically increase in interval  $(\frac{1}{e}, 1)$  with minimal value of  $e$ . Analogously for the function  $g(\lambda) = \frac{1}{\lambda (\ln \frac{1}{\lambda})^k}$ , which monotonically decrease in interval  $(0, \frac{1}{e^k})$  and monotonically increase in interval  $(\frac{1}{e^k}, 1)$  with minimal value of  $(\frac{e}{k})^k$ . Therefore, when  $k = 1$ ,  $C_\lambda$  is bounded by  $\mathcal{O}(\frac{1}{\lambda \ln \frac{1}{\lambda}})$ ; otherwise, when  $\lambda \rightarrow 1$ ,  $C_\lambda$  is bounded by  $\mathcal{O}(\frac{1}{\lambda \ln \frac{1}{\lambda}})$ ; when  $\lambda \rightarrow 0$ ,  $C_\lambda$  is bounded by  $\mathcal{O}(\frac{1}{\lambda (\ln \frac{1}{\lambda})^k})$ .

## D Proof of the Connection

In this section, we will first figure out a fundamental lemma that illustrates the structure of dataset in decentralized setting. Then we will provide proof for the connection between argument stability and primal-dual generalization gap.

**Lemma 4.** Denoting dataset as  $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ , and denoting each local samples stored in  $\mathcal{S}_i$  as  $\mathcal{S}_i = \{\xi_{i,l_i}\}_{l_i=1, \dots, n}$ . Then we have the decomposition equation of the empirical function:

$$F_{\mathcal{S}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n^m} \sum_{l_1=1}^n \dots \sum_{l_m=1}^n \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}, \mathbf{y}; \xi_{i,l_i})$$

*Proof.*

$$\begin{aligned} & \frac{1}{n^m} \sum_{l_1=1}^n \dots \sum_{l_m=1}^n \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}, \mathbf{y}; \xi_{i,l_i}) \\ &= \frac{1}{m} \sum_{i=1}^m \left( \frac{1}{n^{m-1}} \sum_{l_1=1}^n \dots \sum_{l_{i-1}=1}^n \sum_{l_{i+1}=1}^n \dots \sum_{l_m=1}^n \frac{1}{n} \sum_{l_i=1}^n f_i(\mathbf{x}, \mathbf{y}; \xi_{i,l_i}) \right) \\ &= \frac{1}{m} \sum_{i=1}^m \left( \frac{1}{n^{m-1}} \sum_{l_1=1}^n \dots \sum_{l_{i-1}=1}^n \sum_{l_{i+1}=1}^n \dots \sum_{l_m=1}^n F_{\mathcal{S}_i}(\mathbf{x}, \mathbf{y}) \right) \\ &= \frac{1}{m} \sum_{i=1}^m F_{\mathcal{S}_i}(\mathbf{x}, \mathbf{y}) \\ &= F_{\mathcal{S}}(\mathbf{x}, \mathbf{y}) \end{aligned}$$

$\square$

*Proof of Theorem 1.* We let  $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$  where  $\mathcal{S}_i = \{\xi_{i,l}\}_{1 \leq l \leq n}$  and  $\mathcal{S}' = \{\mathcal{S}'_1, \dots, \mathcal{S}'_m\}$  where  $\mathcal{S}'_i = \{\xi'_{i,l}\}_{1 \leq l \leq n}$  be two different datasets while  $\mathcal{S}_i$  and  $\mathcal{S}'_i$  are drawn from the same distribution  $\mathcal{D}_i$ . Further we define  $\mathcal{S}^{(l)} = \{\mathcal{S}_1^{(l_1)}, \mathcal{S}_2^{(l_2)}, \dots, \mathcal{S}_m^{(l_m)}\}$  where  $\mathbf{l} = (l_1, l_2, \dots, l_m)$  and  $\mathcal{S}_i^{(l_i)} = \{\xi_{i,1}, \dots, \xi_{i,l_i-1}, \xi'_{i,l_i}, \xi_{i,l_i+1}, \dots, \xi_{i,n}\}$ . So we can say that there is at most one different sample in each local dataset between  $\mathcal{S}$  and  $\mathcal{S}^{(l)}$ . For function  $F : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ , we denote  $\mathbf{y}^*(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y})$ .

For case **a**, the weak primal-dual generalization gap  $\epsilon_{gen}^w(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathcal{A}_{\mathbf{y}}(\mathcal{S})) = \Delta^w(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathcal{A}_{\mathbf{y}}(\mathcal{S})) - \Delta_{\mathcal{S}}^w(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathcal{A}_{\mathbf{y}}(\mathcal{S}))$ . We first make some adjustments to the definition:

$$\begin{aligned}
& \epsilon_{gen}^w(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \\
&= (\sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}[F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] - \inf_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}[F(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))]) - (\sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}[F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] - \inf_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}[F_S(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))]) \\
&= \sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}[F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] - \sup_{\mathbf{y}' \in \mathcal{Y}} \mathbb{E}[F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] + \inf_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}[F_S(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))] - \inf_{\mathbf{x}' \in \mathcal{X}} \mathbb{E}[F(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))] \\
&\leq \sup_{\mathbf{y}' \in \mathcal{Y}} (\mathbb{E}_{\mathcal{A}, \mathcal{S}}[F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}') - F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] + \sup_{\mathbf{x}' \in \mathcal{X}} (\mathbb{E}_{\mathcal{A}, \mathcal{S}}[F_S(\mathbf{x}', \mathcal{A}_y(\mathcal{S})) - F(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))])
\end{aligned}$$

where the expectation can be detailed into  $\mathbb{E}_{\mathcal{A}, \mathcal{S}}$  in this situation.

For the first term, we have

$$\begin{aligned}
& \mathbb{E}_{\mathcal{A}, \mathcal{S}}[F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}') - F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] \\
&= \frac{1}{n^m} \sum_{l_1=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[F(\mathcal{A}_x(\mathcal{S}^{(l)}), \mathbf{y}')] - \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] \\
&= \frac{1}{n^m} \sum_{l_1=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[\frac{1}{m} \sum_{i=1}^m f_i(\mathcal{A}_x(\mathcal{S}^{(l)}), \mathbf{y}'; \xi_{i, l_i}) - \frac{1}{m} \sum_{i=1}^m f_i(\mathcal{A}_x(\mathcal{S}), \mathbf{y}'; \xi_{i, l_i})] \\
&\leq \frac{1}{n^m} \sum_{l_1=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[\frac{1}{m} \sum_{i=1}^m G \|\mathcal{A}_x(\mathcal{S}^{(l)}) - \mathcal{A}_x(\mathcal{S})\|] \\
&= G \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'} \|\mathcal{A}_x(\mathcal{S}^{(l)}) - \mathcal{A}_x(\mathcal{S})\|
\end{aligned}$$

where the first equation is due to the symmetric distribution between  $\mathcal{S}_i$  and  $\mathcal{S}'_i$  and there are overall  $n^m$  permutations of  $l$ . And the second equation is due to the independence of sample  $\xi_{i, l_i}$  from the training process upon dataset  $\mathcal{S}^{(l)}$  and Lemma 4.

And we can analyze the second term in a similar way:

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}}[F_S(\mathbf{x}', \mathcal{A}_y(\mathcal{S})) - F(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))] \leq G \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'} \|\mathcal{A}_y(\mathcal{S}) - \mathcal{A}_y(\mathcal{S}^{(l)})\|$$

Therefore we can get the weak primal-dual generalization gap that:

$$\epsilon_{gen}^w(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \leq \sqrt{2}G \sup_{\mathcal{S}, \mathcal{S}'} \mathbb{E}_{\mathcal{A}} \left\| \begin{array}{c} \mathcal{A}_x(\mathcal{S}) - \mathcal{A}_x(\mathcal{S}^{(l)}) \\ \mathcal{A}_y(\mathcal{S}) - \mathcal{A}_y(\mathcal{S}^{(l)}) \end{array} \right\| \leq \sqrt{2}G\epsilon$$

For case **b**, the strong primal-dual generalization gap  $\epsilon_{gen}^s(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) = \Delta^s(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) - \Delta_{\mathcal{S}}^s(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S}))$ .

Observing the structure that:

$$\begin{aligned}
& \epsilon_{gen}^s(\mathcal{A}_x(\mathcal{S}), \mathcal{A}_y(\mathcal{S})) \\
&= \mathbb{E}[\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))] - \mathbb{E}[\sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))] \\
&= \mathbb{E}_{\mathcal{A}, \mathcal{S}}[\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}') - \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] + \mathbb{E}_{\mathcal{A}, \mathcal{S}}[\inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathcal{A}_y(\mathcal{S})) - \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathcal{A}_y(\mathcal{S}))]
\end{aligned} \tag{6}$$

where the expectation can be detailed into  $\mathbb{E}_{\mathcal{S}, \mathcal{A}}$  in this situation.

For the first term,

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}, \mathcal{S}}[\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathcal{A}_x(\mathcal{S}), \mathbf{y}')] &\stackrel{(i)}{=} \frac{1}{n^m} \sum_{l_1=1}^n \sum_{l_2=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathcal{A}_x(\mathcal{S}^{(l)}), \mathbf{y}')] \\
&\stackrel{(ii)}{=} \frac{1}{n^m} \sum_{l_1=1}^n \sum_{l_2=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[F(\mathcal{A}_x(\mathcal{S}^{(l)}), \mathbf{y}_{\mathcal{S}^{(l)}}^*)] \\
&\stackrel{(iii)}{=} \frac{1}{n^m} \sum_{l_1=1}^n \sum_{l_2=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'}[\frac{1}{m} \sum_{i=1}^m f_i(\mathcal{A}_x(\mathcal{S}^{(l)}), \mathbf{y}_{\mathcal{S}^{(l)}}^*; \xi_{i, l_i})]
\end{aligned}$$

Since  $\mathcal{S}_i$  and  $\mathcal{S}'_i$  are drawn from the same distribution, and there are  $n^m$  permutations of  $l = \{l_1, \dots, l_m\}$  so we can get equality (i). In equality (ii), we denote  $\mathbf{y}_{\mathcal{S}^{(l)}}^*$  as  $arg \max_{\mathbf{y} \in \mathcal{Y}} F(\mathcal{A}_x(\mathcal{S}), \mathbf{y})$ .

While for equality (iii),  $\xi_{i,l_i}$  is independent from the training process under dataset  $\mathcal{S}^l$  for each local agent respectively.

Using the property of Lipschitz continuous (see Assumption 1), we can further get:

$$\begin{aligned} f_i(\mathcal{A}_{\mathbf{x}}(\mathcal{S}^l), \mathbf{y}_{\mathcal{S}^l}^*; \xi_{i,l_i}) - f_i(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathbf{y}_{\mathcal{S}}^*; \xi_{i,l_i}) &\leq G \left\| \begin{pmatrix} \mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{x}}(\mathcal{S}) \\ \mathbf{y}_{\mathcal{S}^l}^* - \mathbf{y}_{\mathcal{S}}^* \end{pmatrix} \right\| \\ &\stackrel{(a)}{\leq} G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{y}}^2}} \|\mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{x}}(\mathcal{S})\| \end{aligned}$$

where inequality (a) is a conclusion of Lemma 4.3 in [22] that  $\|\mathbf{y}_{\mathcal{S}^l}^* - \mathbf{y}_{\mathcal{S}}^*\| \leq \frac{L}{\mu_{\mathbf{y}}} \|\mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{x}}(\mathcal{S})\|$ .

