Network Effects in Performative Prediction Games

Xiaolu Wang [1] **Chung-Yiu Yau** [1] **Hoi-To Wai** [1]

Abstract

This paper studies the multi-agent performative prediction (Multi-PP) games over multiplex networks. We consider a distributed learning setting where agents partially cooperate on an *agent network*, while during learning, the data samples drawn depend on the prediction models of the agent itself and neighboring agents on a *population network*. The dynamics of Multi-PP games is hence affected by the interplay between both networks. This paper concentrates on this Multi-PP game with the following contributions. Firstly, we analyze sufficient conditions for the existence of the performative stable equilibrium (PSE) and Nash equilibrium (NE) of the Multi-PP games. Secondly, we analyze the changes to the equilibrium induced by perturbed data distributions, and derive the closed-form solutions where the network topologies are explicit. Our results connect the existence of PSE/NE with strengths of agents' cooperation, and the changes of equilibrium solutions across agents with their node centrality, etc. Lastly, we show that a stochastic gradient descent (SGD) based distributed learning procedure finds the PSE under the said sufficient condition. Numerical illustrations on the network effects in Multi-PP games corroborate our findings.

1. Introduction

A recent trend in machine learning is to study *distributed learning* where prediction models are trained on multiple agents from local and privacy-sensitive data (Konečnỳ et al., 2016). In the general setting, the learning process on these agents are applied locally and can be aided by neighbor agents on an *agent network* $\mathcal{G}^{\mathsf{A}}$. This is motivated by the scenario when the agents wish to utilize joint experience to improve generalization performance, especially when the local data distributions are heterogeneous. Variants of this scenario have been considered, e.g., personalized learning with graph regularization (Liu et al., 2017; Vanhaesebrouck et al., 2017; Bellet et al., 2018; Nassif et al., 2020), distributed learning with consensus (Nedic & Ozdaglar, 2009; Lian et al., 2017), etc.

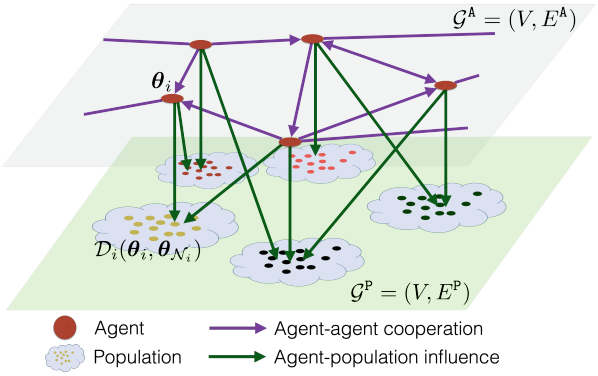

Figure 1: **Overview of the Multi-PP game on a multiplex network.** Each node consists of an agent-population pair. The agent cooperation network $\mathcal{G}^{\mathsf{A}}$ and the population influence network $\mathcal{G}^{\mathsf{P}}$ share the same set of nodes V while have different edge sets E^{A} and E^{P}, respectively. The agent i's risk depends on its own prediction model $\boldsymbol{\theta}_i$ and its incoming neighbors' models $\boldsymbol{\theta}_{\mathcal{M}_i}$ in $\mathcal{G}^{\mathsf{A}}$. The agent i's local population $\mathcal{D}_i$ depends on $\boldsymbol{\theta}_i$ and its incoming neighbors' models $\boldsymbol{\theta}_{\mathcal{N}_i}$ in $\mathcal{G}^{\mathsf{P}}$. (See Section 2 for a detailed description.)

Meanwhile, a salient challenge in optimizing prediction models deployed to real world is that the models themselves may influence the future outcomes/samples observed by the agent. These outcomes will influence the training of the future prediction models, creating a feedback dynamics between the agent who decides the model and the population who decides the outcomes/samples. This paradigm is known as the *performative prediction* as studied in Perdomo et al. (2020). An exemplar setting is when the population consists of *strategic users* who optimize their own outcomes according to a utility function parameterized by the prediction model. Applying standard algorithms such as stochastic gradient descent (SGD) results in a dynamics between the agent and the population. This has motivated recent works to analyze the existence and stability of a fixed point to the dynamics (Perdomo et al., 2020; Drusvyatskiy & Xiao, 2022),

[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. Correspondence to: Hoi-To Wai <htwai@se.cuhk.edu.hk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

the algorithms for finding the fixed point (Mendler-Dünner et al., 2020), and the optimal solution to the performative prediction formulation (Miller et al., 2021; Izzo et al., 2021).