Combining above two inequalities, we can get:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}, \mathcal{S}} [\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathbf{y}')] \\ &\leq \frac{1}{n^m} \sum_{l_1=1}^n \sum_{l_2=1}^n \dots \sum_{l_m=1}^n \mathbb{E}_{\mathcal{A}, \mathcal{S}, \mathcal{S}'} \left[ \frac{1}{m} \sum_{i=1}^m \left( f_i(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathbf{y}_{\mathcal{S}}^*; \xi_{i,l_i}) + G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{y}}^2}} \|\mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{x}}(\mathcal{S})\| \right) \right] \\ &= \mathbb{E}_{\mathcal{A}, \mathcal{S}} [F_{\mathcal{S}}(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathbf{y}_{\mathcal{S}}^*)] + G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{y}}^2}} \mathbb{E}_{\mathcal{A}} [\sup_{\mathcal{S}, \mathcal{S}'} \|\mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{x}}(\mathcal{S})\|] \\ &\leq \mathbb{E}_{\mathcal{A}, \mathcal{S}} \left[ \sup_{\mathbf{y}' \in \mathcal{Y}} F_{\mathcal{S}}(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathbf{y}') \right] + G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{y}}^2}} \mathbb{E}_{\mathcal{A}} [\sup_{\mathcal{S}, \mathcal{S}'} \|\mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{x}}(\mathcal{S})\|] \end{aligned}$$

where we get the last but one inequality from Lemma 4.

We can do a similar operation on the second counterpart since  $f_i$  is also  $\mu_{\mathbf{x}}$ -strongly convex on  $\mathbf{x}$ .

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}} [\inf_{\mathbf{x}' \in \mathcal{X}} F_{\mathcal{S}}(\mathbf{x}', \mathcal{A}_{\mathbf{y}}(\mathcal{S})) - \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathcal{A}_{\mathbf{y}}(\mathcal{S}))] \leq G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{x}}^2}} \mathbb{E}_{\mathcal{A}} [\sup_{\mathcal{S}, \mathcal{S}'} \|\mathcal{A}_{\mathbf{y}}(\mathcal{S}^l) - \mathcal{A}_{\mathbf{y}}(\mathcal{S})\|]$$

So the overall strong primal-dual generalization gap satisfies:

$$\begin{aligned} &\epsilon_{gen}^s(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathcal{A}_{\mathbf{y}}(\mathcal{S})) \\ &\leq G \sqrt{1 + \frac{L^2}{\mu^2}} \sqrt{2} \mathbb{E}_{\mathcal{A}} \left[ \sup_{\mathcal{S}, \mathcal{S}'} \left\| \begin{pmatrix} \mathcal{A}_{\mathbf{x}}(\mathcal{S}) - \mathcal{A}_{\mathbf{x}}(\mathcal{S}^l) \\ \mathcal{A}_{\mathbf{y}}(\mathcal{S}) - \mathcal{A}_{\mathbf{y}}(\mathcal{S}^l) \end{pmatrix} \right\| \right] \\ &\leq G \sqrt{2 + \frac{2L^2}{\mu^2}} \epsilon \end{aligned}$$

where  $\mu = \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ . □

## E Proof in the Strongly-Convex-Strongly-Concave Case

In this section, we will prove the stability and generalization gap of our D-SGDA in the case of  $\mu_{\mathbf{x}}$ -strongly convex and  $\mu_{\mathbf{y}}$ -strongly concave.

### E.1 Proof of Stability

*Proof of Theorem 2.* We use  $(\mathbf{x}^t, \mathbf{y}^t)$  and  $(\hat{\mathbf{x}}^t, \hat{\mathbf{y}}^t)$  to represent the output in the  $t$ -th iteration when applying D-SGDA on any arbitrary neighbouring dataset  $\mathcal{S}$  and  $\mathcal{S}'$  respectively. Since the D-SGDA algorithm is symmetric [4] with respect to the dataset  $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ , we can assume the different samples appear in the last location in each  $\mathcal{S}_i$  without loss of generalization, i.e.,  $\mathcal{S}' = \{\mathcal{S}'_1, \dots, \mathcal{S}'_{m-1}, \mathcal{S}'_m\}$  where  $\mathcal{S}'_i = \{\xi_{i,1}, \xi_{i,2}, \dots, \xi_{i,n-1}, \xi'_{i,n}\}$  differs from  $\mathcal{S}_i = \{\xi_{i,1}, \dots, \xi_{i,n}\}$  in the  $n$ -th data.

First concentrating on the iteration, we have:

$$\begin{aligned}
\begin{pmatrix} \mathbf{x}^{t+1} \\ \mathbf{y}^{t+1} \end{pmatrix} &= ([\mathbf{X}, \mathbf{Y}]^{t+1})^T \frac{\mathbb{1}_m}{m} \\
&= P_{(\mathcal{X}, \mathcal{Y})^m} \left[ \left( \mathbf{W}[\mathbf{X}, \mathbf{Y}]^t - \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \end{pmatrix} \right)^T \right] \frac{\mathbb{1}_m}{m} \\
&= P_{(\mathcal{X}, \mathcal{Y})} \left[ \left( \mathbf{W}[\mathbf{X}, \mathbf{Y}]^t - \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \end{pmatrix} \right)^T \frac{\mathbb{1}_m}{m} \right] \\
&= P_{(\mathcal{X}, \mathcal{Y})} \left[ ([\mathbf{X}, \mathbf{Y}]^t)^T \mathbf{W}^T \frac{\mathbb{1}_m}{m} - \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \end{pmatrix}^T \frac{\mathbb{1}_m}{m} \right] \\
&= P_{(\mathcal{X}, \mathcal{Y})} \left[ \begin{pmatrix} \mathbf{x}^t \\ \mathbf{y}^t \end{pmatrix} - \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t) \end{pmatrix}^T \frac{\mathbb{1}_m}{m} \right] \\
&= P_{(\mathcal{X}, \mathcal{Y})} \left[ \begin{pmatrix} \mathbf{x}^t \\ \mathbf{y}^t \end{pmatrix} - \frac{1}{m} \sum_{i=1}^m \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) \end{pmatrix} \right]
\end{aligned}$$

where in the first equality we use the definition that  $\mathbf{x}^t = \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i^t$ ,  $\mathbf{y}^t = \frac{1}{m} \sum_{i=1}^m \mathbf{y}_i^t$ , and the last equality is due to the fact that  $\mathbf{W} \frac{\mathbb{1}_m}{m} = \frac{\mathbb{1}_m}{m}$ . We use the notation  $P_{(\mathcal{X}, \mathcal{Y})} \left[ \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \right] \in \mathbb{R}^{d_{\mathbf{x}} + d_{\mathbf{y}}}$  as an abbreviation of  $\begin{pmatrix} P_{\mathcal{X}}[\mathbf{x}] \\ P_{\mathcal{Y}}[\mathbf{y}] \end{pmatrix} \in \mathbb{R}^{d_{\mathbf{x}} + d_{\mathbf{y}}}$ , and  $P_{(\mathcal{X}, \mathcal{Y})^m} [[\mathbf{X}, \mathbf{Y}]^T]$  means every column should obey  $P_{(\mathcal{X}, \mathcal{Y})} \left[ \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \right]$ .

In the  $t$ -th iteration, there is a probability of  $C_m^{m_0} (1 - \frac{1}{n})^{m-m_0} (\frac{1}{n})^{m_0}$  that there are  $m - m_0$  agents not selecting the last different samples while the rest  $m_0$  agents selecting exactly the last different samples. Without loss of generalization, we assume the first  $m - m_0$  not selecting and the rest  $m_0$  agents selecting the different samples respectively. So we can get:

$$\begin{aligned}
&\begin{pmatrix} \mathbf{x}^{t+1} - \hat{\mathbf{x}}^{t+1} \\ \mathbf{y}^{t+1} - \hat{\mathbf{y}}^{t+1} \end{pmatrix} \\
&= P_{\mathcal{X}, \mathcal{Y}} \left[ \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} - \frac{1}{m} \sum_{i=1}^m \begin{pmatrix} \eta_{\mathbf{x}, t} (\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) - \nabla_{\mathbf{x}} f_i(\hat{\mathbf{x}}_i^t, \hat{\mathbf{y}}_i^t; \xi_{i, j_t(i)})) \\ -\eta_{\mathbf{y}, t} (\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) - \nabla_{\mathbf{y}} f_i(\hat{\mathbf{x}}_i^t, \hat{\mathbf{y}}_i^t; \xi_{i, j_t(i)})) \end{pmatrix} \right] \\
&\stackrel{(i)}{=} P_{\mathcal{X}, \mathcal{Y}} \left[ \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} - \frac{1}{m} \left[ \sum_{i=1}^{m-m_0} \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) - \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} f_i(\hat{\mathbf{x}}_i^t, \hat{\mathbf{y}}_i^t; \xi_{i, j_t(i)}) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, j_t(i)}) + \eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} f_i(\hat{\mathbf{x}}_i^t, \hat{\mathbf{y}}_i^t; \xi_{i, j_t(i)}) \end{pmatrix} \right. \right. \\
&\quad \left. \left. + \sum_{i=m-m_0+1}^m \begin{pmatrix} \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, n}) - \eta_{\mathbf{x}, t} \nabla_{\mathbf{x}} f_i(\hat{\mathbf{x}}_i^t, \hat{\mathbf{y}}_i^t; \xi'_{i, n}) \\ -\eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i, n}) + \eta_{\mathbf{y}, t} \nabla_{\mathbf{y}} f_i(\hat{\mathbf{x}}_i^t, \hat{\mathbf{y}}_i^t; \xi'_{i, n}) \end{pmatrix} \right] \right]
\end{aligned}$$

where equality (i) is due to the setting that agents from 1 to  $m - m_0$  select  $\xi_{i, j_t(i)}$  which is not the last different one while the rest agents from  $m - m_0 + 1$  to  $m$  select the last different samples, i.e.,  $\xi_{i, n}$  and  $\xi'_{i, n}$  respectively.

Then we can decompose the term inside the projection into several parts as following:

$$\begin{aligned}
& \underbrace{\frac{1}{m} \sum_{i=1}^{m-m_0} \left( \begin{array}{c} \mathbf{x}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) - (\dot{\mathbf{x}}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,j_t(i)})) \\ \mathbf{y}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) - (\dot{\mathbf{y}}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,j_t(i)})) \end{array} \right)}_{I_1} \\
& + \underbrace{\frac{1}{m} \sum_{i=m-m_0+1}^m \left( \begin{array}{c} \mathbf{x}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) - (\dot{\mathbf{x}}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n})) \\ \mathbf{y}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) - (\dot{\mathbf{y}}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n})) \end{array} \right)}_{I_2} \\
& + \underbrace{\frac{1}{m} \sum_{i=1}^{m-m_0} \left( \begin{array}{c} -\eta_{\mathbf{x},t} (\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)})) \\ \eta_{\mathbf{y},t} (\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)})) \end{array} \right)}_{II_1} \\
& - \underbrace{\frac{1}{m} \sum_{i=1}^{m-m_0} \left( \begin{array}{c} -\eta_{\mathbf{x},t} (\nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}_i^t, \dot{\mathbf{y}}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,j_t(i)})) \\ \eta_{\mathbf{y},t} (\nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}_i^t, \dot{\mathbf{y}}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,j_t(i)})) \end{array} \right)}_{II_2} \\
& + \underbrace{\frac{1}{m} \sum_{i=m-m_0+1}^m \left( \begin{array}{c} -\eta_{\mathbf{x},t} (\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,n}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n})) \\ \eta_{\mathbf{y},t} (\nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,n}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n})) \end{array} \right)}_{II'_1} \\
& - \underbrace{\frac{1}{m} \sum_{i=m-m_0+1}^m \left( \begin{array}{c} -\eta_{\mathbf{x},t} (\nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}_i^t, \dot{\mathbf{y}}_i^t; \xi'_{i,n}) - \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n})) \\ \eta_{\mathbf{y},t} (\nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}_i^t, \dot{\mathbf{y}}_i^t; \xi'_{i,n}) - \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n})) \end{array} \right)}_{II'_2}
\end{aligned} \tag{7}$$