In this paper, we concentrate on the multi-agent performative prediction (Multi-PP) setting where each agent uses samples from a (local) population for training the agent's personal prediction model. As in the (single-agent) performative prediction (Perdomo et al., 2020), each population will *react* to the prediction models deployed by the agents. Specifically, we account for a practical scenario where the effects of prediction models are *localized*, i.e., each population is influenced by a subset of the prediction models deployed by the neighbor agents on a *population network* $\mathcal{G}^{\mathtt{P}}$. For example, when the population network is induced by geographical proximity. Together with the agent network $\mathcal{G}^{\mathtt{A}}$, the two networks (which can have different topologies) are coupled to form a *multiplex network* system (De Domenico et al., 2013). Figure 1 presents an illustrative example of the Multi-PP game.

The above Multi-PP setting results in a game among agents over a multiplex network, coupled with a feedback dynamics between the agents and the populations. Under this network game-theoretic framework, this paper inquires the following questions—*How will the network structures (topologies) affect the game's equilibrium? If the data distribution at a local population/agent is perturbed, how will the perturbation affect the equilibrium solution at other agents on the network?* Addressing these questions are important steps towards understanding the role of network structures in distributed learning.

Quite recently, several related Multi-PP settings have been studied in the literature, yet they are different from the current paper. For example, Narang et al. (2022); Piliouras & Yu (2022) studied a setting for the local population whose outcomes/samples generated depend on the models deployed by *all agents* who are not directly influenced by other agents; Li et al. (2022) considered a setting where each local population is affected only by the model deployed by the local (consensus-seeking) agent. These are special cases of the multiplex network considered in this paper. More specifically, let n be the number of agents: Narang et al. (2022); Piliouras & Yu (2022) consider $\mathcal{G}^{\mathtt{A}}$ and $\mathcal{G}^{\mathtt{P}}$ as a graph with n disconnected nodes and a complete graph, respectively; Li et al. (2022) considers $\mathcal{G}^{\mathtt{A}}$ and $\mathcal{G}^{\mathtt{P}}$ as a complete graph and a graph with n disconnected nodes, respectively; Narang et al. (2022) also considers a case when both $\mathcal{G}^{\mathtt{A}}$ and $\mathcal{G}^{\mathtt{P}}$ are complete graphs.

Departing from existing works whose results are restricted to either complete graphs or disconnected graphs, the current paper conducts the first study on Multi-PP games that focuses on the effects of the *multiplex network structures* on the resulting learning dynamics. Our contributions are summarized as:

- We conduct the first study on Multi-PP games with general network structures. Our formulation is inspired by recent popular works on personalized learning, e.g., (Vanhaesebrouck et al., 2017), and accounts for the interplay between the *cooperation* among agents on $\mathcal{G}^{\mathtt{A}}$ and the *influences* of agents on local populations on $\mathcal{G}^{\mathtt{P}}$.
- We consider two concepts of equilibrium solution in Multi-PP games, namely the performative stable equilibrium (PSE) and the Nash equilibrium (NE). We first derive the *sufficient* conditions for the existence and uniqueness of these equilibrium solutions *in relation to* the multiplex network topology, and provide a SGD-based learning procedure for finding the PSE. Interestingly, when the agent cooperation network is *asymmetric*, we show that increasing the strength of agents' cooperation may destabilize the PSE solution to Multi-PP game by adopting a repeated minimization procedure.
- By specializing the Multi-PP games with simple loss functions and sample distributions, we provide exact characterizations for the PSE and NE. Our results include *necessary and sufficient conditions* for their existence and uniqueness based on a repeated minimization procedure, and the *closed form solutions* of the equilibriums featuring explicit dependence on the network structure. We observe that the stability of Multi-PP games can have non-monotonic dependence on the strength of the agents' cooperation involving the network structures, and the form of equilibrium solutions is related to node centrality of the multiplex network.