According to Lemma 1 and the property of Lipschitz continuity (see Assumption 1), we can bound the term  $I_1$  as:

$$\|I_1\| \leq \frac{m-m_0}{m} (1 - \eta_t^{\min} \frac{L\mu}{L+\mu}) \left\| \begin{pmatrix} \mathbf{x}^t - \dot{\mathbf{x}}^t \\ \mathbf{y}^t - \dot{\mathbf{y}}^t \end{pmatrix} \right\|$$

where  $\eta_t^{\min} \triangleq \min\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ ,  $\eta_t^{\max} \triangleq \max\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ , and  $\mu \triangleq \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ . For the term  $I_2$ , it is bounded by:

$$\begin{aligned}
\|I_2\| & \leq \frac{1}{m} \sum_{i=m-m_0+1}^m \left\| \begin{pmatrix} \mathbf{x}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) - (\dot{\mathbf{x}}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n})) \\ \mathbf{y}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) - (\dot{\mathbf{y}}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n})) \end{pmatrix} \right\| \\
& \leq \frac{1}{m} \sum_{i=m-m_0+1}^m \left[ \left\| \begin{pmatrix} \mathbf{x}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) - (\dot{\mathbf{x}}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,n})) \\ \mathbf{y}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) - (\dot{\mathbf{y}}^t + \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,n})) \end{pmatrix} \right\| \right. \\
& \quad \left. + \left\| \begin{pmatrix} \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n}) - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,n}) \\ \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi_{i,n}) - \eta_{\mathbf{y},t} \nabla_{\mathbf{y}} f_i(\dot{\mathbf{x}}^t, \dot{\mathbf{y}}^t; \xi'_{i,n}) \end{pmatrix} \right\| \right] \\
& \leq \frac{m_0}{m} \left[ \left(1 - \eta_t^{\min} \frac{L\mu}{L+\mu}\right) \left\| \begin{pmatrix} \mathbf{x}^t - \dot{\mathbf{x}}^t \\ \mathbf{y}^t - \dot{\mathbf{y}}^t \end{pmatrix} \right\| + 2\eta_t^{\max} G \right]
\end{aligned}$$

For the term  $II_1$  and  $II'_1$ ,

$$\begin{aligned}
\|II_1 + II'_1\| &\leq \frac{1}{m} \sum_{i=1}^{m-m_0} \left[ \eta_t^{max} \left\| \begin{pmatrix} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) \\ \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \end{pmatrix} \right\| \right] \\
&+ \frac{1}{m} \sum_{i=m-m_0+1}^m \left[ \eta_t^{max} \left\| \begin{pmatrix} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,n}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,n}) \\ \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,n}) - \nabla_{\mathbf{y}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,n}) \end{pmatrix} \right\| \right] \\
&\leq \frac{1}{m} \left[ \sum_{i=1}^m \eta_t^{max} L \left\| \begin{pmatrix} \mathbf{x}_i^t - \mathbf{x}^t \\ \mathbf{y}_i^t - \mathbf{y}^t \end{pmatrix} \right\| \right] \\
&\leq \frac{\eta_t^{max} L}{m} \sum_{i=1}^m \left\| \begin{pmatrix} \mathbf{x}_i^t - \mathbf{x}^t \\ \mathbf{y}_i^t - \mathbf{y}^t \end{pmatrix} \right\| \\
&\leq \frac{\eta_t^{max} L}{\sqrt{m}} \left[ \sum_{i=1}^m \|\mathbf{x}_i^t - \mathbf{x}^t\|^2 + \|\mathbf{y}_i^t - \mathbf{y}^t\|^2 \right]^{1/2} \\
&\leq 2\eta_t^{max} LG \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s}
\end{aligned}$$

And it is the same as the term  $II_2 + II'_2$  that  $\|II_2 + II'_2\| \leq 2\eta_t^{max} LG \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s}$ . Taking expectation on the choice of each local sample, we can get:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^{t+1} - \hat{\mathbf{x}}^{t+1} \\ \mathbf{y}^{t+1} - \hat{\mathbf{y}}^{t+1} \end{pmatrix} \right\| \\
&\leq \sum_{m_0=0}^m C_m^{m_0} \left(1 - \frac{1}{n}\right)^{m-m_0} \left(\frac{1}{n}\right)^{m_0} \left( \left(1 - \eta_t^{min} \frac{L\mu}{L+\mu}\right) \mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| + \frac{m_0}{m} 2\eta_t^{max} G \right. \\
&\quad \left. + 4\eta_t^{max} LG \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \right) \\
&\leq \left(1 - \eta_t^{min} \frac{L\mu}{L+\mu}\right) \mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| + 4\eta_t^{max} LG \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \\
&\quad + \sum_{m_0=0}^m C_m^{m_0} \left(1 - \frac{1}{n}\right)^{m-m_0} \left(\frac{1}{n}\right)^{m_0} \frac{m_0}{m} 2\eta_t^{max} G \\
&= \left(1 - \eta_t^{min} \frac{L\mu}{L+\mu}\right) \mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| + 4\eta_t^{max} LG \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} + \frac{2\eta_t^{max} G}{n}
\end{aligned} \tag{8}$$

where the last inequality is due to the property of binomial coefficient that  $C_m^{m_0} \cdot \frac{m_0}{m} = C_{m-1}^{m_0-1}$ . Recursively applying the above inequality, we can get:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathcal{A}_{\mathbf{x}}(\mathcal{S}) - \mathcal{A}_{\mathbf{x}}(\mathcal{S}') \\ \mathcal{A}_{\mathbf{y}}(\mathcal{S}) - \mathcal{A}_{\mathbf{y}}(\mathcal{S}') \end{pmatrix} \right\| \\
&\leq \frac{2G}{n} \sum_{k=0}^{T-1} \eta_k^{max} \prod_{s=k+1}^{T-1} \left(1 - \eta_s^{min} \frac{L\mu}{L+\mu}\right) + 4GL \sum_{k=1}^{T-1} \left( \eta_k^{max} \sum_{s=0}^{k-1} \eta_s^{max} \lambda^{k-1-s} \right) \prod_{j=k+1}^{T-1} \left(1 - \eta_j^{min} \frac{L\mu}{L+\mu}\right)
\end{aligned} \tag{9}$$

where the initial difference is zero that  $\mathbf{X}^0 = 0$  and  $\mathbf{Y}^0 = 0$ . And we use the last iterate as the output of Algorithm  $\mathcal{A}$ , i.e.,  $(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathcal{A}_{\mathbf{y}}(\mathcal{S})) = (\mathbf{x}^T, \mathbf{y}^T)$ .

For case **a**. when  $\eta_{x,t}$  and  $\eta_{y,t}$  are fixed, we further get:

$$\begin{aligned}
\epsilon_{sta}^{arg}(\mathcal{A}) &\leq \frac{2G}{n} \sum_{k=0}^{T-1} \eta^{max} \left(1 - \eta^{min} \frac{L\mu}{L+\mu}\right)^{T-k-1} + 4GL \sum_{k=1}^{T-1} \left( \eta^{max} \sum_{s=0}^{k-1} \eta^{max} \lambda^{k-1-s} \right) \left(1 - \eta^{min} \frac{L\mu}{L+\mu}\right)^{T-k-1} \\
&\leq \left( 4\eta^{max} GL \eta^{max} \frac{1}{1-\lambda} + \frac{2\eta^{max} G}{n} \right) \sum_{k=0}^{T-1} \left(1 - \eta^{min} \frac{L\mu}{L+\mu}\right)^k \\
&\leq 2G \frac{L+\mu}{\eta^{min} L\mu} \left( \frac{2(\eta^{max})^2 L}{1-\lambda} + \frac{\eta^{max}}{n} \right)
\end{aligned}$$

Further when  $\eta^{max} = \eta^{min} \triangleq \eta$ , the argument stability turns out to be

$$\epsilon_{sta}^{arg}(\mathcal{A}) \leq 2G \frac{L+\mu}{L\mu} \left( \frac{2\eta L}{1-\lambda} + \frac{1}{n} \right)$$

For case **b**. when  $\eta_t^{max} = \frac{1}{\mu(t+1)^c}$  and  $\eta_t^{min} = \frac{1}{\mu(t+1)}$ ,  $c \leq 1$ . Since  $1 \pm a \leq \exp\{a\}$ , we have:

$$\begin{aligned}
\prod_{j=k+1}^{T-1} \left(1 - \eta_j^{min} \frac{L\mu}{L+\mu}\right) &= \prod_{j=k+1}^{T-1} \left(1 - \frac{1}{j+1} \frac{L}{L+\mu}\right) \\
&\leq \prod_{j=k+1}^{T-1} \exp\left\{-\frac{1}{j+1} \frac{L}{L+\mu}\right\} \\
&= \exp\left\{\sum_{j=k+1}^{T-1} -\frac{1}{j+1} \frac{L}{L+\mu}\right\} \\
&\leq \exp\left\{-\frac{L}{L+\mu} (\ln T - \ln(k+1))\right\} \\
&= \left(\frac{k+1}{T}\right)^{\frac{L}{L+\mu}}
\end{aligned}$$

Then back to inequality (9), we can simplify it as:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathcal{A}_{\mathbf{x}}(\mathcal{S}) - \mathcal{A}_{\mathbf{x}}(\mathcal{S}') \\ \mathcal{A}_{\mathbf{y}}(\mathcal{S}) - \mathcal{A}_{\mathbf{y}}(\mathcal{S}') \end{pmatrix} \right\| \\
&\leq \frac{2G}{n} \sum_{k=0}^{T-1} \frac{1}{\mu(k+1)^c} \left(\frac{k+1}{T}\right)^{\frac{L}{L+\mu}} + 4GL \sum_{k=1}^{T-1} \left( \frac{1}{\mu(k+1)^c} \sum_{s=0}^{k-1} \frac{1}{\mu(s+1)^c} \lambda^{k-1-s} \right) \left(\frac{k+1}{T}\right)^{\frac{L}{L+\mu}} \\
&\leq \frac{2G}{\mu n T^{\frac{L}{L+\mu}}} \sum_{k=0}^{T-1} \frac{1}{(k+1)^{c-\frac{L}{L+\mu}}} + \frac{4GL}{\mu^2 T^{\frac{L}{L+\mu}}} \sum_{k=1}^{T-1} \frac{1}{(k+1)^{c-\frac{L}{L+\mu}}} \frac{C_{\lambda}}{k^c}
\end{aligned}$$

□

## E.2 Proof of the Empirical Risk

Then we are going to present the optimization error of D-SGDA in the SC-SC condition.