Lastly, our result reveals the intricate effects of enabling cooperation among agents while accounting for the reaction of population to models deployed on a Multi-PP game over a multiplex network. This gives a new perspective to the study of distributed learning.

Related Works: This paper is related to works on network games (Galeotti et al., 2010; Parise & Ozdaglar, 2019). Particularly, characterizing the equilibrium in network games is important to consumer networks (Candogan et al., 2012), financial networks (Acemoglu et al., 2015), and interventions in economic networks (Bramoullé et al., 2014), etc., and distributed algorithms for reaching these equilibrium have been studied (Mazumdar et al., 2020). Besides, the performative prediction setup includes strategic classification as a special case, which can be studied through the Stackelberg game framework (Hardt et al., 2016; Zrnic et al., 2021).

In this light, our results can be regarded as an extension of the above works to multiplex games. Note that most prior results in the multiplex setting are empirical (Allen et al., 2018; De Domenico et al., 2013; Gómez-Gardenes et al., 2012), with focus on social and economical networks.

2. Problem Setup

This section describes the n-agent Multi-PP games over multiplex networks. The agent network $\mathcal{G}^{\mathtt{A}}$ (resp. the population network $\mathcal{G}^{\mathtt{P}}$) is a directed graph represented by the adjacency matrix $\boldsymbol{A} \in \mathbb{R}_+^{n\times n}$ (resp. $\boldsymbol{P} \in \mathbb{R}_+^{n\times n}$). We follow the convention that if $A_{ij} > 0$ (resp. $P_{ij} > 0$), then there is an edge from j to i in $\mathcal{G}^{\mathtt{A}}$ (resp. $\mathcal{G}^{\mathtt{P}}$). Intuitively, A_{ij} (resp. P_{ij}) represents the influence from node j in the agent (resp. population) network on node i. Define $\mathcal{M}_i := \{j : j \neq i \text{ and } A_{ij} > 0\}$ (resp. $\mathcal{N}_i := \{j : j \neq i \text{ and } P_{ij} > 0\}$) as the set of incoming neighbors of node i in $\mathcal{G}^{\mathtt{A}}$ (resp. $\mathcal{G}^{\mathtt{P}}$).

Let $p_1, \ldots, p_n \in \mathbb{Z}_{++}$ be the dimensions of the prediction models and $p = \sum_{i=1}^n p_i$. We also define $\boldsymbol{\theta}_{\mathcal{M}_i} := [\boldsymbol{\theta}_j]_{j\in\mathcal{M}_i}$ and $\boldsymbol{\theta}_{\mathcal{N}_i} := [\boldsymbol{\theta}_j]_{j\in\mathcal{N}_i}$. We consider a Multi-PP game where agent i seeks to minimize a local risk function:

$$\begin{aligned}\min_{\boldsymbol{\theta}_i\in\mathbb{R}^{p_i}} & F_i(\boldsymbol{\theta}_i, [\boldsymbol{\theta}_j]_{j\in\mathcal{M}_i\cup\mathcal{N}_i}) \\ & := \mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}_{\mathcal{N}_i})}\left[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i)\right],\end{aligned} \tag{1}$$

where $[\boldsymbol{\theta}_j]_{j\in\mathcal{M}_i\cup\mathcal{N}_i}$ is given. The local performative risk function F_i depends on the joint prediction model $\boldsymbol{\theta} := [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n] \in \mathbb{R}^p$ in two ways. First, $f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i)$ is the sampled risk function that evaluates the prediction models deployed by agent i and its incoming neighbors in $\mathcal{G}^{\mathtt{A}}$, i.e., $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}$, with respect to (w.r.t.) a sample $\boldsymbol{Z}_i$ on some metric space $\mathcal{Z}_i$. Second, the distribution of $\boldsymbol{Z}_i$ is given by the *distribution mapping* $\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})$ that is a probability measure on $\mathcal{Z}_i$ and it encodes how the samples generated from the i-th population reacts to the models deployed by agent i and its incoming neighbors in $\mathcal{G}^{\mathtt{P}}$. Observing that the risk function F_i depends on the two graphs $\mathcal{G}^{\mathtt{A}}, \mathcal{G}^{\mathtt{P}}$, Problem (1) yields a multiplex network game (Gómez-Gardenes et al., 2012; Allen et al., 2018).