**Theorem 7 (Optimization error).** *Under assumption 1,2,3, each local function is  $f_i$  is  $\mu_{\mathbf{x}}\text{SC}-\mu_{\mathbf{y}}\text{SC}$ . We further bound the restriction set by  $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\| \leq C_{\mathbf{x}}$  and  $\sup_{\mathbf{y} \in \mathcal{Y}} \|\mathbf{y}\| \leq C_{\mathbf{y}}$ , then we have the strong primal-dual empirical risk over the dataset  $\mathcal{S}$  on the average output in  $T$  iterations as following:*

a. When learning rates  $\eta_{x,t}$  and  $\eta_{y,t}$  are fixed,

$$\Delta_{\mathcal{S}}^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \leq \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min} T} + \eta^{max} G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1-\lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}}.$$

b. When learning rates are varying that  $\eta_{\mathbf{x},t} = \frac{1}{\mu_{\mathbf{x}}(t+1)^{c_{\mathbf{x}}}}$  and  $\eta_{\mathbf{y},t} = \frac{1}{\mu_{\mathbf{y}}(t+1)^{c_{\mathbf{y}}}}$ ,

$$\Delta_S^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \leq \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + T_{\mathbf{x}} + T_{\mathbf{y}} + T_{max}.$$

where  $\alpha \in \{\mathbf{x}, \mathbf{y}\}$ ,  $k_{min} = \min\{c_{\mathbf{x}}, c_{\mathbf{y}}\}$  and  $\mu = \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ :

$$T_{\alpha} = \begin{cases} \frac{G^2}{2\mu_{\alpha}} \frac{1+\ln T}{T} & c_{\alpha} = 1 \\ \frac{G^2}{2\mu_{\alpha}(1-c_{\alpha})T^{c_{\alpha}}} & 0 < c_{\alpha} < 1 \end{cases}; \quad T_{max} = \begin{cases} \frac{4GLC_{\lambda}(C_{\mathbf{x}}+C_{\mathbf{y}})\ln T}{\mu T} & k_{min} = 1 \\ \frac{4GLC_{\lambda}(C_{\mathbf{x}}+C_{\mathbf{y}})}{\mu(1-k_{min})T^{k_{min}}} & 0 < k_{min} < 1 \end{cases}.$$

*Proof of Theorem 7.* First, we should notice that when each  $f_i$  owns the property of  $\mu_{\mathbf{x}}$ -strong convexity and  $\mu_{\mathbf{y}}$ -strong concavity, then the linear summation should inherit the properties, i.e.,  $F_S(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n f_i(\mathbf{x}, \mathbf{y}; \xi_{i,l})$  is  $\mu_{\mathbf{x}}$ -strongly convex on  $\mathbf{x}$  and  $\mu_{\mathbf{y}}$ -strongly concave on  $\mathbf{y}$ .

First observe the update rule of  $\mathbf{x}$  (see Algorithm 1), for any given  $\mathbf{x} \in \mathcal{X}$ :

$$\begin{aligned} & \|\mathbf{x}^{t+1} - \mathbf{x}\|^2 \\ & \leq \|\mathbf{x}^t - \frac{1}{m} \sum_{i=1}^m \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \mathbf{x}\|^2 \\ & = \|\mathbf{x}^t - \mathbf{x}\|^2 + \left\| \frac{1}{m} \sum_{i=1}^m \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \right\|^2 - 2 \langle \mathbf{x}^t - \mathbf{x}, \frac{1}{m} \sum_{i=1}^m \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle \end{aligned} \quad (10)$$

where for the first inequality, we can recall how we transform in the proof for stability (see Appendix E.1) that  $\mathbf{x}^{t+1} = (\mathbf{X}^{t+1})^T \frac{\mathbf{1}_m}{m} = (\mathbf{W} \mathbf{X}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t))^T \frac{\mathbf{1}_m}{m} = \mathbf{x}^t - \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{X}^t, \mathbf{Y}^t; \boldsymbol{\xi}^t)^T \frac{\mathbf{1}_m}{m}$  regardless of the projection.

For the second term, we have:

$$\begin{aligned} & \left\| \frac{1}{m} \sum_{i=1}^m \eta_{\mathbf{x},t} \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \right\|^2 \\ & = \frac{\eta_{\mathbf{x},t}^2}{m^2} \left[ \sum_{i=1}^m \|\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)})\|^2 + \sum_{i \neq k} \langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}), \nabla_{\mathbf{x}} f_k(\mathbf{x}_k^t, \mathbf{y}_k^t; \xi_{k,j_t(k)}) \rangle \right] \\ & \leq \frac{\eta_{\mathbf{x},t}^2}{m^2} (mG^2 + (m^2 - m)G^2) \\ & = \eta_{\mathbf{x},t}^2 G^2 \end{aligned} \quad (11)$$

We decompose the third term in association with the empirical function that:

$$\begin{aligned} & 2\eta_{\mathbf{x},t} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle \\ & = 2\eta_{\mathbf{x},t} \langle \mathbf{x} - \mathbf{x}^t, \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle + 2\eta_{\mathbf{x},t} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \\ & \leq 2\eta_{\mathbf{x},t} (F_S(\mathbf{x}, \mathbf{y}^t) - F_S(\mathbf{x}^t, \mathbf{y}^t)) - \eta_{\mathbf{x},t} \mu_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}^t\|^2 \\ & \quad + 2\eta_{\mathbf{x},t} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \end{aligned} \quad (12)$$

where the last inequality is due to the strong convexity of the empirical function  $F_S$  on parameter  $\mathbf{x}$ .

Then we combine the inequalities (10) (11) (12) and we can get:

$$\begin{aligned} 2\eta_{\mathbf{x},t} (F_S(\mathbf{x}^t, \mathbf{y}^t) - F_S(\mathbf{x}, \mathbf{y}^t)) & \leq (1 - \eta_{\mathbf{x},t} \mu_{\mathbf{x}}) \|\mathbf{x} - \mathbf{x}^t\|^2 - \|\mathbf{x}^{t+1} - \mathbf{x}\|^2 + \eta_{\mathbf{x},t}^2 G^2 \\ & \quad + 2\eta_{\mathbf{x},t} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \end{aligned} \quad (13)$$

For case **a**. when learning rates  $\eta_{\mathbf{x},t}$  and  $\eta_{\mathbf{y},t}$  are fixed and we write them as  $\eta_{\mathbf{x}}$  and  $\eta_{\mathbf{y}}$  respectively. Taking summation over the above inequality from  $t = 0$  to  $t = T$  and using the concavity of the empirical function on  $\mathbf{y}$ :

$$\begin{aligned}
& \sum_{t=0}^{T-1} 2\eta_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) - F_S(\mathbf{x}, \mathbf{y}_{ave}^T) \\
& \leq \sum_{t=0}^{T-1} 2\eta_{\mathbf{x}} (F_S(\mathbf{x}^t, \mathbf{y}^t) - F_S(\mathbf{x}, \mathbf{y}^t)) \\
& \leq (1 - \eta_{\mathbf{x}} \mu_{\mathbf{x}}) \|\mathbf{x}\|^2 + \eta_{\mathbf{x}}^2 G^2 T + 2\eta_{\mathbf{x}} \sum_{t=0}^{T-1} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \\
& \leq (1 - \eta_{\mathbf{x}} \mu_{\mathbf{x}}) \|\mathbf{x}\|^2 + \eta_{\mathbf{x}}^2 G^2 T + 2\eta_{\mathbf{x}} \sum_{t=0}^{T-1} \langle \mathbf{x}^t, \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle \\
& \quad + 2\eta_{\mathbf{x}} \|\mathbf{x}\| \left\| \sum_{t=0}^{T-1} \left( \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \right) \right\|
\end{aligned}$$

Then we take expectation on the randomness of the algorithm (notice that we do not take expectation on the randomness of dataset) on both sides of above inequality and choose the infinity of  $\mathbf{x}$  on the left side since the above inequality holds for any  $\mathbf{x} \in \mathcal{X}$ :

$$\begin{aligned}
& \sum_{t=0}^{T-1} 2\eta_{\mathbf{x}} \mathbb{E}_{\mathcal{A}} [F_S(\mathbf{x}^t, \mathbf{y}^t) - \inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T)] \\
& \leq (1 - \eta_{\mathbf{x}} \mu_{\mathbf{x}}) C_{\mathbf{x}}^2 + \eta_{\mathbf{x}}^2 G^2 T + 2\eta_{\mathbf{x}} \underbrace{\sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} \langle \mathbf{x}^t, \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle}_I \\
& \quad + 2C_{\mathbf{x}} \eta_{\mathbf{x}} \underbrace{\mathbb{E}_{\mathcal{A}} \left\| \sum_{t=0}^{T-1} \left( \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \right) \right\|}_{II}
\end{aligned} \tag{14}$$

For term  $I$ , it is decomposed as:

$$\begin{aligned}
& \sum_{t=0}^{T-1} \mathbb{E}_{j_t} \langle \mathbf{x}^t, \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle \\
& = \sum_{t=0}^{T-1} \mathbb{E}_{j_t} \langle \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,l}) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) \rangle \\
& \quad + \sum_{t=0}^{T-1} \mathbb{E}_{j_t} \langle \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle \\
& \stackrel{(i)}{\leq} 2C_{\mathbf{x}} \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} \left\| \frac{1}{m} \sum_{i=1}^m (\nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)})) \right\| \\
& \leq 2C_{\mathbf{x}} \sum_{t=0}^{T-1} \frac{L}{m} \mathbb{E}_{\mathcal{A}} \left[ \sum_{i=1}^m (\|\mathbf{x}^t - \mathbf{x}_i^t\|^2 + \|\mathbf{y}^t - \mathbf{y}_i^t\|^2)^{1/2} \right] \\
& \stackrel{(ii)}{\leq} 4C_{\mathbf{x}} GL \sum_{t=0}^{T-1} \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s}
\end{aligned}$$

where inequality (i) is because of the fact that taking expectation on the randomness of  $j_t$  is equal to the empirical result on the dataset  $\mathcal{S}$  and inequality (ii) is due to Cauchy-Schwarz inequality and Lemma 2.

Then for the term  $II$ , we decompose its corresponding quadratic one:

$$\begin{aligned}
& \left( \mathbb{E}_{\mathcal{A}} \left\| \sum_{t=0}^{T-1} \left( \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \right) \right\| \right)^2 \\
& \leq \mathbb{E}_{\mathcal{A}} \left\| \sum_{t=0}^{T-1} \left( \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \right) \right\|^2 \\
& = \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} \left\| \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,l}) \right\|^2 \\
& \quad + \sum_{t \neq t'} \mathbb{E}_{\mathcal{A}} \left\langle \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,l}), \right. \\
& \quad \quad \left. \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{t'}, \mathbf{y}_i^{t'}; \xi_{i,j_{t'}(i)}) - \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n \nabla_{\mathbf{x}} f_i(\mathbf{x}^{t'}, \mathbf{y}^{t'}; \xi_{i,l}) \right\rangle \\
& \leq 4G^2T + \sum_{t \neq t'} \mathbb{E}_{\mathcal{A}} \left\langle \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) \right. \\
& \quad \quad + \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}) - \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,l}), \\
& \quad \quad \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{t'}, \mathbf{y}_i^{t'}; \xi_{i,j_{t'}(i)}) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^{t'}, \mathbf{y}^{t'}; \xi_{i,j_{t'}(i)}) \\
& \quad \quad \left. + \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^{t'}, \mathbf{y}^{t'}; \xi_{i,j_{t'}(i)}) - \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{l=1}^n \nabla_{\mathbf{x}} f_i(\mathbf{x}^{t'}, \mathbf{y}^{t'}; \xi_{i,l}) \right\rangle \\
& \stackrel{(a)}{=} 4G^2T + \sum_{t \neq t'} \mathbb{E}_{\mathcal{A}} \left\langle \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)}), \right. \\
& \quad \quad \left. \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{t'}, \mathbf{y}_i^{t'}; \xi_{i,j_{t'}(i)}) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}^{t'}, \mathbf{y}^{t'}; \xi_{i,j_{t'}(i)}) \right\rangle \\
& \leq 4G^2T + \sum_{t \neq t'} \mathbb{E}_{\mathcal{A}} \left( \left\| \frac{1}{m} \sum_{i=1}^m (\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}^t, \mathbf{y}^t; \xi_{i,j_t(i)})) \right\| \right. \\
& \quad \quad \left. \left\| \frac{1}{m} \sum_{i=1}^m (\nabla_{\mathbf{x}} f_i(\mathbf{x}_i^{t'}, \mathbf{y}_i^{t'}; \xi_{i,j_{t'}(i)}) - \nabla_{\mathbf{x}} f_i(\mathbf{x}^{t'}, \mathbf{y}^{t'}; \xi_{i,j_{t'}(i)})) \right\| \right) \\
& \stackrel{(b)}{\leq} 4G^2T + \sum_{t \neq t'} \frac{L^2}{m} \mathbb{E}_{\mathcal{A}} \left[ \sum_{i=1}^m (\|\mathbf{x}^t - \mathbf{x}_i^t\|^2 + \|\mathbf{y}^t - \mathbf{y}_i^t\|^2) \cdot \sum_{i=1}^m (\|\mathbf{x}^{t'} - \mathbf{x}_i^{t'}\|^2 + \|\mathbf{y}^{t'} - \mathbf{y}_i^{t'}\|^2) \right]^{1/2} \\
& \leq 4G^2T + \sum_{t \neq t'} 4G^2L^2 \left( \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \right) \left( \sum_{s=0}^{t'-1} \eta_s^{max} \lambda^{t-1-s} \right)
\end{aligned}$$

where equality (a) owes to that taking expectation on the randomness of  $j_t$  or  $j_{t'}$  equals to the empirical risk and inequality (b) is due to the Lipschitz smoothness of each  $f_i$  and Cauchy-Schwarz inequality.

Combining above inequalities of term  $I$  and  $II$  into the inequality (14) and we can summarize as:

$$\begin{aligned} & \sum_{t=0}^{T-1} 2\eta_{\mathbf{x}} \mathbb{E}_{\mathcal{A}} [F_S(\mathbf{x}^t, \mathbf{y}^t) - \inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T)] \\ & \leq (1 - \eta_{\mathbf{x}} \mu_{\mathbf{x}}) C_{\mathbf{x}}^2 + \eta_{\mathbf{x}}^2 G^2 T + 4\eta_{\mathbf{x}} C_{\mathbf{x}} G L \sum_{t=0}^{T-1} \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \\ & \quad + 2C_{\mathbf{x}} \eta_{\mathbf{x}} \sqrt{4G^2 T + 4G^2 L^2 \sum_{t \neq t'} \left( \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \right) \left( \sum_{s=0}^{t'-1} \eta_s^{max} \lambda^{t'-1-s} \right)} \end{aligned}$$

Dividing both sides by  $T$  and we can get:

$$\begin{aligned} & \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} [F_S(\mathbf{x}^t, \mathbf{y}^t)] - \mathbb{E}_{\mathcal{A}} [\inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T)] \\ & \leq \frac{(1 - \eta_{\mathbf{x}} \mu_{\mathbf{x}}) C_{\mathbf{x}}^2}{2\eta_{\mathbf{x}} T} + \frac{\eta_{\mathbf{x}} G^2}{2} + \frac{4C_{\mathbf{x}} G L \eta^{max}}{1 - \lambda} + \frac{2C_{\mathbf{x}} G}{\sqrt{T}} \end{aligned}$$

And we can get the other-hand result in the same symmetric way:

$$\begin{aligned} & \mathbb{E}_{\mathcal{A}} [\sup_{\mathbf{y} \in \mathcal{Y}} F_S(\mathbf{x}_{ave}^T, \mathbf{y})] - \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} [F_S(\mathbf{x}^t, \mathbf{y}^t)] \\ & \leq \frac{(1 - \eta_{\mathbf{y}} \mu_{\mathbf{y}}) C_{\mathbf{y}}^2}{2\eta_{\mathbf{y}} T} + \frac{\eta_{\mathbf{y}} G^2}{2} + \frac{4C_{\mathbf{y}} G L \eta^{max}}{1 - \lambda} + \frac{2C_{\mathbf{y}} G}{\sqrt{T}} \end{aligned}$$

Combining above two inequalities we can get the result:

$$\Delta_S^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \leq \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min} T} + \eta^{max} G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}}) G L \eta^{max}}{1 - \lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}}) G}{\sqrt{T}}$$

Then we come to the case **b.** when  $\eta_{\mathbf{x},t} = \frac{1}{\mu_{\mathbf{x}} \cdot (t+1)^{c_{\mathbf{x}}}}$  and  $\eta_{\mathbf{y},t} = \frac{1}{\mu_{\mathbf{y}} \cdot (t+1)^{c_{\mathbf{y}}}}$ . Back to the inequality (13), we simplify it into:

$$\begin{aligned} & \frac{2}{\mu_{\mathbf{x}} (t+1)^{c_{\mathbf{x}}}} (F_S(\mathbf{x}^t, \mathbf{y}^t) - F_S(\mathbf{x}, \mathbf{y}^t)) \leq (1 - \frac{1}{(t+1)^{c_{\mathbf{x}}}}) \|\mathbf{x} - \mathbf{x}^t\|^2 - \|\mathbf{x}^{t+1} - \mathbf{x}\|^2 + \frac{G^2}{\mu_{\mathbf{x}}^2 (t+1)^{2c_{\mathbf{x}}}} \\ & \quad + \frac{2}{\mu_{\mathbf{x}} (t+1)^{c_{\mathbf{x}}}} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \\ & F_S(\mathbf{x}^t, \mathbf{y}^t) - F_S(\mathbf{x}, \mathbf{y}^t) \leq \frac{\mu_{\mathbf{x}}}{2} ((t+1)^{c_{\mathbf{x}}} - 1) \|\mathbf{x} - \mathbf{x}^t\|^2 - \frac{\mu_{\mathbf{x}}}{2} (t+1)^{c_{\mathbf{x}}} \|\mathbf{x}^{t+1} - \mathbf{x}\|^2 + \frac{G^2}{2\mu_{\mathbf{x}} (t+1)^{c_{\mathbf{x}}}} \\ & \quad + \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \end{aligned}$$

Taking summation on the above inequality from  $t = 0$  to  $t = T - 1$  and making use of the convexity of  $F_S$  on the second parameter that:

$$\begin{aligned} & \sum_{t=0}^{T-1} [F_S(\mathbf{x}^t, \mathbf{y}^t) - F_S(\mathbf{x}, \mathbf{y}_{ave}^T)] \\ & \leq \frac{G^2}{2\mu_{\mathbf{x}}} \sum_{t=0}^{T-1} \frac{1}{(t+1)^{c_{\mathbf{x}}}} + \sum_{t=0}^{T-1} \langle \mathbf{x} - \mathbf{x}^t, \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \rangle \\ & \leq \frac{G^2}{2\mu_{\mathbf{x}}} \sum_{t=0}^{T-1} \frac{1}{(t+1)^{c_{\mathbf{x}}}} + \sum_{t=0}^{T-1} \langle \mathbf{x}^t, \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) - \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) \rangle \\ & \quad + \|\mathbf{x}\| \left\| \sum_{t=0}^{T-1} \left( \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f_i(\mathbf{x}_i^t, \mathbf{y}_i^t; \xi_{i,j_t(i)}) - \nabla_{\mathbf{x}} F_S(\mathbf{x}^t, \mathbf{y}^t) \right) \right\| \end{aligned}$$

Next we will take expectation on the randomness of the algorithm on both sides and proceed in the same way as we do in case **a**. that:

$$\begin{aligned} & \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} \left[ F_S(\mathbf{x}^t, \mathbf{y}^t) - \inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T) \right] \\ & \leq \frac{G^2}{2\mu_{\mathbf{x}}} \sum_{t=0}^{T-1} \frac{1}{(t+1)^{c_{\mathbf{x}}}} + 2C_{\mathbf{x}}GL \sum_{t=0}^{T-1} \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \\ & \quad + C_{\mathbf{x}} \sqrt{4G^2T + 4G^2L^2 \sum_{t \neq t'} \left( \sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s} \right) \left( \sum_{s=0}^{t'-1} \eta_s^{max} \lambda^{t-1-s} \right)} \end{aligned}$$

Without loss of generalization, we assume  $\eta_t^{max} = \eta_{\mathbf{x},t} = \frac{1}{\mu_{\mathbf{x}}(t+1)^{c_{\mathbf{x}}}}$ . Therefore the summation  $\sum_{s=0}^{t-1} \eta_s^{max} \lambda^{t-1-s}$  comes out to be  $\sum_{s=0}^{t-1} \frac{\lambda^{t-1-s}}{\mu_{\mathbf{x}}(t+1)^{c_{\mathbf{x}}}} \leq \frac{C_{\lambda}}{\mu_{\mathbf{x}} t^{c_{\mathbf{x}}}}$ . Then we have to analyse in different categories for the value of  $c_{\mathbf{x}}$ .

First when  $c_{\mathbf{x}} = 1$ , the summation result turns out to be  $\sum_{t=0}^{T-1} \frac{1}{t+1} \leq 1 + \ln T$ ,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} [F_S(\mathbf{x}^t, \mathbf{y}^t)] - \inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T) \leq \frac{G^2}{2\mu_{\mathbf{x}}} \frac{1 + \ln T}{T} + \frac{4GLC_{\mathbf{x}}C_{\lambda} \ln T}{\mu_{\mathbf{x}}T} + \frac{2GC_{\mathbf{x}}}{\sqrt{T}}$$

When  $0 < c_{\mathbf{x}} < 1$ ,  $\sum_{t=0}^{T-1} \frac{1}{(t+1)^{c_{\mathbf{x}}}} \leq \frac{T^{1-c_{\mathbf{x}}}}{1-c_{\mathbf{x}}}$ , then we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{A}} [F_S(\mathbf{x}^t, \mathbf{y}^t)] - \inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T) \leq \frac{G^2}{2\mu_{\mathbf{x}}(1-c_{\mathbf{x}})T^{c_{\mathbf{x}}}} + \frac{4GLC_{\mathbf{x}}C_{\lambda}}{\mu_{\mathbf{x}}(1-c_{\mathbf{x}})T^{c_{\mathbf{x}}}} + \frac{2GC_{\mathbf{x}}}{\sqrt{T}}$$

It is in a similar way to get the symmetric result on the other hand and combining both results we can obtain the strong primal-dual empirical risk:

$$\mathbb{E}_{\mathcal{A}} [\sup_{\mathbf{y} \in \mathcal{Y}} F_S(\mathbf{x}_{ave}^T, \mathbf{y}) - \inf_{\mathbf{x} \in \mathcal{X}} F_S(\mathbf{x}, \mathbf{y}_{ave}^T)] \leq \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + T_{\mathbf{x}} + T_{\mathbf{y}} + T_{max}$$

where

$$T_{\mathbf{x}} = \begin{cases} \frac{G^2}{2\mu_{\mathbf{x}}} \frac{1 + \ln T}{T} & c_{\mathbf{x}} = 1 \\ \frac{G^2}{2\mu_{\mathbf{x}}(1-c_{\mathbf{x}})T^{c_{\mathbf{x}}}} & 0 < c_{\mathbf{x}} < 1 \end{cases}; \quad T_{\mathbf{y}} = \begin{cases} \frac{G^2}{2\mu_{\mathbf{y}}} \frac{1 + \ln T}{T} & c_{\mathbf{y}} = 1 \\ \frac{G^2}{2\mu_{\mathbf{y}}(1-c_{\mathbf{y}})T^{c_{\mathbf{y}}}} & 0 < c_{\mathbf{y}} < 1 \end{cases}$$

and

$$c_{min} = \min\{c_{\mathbf{x}}, c_{\mathbf{y}}\}; \quad \mu = \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}; \quad T_{max} = \begin{cases} \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}}) \ln T}{\mu T} & c_{min} = 1 \\ \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu(1-c_{min})T^{c_{min}}} & 0 < c_{min} < 1 \end{cases}$$

□

### E.3 Proof of Strong/Weak Primal-Dual Population Risk

In this part, we are going to prove the strong and weak primal-dual population risk of the algorithm D-SGDA. Actually this is an obvious result as a summary of above lemmas and theorems.

*Proof of Theorem 3.* As we introduced the population risk (see Def. 2), we decompose the population risk into generalization gap and empirical risk that:

$$\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) = (\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) - \Delta_S^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)) + \Delta_S^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$$

Notice that we use the average iterate  $(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$  instead of the last iterate to denote the output of D-SGDA. The first term is the averaged version of the strong primal-dual generalization gap and we will make some adjustments.

First we have the argument stability bound of D-SGDA that  $\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^T - \hat{\mathbf{x}}^T \\ \mathbf{y}^T - \hat{\mathbf{y}}^T \end{pmatrix} \right\| \leq \epsilon_{sta}^{arg}(\mathcal{A})$ , where  $(\mathbf{x}^T, \mathbf{y}^T)$  and  $(\hat{\mathbf{x}}^T, \hat{\mathbf{y}}^T)$  denote the  $T$ -th output when D-SGDA is executed on the neighboring dataset

respectively. So we can make use of the convexity of the norm and obtain the argument stability of averaged-version:

$$\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}_{ave}^T - \dot{\mathbf{x}}_{ave}^T \\ \mathbf{y}_{ave}^T - \dot{\mathbf{y}}_{ave}^T \end{pmatrix} \right\| \leq \mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^T - \dot{\mathbf{x}}^T \\ \mathbf{y}^T - \dot{\mathbf{y}}^T \end{pmatrix} \right\| \leq \epsilon_{sta}^{arg}(\mathcal{A}) \quad (15)$$

Then the strong primal-dual generalization gap should be  $G\sqrt{2 + \frac{2L^2}{\mu^2}\epsilon_{sta}^{arg}(\mathcal{A})}$  following Thm. 1 when each  $f_i$  is strongly convex w.r.t.  $\mathbf{x}$  and strongly concave w.r.t.  $\mathbf{y}$ .

At last, the strong primal-dual empirical risk  $\mathbb{E}_{\mathcal{A}}[\Delta_S^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)]$  is studied in Thm. 7. So we can combine above two bounds to analyze the strong primal-dual population risk in different categories when learning rates are fixed:

$$\begin{aligned} & \mathbb{E}_{\mathcal{A}}[\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)] \\ & \leq G\sqrt{2 + \frac{2L^2}{\mu^2}} \left( 2G \frac{L + \mu}{\eta^{min} L \mu} \left( \frac{2(\eta^{max})^2 L}{1 - \lambda} + \frac{\eta^{max}}{n} \right) + \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min} T} + \eta^{max} G^2 \right) \\ & \quad + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1 - \lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}} \end{aligned}$$

And when learning rates are varying that  $\eta_t^{min} = \frac{1}{\mu(t+1)}$  and  $\eta_t^{max} = \frac{1}{\mu(t+1)^c}$ ,  $c = 1$ , requiring  $2c > \frac{L}{L+\mu} + 1$ :

$$\begin{aligned} & \mathbb{E}_{\mathcal{A}}[\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)] \\ & \leq G\sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G}{\mu n T^{\frac{L}{L+\mu}}} \sum_{k=0}^{T-1} \frac{1}{(k+1)^{c - \frac{L}{L+\mu}}} + \frac{4GL}{\mu^2 T^{\frac{L}{L+\mu}}} \sum_{k=1}^{T-1} \frac{1}{(k+1)^{c - \frac{L}{L+\mu}}} \frac{C_{\lambda}}{k^c} \right) \\ & \quad + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2}{\mu} \frac{1 + \ln T}{T} + \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}}) \ln T}{\mu T} \\ & \leq G\sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G}{\mu^{\frac{L}{L+\mu}} n} + \frac{4GLC_{\lambda}}{\mu^2(1 - \frac{L}{L+\mu})} \frac{1}{T^{\frac{L}{L+\mu}}} \right) + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2}{\mu} \frac{1 + \ln T}{T} \\ & \quad + \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}}) \ln T}{\mu T} \end{aligned}$$

When  $c < 1$  and requiring  $2c > \frac{L}{L+\mu} + 1$ :

$$\begin{aligned} & \mathbb{E}_{\mathcal{A}}[\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)] \\ & \leq G\sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G}{\mu n T^{\frac{L}{L+\mu}}} \sum_{k=0}^{T-1} \frac{1}{(k+1)^{c - \frac{L}{L+\mu}}} + \frac{4GL}{\mu^2 T^{\frac{L}{L+\mu}}} \sum_{k=1}^{T-1} \frac{1}{(k+1)^{c - \frac{L}{L+\mu}}} \frac{C_{\lambda}}{k^c} \right) \\ & \quad + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2}{2\mu} \left( \frac{1 + \ln T}{T} + \frac{1}{(1-c)T^c} \right) + \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu(1-c)T^c} \\ & \leq G\sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G}{\mu(1-c + \frac{L}{L+\mu})} \frac{T^{1-c}}{n} + \frac{4GLC_{\lambda}}{\mu^2(2c - \frac{L}{L+\mu} - 1)} \frac{1}{T^{\frac{L}{L+\mu}}} \right) + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} \\ & \quad + \frac{G^2}{2\mu} \left( \frac{1 + \ln T}{T} + \frac{1}{(1-c)T^c} \right) + \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu(1-c)T^c} \end{aligned}$$

When  $c < 1$  and  $2c = \frac{L}{L+\mu} + 1$ :

$$\begin{aligned}
& \mathbb{E}_{\mathcal{A}}[\Delta^s(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)] \\
& \leq G \sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G}{\mu n T^{\frac{L}{L+\mu}}} \sum_{k=0}^{T-1} \frac{1}{(k+1)^{c-\frac{L}{L+\mu}}} + \frac{4GL}{\mu^2 T^{\frac{L}{L+\mu}}} \sum_{k=1}^{T-1} \frac{1}{(k+1)^{c-\frac{L}{L+\mu}}} \frac{C_\lambda}{k^c} \right) \\
& \quad + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2}{2\mu} \left( \frac{1 + \ln T}{T} + \frac{1}{(1-c)T^c} \right) + \frac{4GLC_\lambda(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu(1-c)T^c} \\
& \leq G \sqrt{2 + \frac{2L^2}{\mu^2}} \left( \frac{2G T^{1-c}}{c\mu n} + \frac{4GLC_\lambda}{\mu^2} \frac{\ln T}{T^{\frac{L}{L+\mu}}} \right) + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2}{2\mu} \left( \frac{1 + \ln T}{T} + \frac{1}{(1-c)T^c} \right) \\
& \quad + \frac{4GLC_\lambda(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu(1-c)T^c}
\end{aligned}$$

□

## F Proof in the Convex-Concave Case

In this section, we will provide corresponding proof for argument stability, optimization error and weak primal-dual population risk in the C-C condition.

### F.1 Proof of Stability

*Proof of Theorem 4.* Analogous to Eq. (9) in the proof for SC-SC (see Appendix E.1), considering C-C a special case for  $\mu_{\mathbf{x}}\text{SC}-\mu_{\mathbf{y}}\text{SC}$  when  $\mu_{\mathbf{x}} = 0, \mu_{\mathbf{y}} = 0$ .

Thus we can get the result:

$$\begin{aligned}
& \mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathcal{A}_{\mathbf{x}}(\mathcal{S}) - \mathcal{A}_{\mathbf{x}}(\mathcal{S}') \\ \mathcal{A}_{\mathbf{y}}(\mathcal{S}) - \mathcal{A}_{\mathbf{y}}(\mathcal{S}') \end{pmatrix} \right\| \\
& \leq \frac{2G}{n} \sum_{k=0}^{T-1} \eta_k^{max} + 4GL \sum_{k=1}^{T-1} \left( \eta_k^{max} \sum_{s=0}^{k-1} \eta_s^{max} \lambda^{k-1-s} \right)
\end{aligned}$$

□

### F.2 Proof of the Empirical Risk

Similar to Thm. 7, considering C-C as a special case of SC-SC for  $\mu_{\mathbf{x}} = \mu_{\mathbf{y}} = 0$ , we can get the weak primal-dual empirical risk in the following corollary, where we use the Jensen's inequality that  $\Delta_S^w(\mathbf{x}, \mathbf{y}) \leq \Delta_S^s(\mathbf{x}, \mathbf{y})$ .

**Corollary 1.** *Under assumption 1,2,3 and the restriction that  $\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\| \leq C_{\mathbf{x}}$  and  $\sup_{\mathbf{y} \in \mathcal{Y}} \|\mathbf{y}\| \leq C_{\mathbf{y}}$ , each local function is  $f_i$  is C-C. We have the weak primal-dual empirical risk over the dataset  $\mathcal{S}$  on the average output in  $T$  iterations as following for fixed learning rates:*

$$\Delta_S^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \leq \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min}T} + \eta^{max}G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1-\lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}}.$$

### F.3 Proof of Weak Primal-Dual Population Risk

*Proof of Theorem 5.* When each local function  $f_i$  is not strongly convex or strongly concave, we can not get access to the strong primal-dual generalization gap but weak primal-dual generalization gap. Following the same step in above proof that:

$$\Delta^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) = (\Delta^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) - \Delta_S^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)) + \Delta_S^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$$

Analogously we have the weak primal-dual generalization gap according to Thm. 1 that:

$$\Delta^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) - \Delta_S^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \leq \sqrt{2}G\epsilon_{sta}^{arg}(\mathcal{A})$$

where  $\epsilon_{sta}^{arg}(\mathcal{A}) \leq \frac{2G\eta^{max}T}{n} + \frac{4GL(\eta^{max})^2T}{1-\lambda}$  follows Thm. 4 when learning rates are fixed.

In the case without strong convexity or strong concavity, we select the fixed learning rates and following the Thm. 7, we can bound the weak primal-dual empirical risk as:

$$\Delta_{\mathcal{S}}^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) \leq \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min}T} + \eta^{max}G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1-\lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}}$$

Finally we combine above patterns and we can get the weak primal-dual population risk:

$$\begin{aligned} \Delta^w(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T) &\leq \sqrt{2}G\left(\frac{2G\eta^{max}T}{n} + \frac{4GL(\eta^{max})^2T}{1-\lambda}\right) \\ &\quad + \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min}T} + \eta^{max}G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1-\lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}} \end{aligned}$$

□

## G Proof in Nonconvex-Nonconcave Case

In this section, we will provide proof for weak stability and therefore we can derive the weak primal-dual generalization gap following Thm. 1 in the NC-NC condition.

### G.1 Important Lemmas

Before we present the proof for the stability bound in nonconvex-nonconcave case, we should first introduce an important lemma which describes the fact that D-SGDA will run several iterations before encountering the different samples. We extend the Lemma 3.11 in [13] and make adjustments on Lemma F.1 in [19] to fit our decentralized setting.

**Lemma 5.** *Let  $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$  and  $\mathcal{S}' = \{\mathcal{S}'_1, \dots, \mathcal{S}'_m\}$  be any arbitrary neighboring datasets,  $(\mathbf{x}^t, \mathbf{y}^t)$  and  $(\hat{\mathbf{x}}^t, \hat{\mathbf{y}}^t)$  represent output in  $t$ -th iteration under dataset  $\mathcal{S}$  and  $\mathcal{S}'$  respectively. We further require each local function is bounded that  $|f_i(\mathbf{x}, \mathbf{y}; \xi)| \leq B, \forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$ . Denoting  $\delta_t = \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\|$ , then we have:*

$$\mathbb{E}_{\mathcal{A}}[f(\mathbf{x}^t, \mathbf{y}'; \xi) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \xi) + f(\mathbf{x}', \mathbf{y}^t; \xi) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \xi)] \leq \sqrt{2}G\mathbb{E}_{\mathcal{A}} \left[ \delta_t \mid \delta_{t_0} = 0 \right] + \frac{Bmt_0}{n}.$$

*Proof.* First according to the property of Lipschitz continuity (see Assumption 1):

$$\begin{aligned} &f(\mathbf{x}^t, \mathbf{y}'; \xi) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \xi) + f(\mathbf{x}', \mathbf{y}^t; \xi) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \xi) \\ &\leq G\|\mathbf{x}^t - \hat{\mathbf{x}}^t\| + G\|\mathbf{y}^t - \hat{\mathbf{y}}^t\| \\ &\leq G\sqrt{2}\delta_t \end{aligned}$$

Then we decompose the expectation by the law of total expectation:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}}[f(\mathbf{x}^t, \mathbf{y}'; \xi) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \xi) + f(\mathbf{x}', \mathbf{y}^t; \xi) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \xi)] \\ &= \mathbb{P}(\delta_{t_0} = 0)\mathbb{E}_{\mathcal{A}} \left[ f(\mathbf{x}^t, \mathbf{y}'; \xi) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \xi) + f(\mathbf{x}', \mathbf{y}^t; \xi) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \xi) \mid \delta_{t_0} = 0 \right] \\ &\quad + \mathbb{P}(\delta_{t_0} \neq 0)\mathbb{E}_{\mathcal{A}} \left[ f(\mathbf{x}^t, \mathbf{y}'; \xi) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \xi) + f(\mathbf{x}', \mathbf{y}^t; \xi) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \xi) \mid \delta_{t_0} \neq 0 \right] \\ &\leq \sqrt{2}G\mathbb{E}_{\mathcal{A}} \left[ \delta_t \mid \delta_{t_0} = 0 \right] + B\mathbb{P}(\delta_{t_0} \neq 0) \end{aligned}$$

While the event that  $\delta_{t_0} \neq 0$  means the training process has already encountered the different samples before  $t_0$ :

$$\mathbb{P}(\delta_{t_0} \neq 0) \leq \sum_{t=1}^{t_0} \sum_{k=1}^m C_m^k \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k} = t_0 \left(1 - \left(1 - \frac{1}{n}\right)^m\right) \leq \frac{mt_0}{n}$$

Combining above inequalities and we can prove the Lemma. □

## G.2 Proof of Stability

*Proof of Theorem 6.* We are under the same setting as in the proof for SC-SC (see Appendix E.1) that  $(\mathbf{x}^t, \mathbf{y}^t)$  and  $(\hat{\mathbf{x}}^t, \hat{\mathbf{y}}^t)$  representing the  $t$ -th output over neighboring dataset  $\mathcal{S}$  and  $\mathcal{S}'$  respectively. And we assume each local dataset  $\mathcal{S}'_i$  in  $\mathcal{S}' = \{\mathcal{S}'_1, \dots, \mathcal{S}'_m\}$  differs from  $\mathcal{S}$  by the last sample without loss of generalization, i.e.,  $\mathcal{S}_i = \{\xi_{i,1}, \dots, \xi_{i,n}\}$  while  $\mathcal{S}'_i = \{\xi_{i,1}, \dots, \xi_{i,n-1}, \xi'_{i,n}\}$ . Analogous to the decomposition equality (7), we will bound the term  $I_1$  and  $I_2$  without convexity or concavity. Referring to Lemma 1, we have:

$$\begin{aligned} \|I_1\| &\leq \frac{m - m_0}{m} (1 + \eta_t^{\min} L) \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| \\ \|I_2\| &\leq \frac{m_0}{m} \left[ (1 + \eta_t^{\min} L) \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| + 2\eta_t^{\max} G \right] \end{aligned}$$

where we use the same denotation that  $\eta_t^{\min} = \min\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$  and  $\eta_t^{\max} = \max\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ . So we can get the similar result as the inequality (8):

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} \left[ \left\| \begin{pmatrix} \mathbf{x}^{t+1} - \hat{\mathbf{x}}^{t+1} \\ \mathbf{y}^{t+1} - \hat{\mathbf{y}}^{t+1} \end{pmatrix} \right\| \middle| \delta_{t_0} = 0 \right] \\ &\leq \sum_{m_0=0}^m C_m^{m_0} \left(1 - \frac{1}{n}\right)^{m-m_0} \left(\frac{1}{n}\right)^{m_0} \left( (1 + \eta_t^{\min} L) \mathbb{E}_{\mathcal{A}} \left[ \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| \middle| \delta_{t_0} = 0 \right] \right. \\ &\quad \left. + \frac{m_0}{m} 2\eta_t^{\max} G + 4\eta_t^{\max} LG \sum_{s=0}^{t-1} \eta_s^{\max} \lambda^{t-1-s} \right) \\ &\leq (1 + \eta_t^{\min} L) \mathbb{E}_{\mathcal{A}} \left[ \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| \middle| \delta_{t_0} = 0 \right] + 4\eta_t^{\max} LG \sum_{s=0}^{t-1} \eta_s^{\max} \lambda^{t-1-s} \\ &\quad + \sum_{m_0=0}^m C_m^{m_0} \left(1 - \frac{1}{n}\right)^{m-m_0} \left(\frac{1}{n}\right)^{m_0} \frac{m_0}{m} 2\eta_t^{\max} G \\ &= (1 + \eta_t^{\min} L) \mathbb{E}_{\mathcal{A}} \left[ \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| \middle| \delta_{t_0} = 0 \right] + 4\eta_t^{\max} LG \sum_{s=0}^{t-1} \eta_s^{\max} \lambda^{t-1-s} + \frac{2\eta_t^{\max} G}{n} \end{aligned}$$

Recursively applying above inequalities from  $t = t_0$  to  $t = T - 1$ :

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} \left[ \left\| \begin{pmatrix} \mathcal{A}_{\mathbf{x}}(\mathcal{S}) - \mathcal{A}_{\mathbf{x}}(\mathcal{S}') \\ \mathcal{A}_{\mathbf{y}}(\mathcal{S}) - \mathcal{A}_{\mathbf{y}}(\mathcal{S}') \end{pmatrix} \right\| \middle| \delta_{t_0} = 0 \right] \\ &\leq \frac{2G}{n} \sum_{k=t_0}^{T-1} \eta_k^{\max} \prod_{s=k+1}^{T-1} (1 + \eta_s^{\min} L) + 4GL \sum_{k=t_0}^{T-1} \left( \eta_k^{\max} \sum_{s=0}^{k-1} \eta_s^{\max} \lambda^{k-1-s} \right) \prod_{j=k+1}^{T-1} (1 + \eta_j^{\min} L) \end{aligned} \quad (16)$$

where we use the fact that  $\delta_{t_0} = \left\| \begin{pmatrix} \mathbf{x}^t - \hat{\mathbf{x}}^t \\ \mathbf{y}^t - \hat{\mathbf{y}}^t \end{pmatrix} \right\| = 0$ .

When learning rates are fixed that  $\eta_{\mathbf{x},t} = \eta_{\mathbf{x}}$  and  $\eta_{\mathbf{y},t} = \eta_{\mathbf{y}}$ , then we have:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} \left[ \delta_t \middle| \delta_{t_0} = 0 \right] \\ &\leq \frac{2G\eta^{\max} ((1 + \eta^{\min} L)^{T-t_0} - 1)}{n\eta^{\min} L} + \frac{4GL(\eta^{\max})^2 ((1 + \eta^{\min} L)^{T-t_0} - 1)}{(1 - \lambda)\eta^{\min} L} \end{aligned}$$

Then combining with Lemma 5, we have the following result:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} [f(\mathbf{x}^t, \mathbf{y}'; \xi) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \xi) + f(\mathbf{x}', \mathbf{y}^t; \xi) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \xi)] \\ &\leq \sqrt{2}G \left( \frac{2G\eta^{\max}}{n} + \frac{4GL\eta^{\max 2}}{1 - \lambda} \right) (T - t_0) + \frac{Bmt_0}{n} \end{aligned}$$

where it can obtain the optimal of  $2\sqrt{2}G^2 \left( \frac{\eta^{\max T}}{n} + \frac{2L\eta^{\max 2 T}}{1 - \lambda} \right)$  when  $t_0 = 0$ .

When learning rates are varying that  $\eta_t^{min} = \frac{1}{t+1}$  and  $\eta_t^{max} = \frac{1}{(t+1)^c}$ ,  $c \leq 1$ , then we can simplify the production by  $1 \pm a \leq \exp\{a\}$ :

$$\begin{aligned} \prod_{j=k+1}^{T-1} (1 + \eta_j^{min} L) &= \prod_{j=k+1}^{T-1} \left(1 + \frac{L}{j+1}\right) \leq \prod_{j=k+1}^{T-1} \exp\left\{\frac{L}{j+1}\right\} = \exp\left\{\sum_{j=k+1}^{T-1} \frac{L}{j+1}\right\} \\ &\leq \exp\left\{L \ln \frac{T}{k+1}\right\} = \left(\frac{T}{k+1}\right)^L \end{aligned}$$

Then back to the inequality (16), we can obtain:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} \left[ \delta_t \mid \delta_{t_0} = 0 \right] \\ &\leq \frac{2G}{n} \sum_{k=t_0}^{T-1} \frac{1}{(k+1)^c} \left(\frac{T}{k+1}\right)^L + 4GL \sum_{k=t_0}^{T-1} \frac{1}{(k+1)^c} \frac{C_\lambda}{k^c} \left(\frac{T}{k+1}\right)^L \\ &= \frac{2GT^L}{n} \sum_{k=t_0}^{T-1} \frac{1}{(k+1)^{c+L}} + 4GLC_\lambda T^L \sum_{k=t_0}^{T-1} \frac{1}{(k+1)^{c+L} k^c} \end{aligned} \quad (17)$$

Requiring  $c+L > 1$  for convergence and combining with Lemma 5, we have the following inequality:

$$\begin{aligned} &\mathbb{E}_{\mathcal{A}} [f(\mathbf{x}^t, \mathbf{y}'; \boldsymbol{\xi}) - f(\hat{\mathbf{x}}^t, \mathbf{y}'; \boldsymbol{\xi}) + f(\mathbf{x}', \mathbf{y}^t; \boldsymbol{\xi}) - f(\mathbf{x}', \hat{\mathbf{y}}^t; \boldsymbol{\xi})] \\ &\leq \sqrt{2G} \left( \frac{2GT^L}{n} \sum_{k=t_0}^{T-1} \frac{1}{(k+1)^{c+L}} + 4GLC_\lambda T^L \sum_{k=t_0}^{T-1} \frac{1}{(k+1)^{c+L} k^c} \right) + \frac{Bmt_0}{n} \\ &\leq \frac{2\sqrt{2}G^2 T^L}{(c+L-1)n} \frac{1}{t_0^{c+L-1}} + \frac{4\sqrt{2}G^2 LC_\lambda T^L}{(2c+L-1)} \frac{1}{t_0^{2c+L-1}} + \frac{Bmt_0}{n} \\ &\leq (c+L)(c+L-1)^{\frac{1}{c+L}} \left( \frac{2\sqrt{2}G^2 T^L}{(c+L-1)n} + \frac{4\sqrt{2}G^2 LC_\lambda T^L}{2c+L-1} \right)^{\frac{1}{c+L}} \left( \frac{Bm}{n} \right)^{1-\frac{1}{c+L}} \end{aligned}$$

where for the last inequality, when  $t_0 = \left( \frac{(c+L-1) \left( \frac{2\sqrt{2}G^2 T^L}{(c+L-1)n} + \frac{4\sqrt{2}G^2 LC_\lambda T^L}{2c+L-1} \right)}{\frac{Bm}{n}} \right)^{\frac{1}{c+L}}$ , it can obtain minimal.

Taking supremum over parameters  $\mathbf{x}, \mathbf{y}$  separately and over the random sample, we can get the weak stability for D-SGDA in the NC-NC condition.  $\square$

## H Primal Metric

In addition to the primal-dual population risk (see Def. 2) as well as the corresponding generalization gap (see Def. 3), primal measure extends the relative concept in single variable minimization problem.

The excess primal population risk is defined as  $\Delta^{ex}(\mathbf{x}) = \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}', \mathbf{y}')$ ; and the excess primal empirical risk is  $\Delta_S^{ex}(\mathbf{x}) = \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}', \mathbf{y}')$ , for a randomized model  $(\mathbf{x}, \mathbf{y})$ .

There are two ways to define the corresponding generalization gap. One is to directly subtract the excess primal empirical risk from the excess primal population risk, defined as the excess primal generalization gap:  $\epsilon_{gen}^{ex}(\mathbf{x}) = \Delta^{ex}(\mathbf{x}) - \Delta_S^{ex}(\mathbf{x})$  (called primal gap in [30]). Another one is to neglect the difference between the saddle point of  $F$  and  $F_S$ , defined as the primal generalization gap:  $\epsilon_{gen}^{pr}(\mathbf{x}) = \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}') - \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}, \mathbf{y}')$ .

**Remark 11.** Our definition of strong primal-dual population risk (see Def. 2) has included the expectation inside for assistance with the definition of weak primal-dual population risk (see Def. 2). Although the definitions in [19] hold different forms, which do not contain the expectation inside, while our theoretical result aims at the same value.

Ozdaglar et al. [30] points out that the excess primal generalization gap  $\epsilon_{gen}^{ex}$  can act as a better metric to characterize the generalizability in the nonconvex condition. It is our limitation that we do not

calculate the corresponding excess primal generalization gap and population risk for nonconvex case in our paper.

While we omit the primal generalization gap  $\epsilon_{gen}^{pr}$  and excess primal population risk  $\Delta^{ex}$  under C-C condition in the main text, for it can be derived from the corresponding proof of strong primal-dual risk. And we will illustrate them in the following as a corollary.

**Corollary 2.** *For an  $\epsilon$ -argument stable decentralized algorithm  $\mathcal{A}$ , under Assumption 1, 2, when each  $f_i$  is  $\mu_{\mathbf{y}}$ -strongly concave on the second parameter, we have the primal generalization gap:*

$$\mathbb{E}_{\mathcal{A}, \mathcal{S}}[\epsilon_{gen}^{pr}(\mathcal{A}_{\mathbf{x}}(\mathcal{S}))] \leq G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{y}}^2} \epsilon}.$$

*Proof of Corollary 2.*  $\epsilon_{gen}^{pr}(\mathcal{A}_{\mathbf{x}}(\mathcal{S}))$  is exactly the first counterpart of the strong primal-dual generalization gap  $\epsilon_{gen}^s(\mathcal{A}_{\mathbf{x}}(\mathcal{S}), \mathcal{A}_{\mathbf{y}}(\mathcal{S}))$  in Eq. (6). And referring to the proof for case **b.** in Thm. 1 (see Appendix D), we can get the result as above.  $\square$

**Corollary 3.** *Under Assumption 1,2,3, when each  $f_i$  is  $\mu_{\mathbf{x}}SC$ - $\mu_{\mathbf{y}}SC$ , we have the excess primal population risk as follows, where  $\eta_t^{max} \triangleq \max\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ ,  $\eta_t^{min} \triangleq \min\{\eta_{\mathbf{x},t}, \eta_{\mathbf{y},t}\}$ ,  $\mu = \min\{\mu_{\mathbf{x}}, \mu_{\mathbf{y}}\}$ , and  $(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$  is defined in Eq. (3):*

a. *for fixed learning rates,*

$$\begin{aligned} \mathbb{E}[\Delta^{pr}(\mathbf{x}_{ave}^T)] &\leq 2G \sqrt{1 + \frac{L^2}{\mu^2} \left( 2G \frac{L + \mu}{\eta^{min} L \mu} \left( \frac{2(\eta^{max})^2 L}{1 - \lambda} + \frac{\eta^{max}}{n} \right) + \frac{C_{\mathbf{x}}^2 + C_{\mathbf{y}}^2}{2\eta^{min} T} \right)} \\ &\quad + \eta^{max} G^2 + \frac{4(C_{\mathbf{x}} + C_{\mathbf{y}})GL\eta^{max}}{1 - \lambda} + \frac{2(C_{\mathbf{x}} + C_{\mathbf{y}})G}{\sqrt{T}}. \end{aligned}$$

b. *for decaying learning rates that  $\eta_t^{min} = \frac{1}{\mu(t+1)}$  and  $\eta_t^{max} = \frac{1}{\mu(t+1)^c}$  with  $c \leq 1$  and  $2c \geq \frac{L}{L+\mu} + 1$ ,*

$$\begin{aligned} &\mathbb{E}[\Delta^{pr}(\mathbf{x}_{ave}^T)] \\ &\leq 2G \sqrt{1 + \frac{L^2}{\mu^2} \left( \frac{2G}{\mu(1-c + \frac{L}{L+\mu})} \frac{T^{1-c}}{n} + \frac{4GLC_{\lambda}}{\mu^2 T^{\frac{L}{L+\mu}}} \left( \frac{\mathbf{1}_{2c \neq L/(L+\mu)+1}}{2c - \frac{L}{L+\mu} - 1} + \ln T \cdot \mathbf{1}_{2c=L/(L+\mu)+1} \right) \right)} \\ &\quad + \frac{2G(C_{\mathbf{x}} + C_{\mathbf{y}})}{\sqrt{T}} + \frac{G^2(1 + \ln T)}{2\mu T} + \frac{\mathbf{1}_{c \neq 1}}{(1-c)T^c} + \frac{(1 + \ln T)\mathbf{1}_{c=1}}{T} + \frac{4GLC_{\lambda}(C_{\mathbf{x}} + C_{\mathbf{y}})}{\mu T^c} \left( \frac{\mathbf{1}_{c \neq 1}}{1-c} + \ln T \cdot \mathbf{1}_{c=1} \right). \end{aligned}$$

*Proof of Corollary 3.* Firstly we already know the argument stability bound for D-SGDA (denoted as  $\mathcal{A}$ ) on the last iterate. Then the averaged output follows due to the convexity of the norm:

$$\mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}_{ave}^T - \hat{\mathbf{x}}_{ave}^T \\ \mathbf{y}_{ave}^T - \hat{\mathbf{y}}_{ave}^T \end{pmatrix} \right\| \leq \mathbb{E}_{\mathcal{A}} \left\| \begin{pmatrix} \mathbf{x}^T - \hat{\mathbf{x}}^T \\ \mathbf{y}^T - \hat{\mathbf{y}}^T \end{pmatrix} \right\| \leq \epsilon_{sta}^{arg}(\mathcal{A})$$

Then we have the primal generalization gap according to Corollary 2 that  $\mathbb{E}_{\mathcal{A}, \mathcal{S}}[\epsilon_{gen}^{pr}(\mathbf{x}_{ave}^T)] \leq G \sqrt{1 + \frac{L^2}{\mu_{\mathbf{y}}^2} \epsilon_{sta}^{arg}(\mathcal{A})}$ .

We decompose the primal population risk for the averaged output  $(\mathbf{x}_{ave}^T, \mathbf{y}_{ave}^T)$  as follows:

$$\begin{aligned} &\sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}_{ave}^T, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}', \mathbf{y}') \\ &= \left( \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}_{ave}^T, \mathbf{y}') - \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}_{ave}^T, \mathbf{y}') \right) + \left( \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}_{ave}^T, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathbf{y}_{ave}^T) \right) \\ &\quad + \left( \inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathbf{y}_{ave}^T) - \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathbf{y}_{ave}^T) \right) + \left( \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathbf{y}_{ave}^T) - \inf_{\mathbf{x}' \in \mathcal{X}} \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}', \mathbf{y}') \right) \\ &\leq \left( \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}_{ave}^T, \mathbf{y}') - \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}_{ave}^T, \mathbf{y}') \right) + \left( \sup_{\mathbf{y}' \in \mathcal{Y}} F_S(\mathbf{x}_{ave}^T, \mathbf{y}') - \inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathbf{y}_{ave}^T) \right) \\ &\quad + \left( \inf_{\mathbf{x}' \in \mathcal{X}} F_S(\mathbf{x}', \mathbf{y}_{ave}^T) - \inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathbf{y}_{ave}^T) \right) \end{aligned}$$

where the inequality is due to:  $\inf_{\mathbf{x}' \in \mathcal{X}} F(\mathbf{x}', \mathcal{A}_{\mathbf{y}}(\mathcal{S})) \leq \inf_{\mathbf{x}' \in \mathcal{X}} \sup_{\mathbf{y}' \in \mathcal{Y}} F(\mathbf{x}', \mathbf{y}')$ .

The first term is the primal generalization gap. And notice that the third term is analogous to the contrast side of the primal generalization gap. So both of them can be bounded by  $\epsilon_{gen}^{pr}(\mathbf{x}_{ave}^T)$ .

While the second term is the strong primal-dual empirical risk referring to the proof of Thm. 7 (see Appendix E.2).

Overall we can get the excess primal population risk, almost the same with the strong primal-dual population risk for SC-SC case (see Appendix E.3) except for a  $\sqrt{2}$ -times factor in the argument stability error. And our bound analysis for the excess primal population risk is consistent with strong primal-dual population risk (see Remark 6 below Thm. 3).

□

## I Additional Experiments

In this paper, we include two experiments including solving the AUC problem on svmguide and w5a by the decentralized SGD to verify the conclusions for the Convex-Concave case and the generative adversarial network training for the Nonconvex-Nonconcave Case on MNIST.

### I.1 General setup.

Different from the stability and generalization analysis of the way in [19], we need to deal with learning in a decentralized manner. In our experiments, we denote the total number of clients as  $N_c$  and  $\mathcal{S} = \{S_1, S_2, \dots, S_{N_c}\}$  as the set of samples, where  $S_i$  represents the observations stored in the  $i$ -th client. And we let  $N_{S_i}$  denote the size of  $S_i$ . We follow the same experimental setting as outlined in [13, 19] to build a neighboring/perturbing dataset  $\mathcal{S}'$ , which is constructed by individually changing one observation on each node. That is to say, for each  $S_i$ ,  $S'_i$  is constructed by randomly changing one element in  $S_i$ . Then, we deploy the totally same sub-model on each client and initialize them to the same starting point. Then, each sub-model is trained on its local data. After each iteration, each sub-model is communicated with some other clients as per a predefined communication topology. To evaluate the distance between two models trained on  $\mathcal{S}$  and  $\mathcal{S}'$ , after finishing training, we obtain an ensemble model by averaging all sub-models collected from all clients.

### I.2 Detailed implementations.

For the AUC problem, we get two model sequences  $\{(w, v)\}$  and  $\{(w', v')\}$ . Then, we calculate the Euclidean distance  $\Delta = (\|w - w'\|_2^2 + \|v - v'\|_2^2)^{1/2}$ . For training each model on each client, the algorithm we used in our experiment is SOLAM [41], which is the SGDA designed for the minimax AUC problem. We repeat the experiments 10 times to report the average results as well as the standard deviation.

For the generative adversarial learning problem, we just take the vanilla GAN structure, of which the generator and the discriminator comprise 4 fully connected layers, respectively. The leaky ReLU is taken before the output layer. Following [19], we ignore all forms of regularization such as the weight decay or dropout, and data augmentation tricks. We take 3 different seeds and 3 different ways to construct  $\mathcal{S}'$ , which means changing different observations (total 9 runs). To evaluate the model distance, we also take the Euclidean distance between the generator and discriminator separately.

Our implementation is highly based on the two source codes<sup>23</sup>.

<sup>2</sup><https://github.com/zhenhuan-yang/minimax-stability>

<sup>3</sup><https://github.com/Raiden-Zhu/Generalization-of-DSGD>