In (1), each agent aims to minimize its individual *performative risk* $F_i(\boldsymbol{\theta}_i, [\boldsymbol{\theta}_j]_{j\in\mathcal{M}_i\cup\mathcal{N}_i})$ that depends on the joint prediction model $[\boldsymbol{\theta}_j]_{j\in\{i\}\cup\mathcal{M}_i\cup\mathcal{N}_i}$. Inspired by the studies on personalized learning (Vanhaesebrouck et al., 2017; Bellet et al., 2018), we consider risk function of the following form:

$$f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i) := \ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) + \frac{\rho_i}{2}\sum_{j\in\mathcal{M}_i} A_{ij}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2, \tag{2}$$

where $\ell_i : \mathbb{R}^{p_i} \times \mathcal{Z}_i \mapsto \mathbb{R}$ is the loss function that depends only on the local model $\boldsymbol{\theta}_i$ and the second term is the so called *graph regularization* term with $\rho_i \in [0, \infty)$ that promotes similarity between $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, $j \in \mathcal{M}_i$. If $\mathcal{G}^{\mathtt{A}}$ is a connected graph, then the equilibrium solution of minimizing (2) with $\rho_i \to 0$ provides purely local prediction models while $\rho_i \to \infty$ provides the common (i.e., consensual) prediction model. Note that in personalized learning, it is common for $\mathcal{M}_i$ to include agents with similar target models as agent i (e.g., the agents are close in terms of their geographical locations) to maximize performance.

In the sequel, we set without loss of generality that $\boldsymbol{A}$ is row stochastic with zero diagonal, i.e., $A_{ii} = 0$ and $\sum_{j=1}^n A_{ij} = 1$ for $i \in [n] := \{1, \ldots, n\}$.[1] We also make the following assumptions that are common in the performative prediction literature, e.g., Perdomo et al. (2020); Mendler-Dünner et al. (2020); Drusvyatskiy & Xiao (2022); Li et al. (2022); Narang et al. (2022)

Assumption 2.1. For each $i \in [n]$, the following hold:

i) There exist a constant $\mu_i \geq 0$, such that for any given $\boldsymbol{Z}_i \in \mathcal{Z}_i$, $\ell_i(\cdot; \boldsymbol{Z}_i)$ is μ_i-strongly convex;
ii) $\ell_i(\cdot; \boldsymbol{Z}_i)$ is C^1-smooth for any $\boldsymbol{Z}_i \in \mathcal{Z}_i$ and there exists a constant $L_i > 0$ such that for any $\boldsymbol{\theta}_i, \boldsymbol{\theta}_i' \in \mathbb{R}^{p_i}$ and $\boldsymbol{Z}_i, \boldsymbol{Z}_i' \in \mathcal{Z}_i$, it holds that

$$\begin{aligned}&\|\nabla\ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) - \nabla\ell_i(\boldsymbol{\theta}_i'; \boldsymbol{Z}_i')\|_2 \\ &\leq L_i(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|_2 + \|\boldsymbol{Z}_i - \boldsymbol{Z}_i'\|_2).\end{aligned}$$

Assumption 2.1 imposes the convexity and smoothness properties on the loss function. It is worth noting that in Assumption 2.1(i), we allow $\mu_i = 0$, in which case $\ell_i(\cdot; \boldsymbol{Z}_i)$ is convex but not strongly convex. In spite of the convexity of ℓ_i, F_i can be generally non-convex in the first argument $\boldsymbol{\theta}_i$. We remark that in Assumption 2.1(ii), the Lipschitz continuity w.r.t. the first argument of $\nabla\ell_i(\cdot;\cdot)$ is not required in some of our results.

We also require the following condition on the distribution mapping. Let $W_1(\mathcal{D}, \mathcal{D}')$ denote the Wasserstein 1-distance between two probability measures $\mathcal{D}$, $\mathcal{D}'$, we impose the following Lipschitz-like property:

Assumption 2.2. For $i \in [n]$, there exists $\varepsilon_i \geq 0$ such that

$$W_1\left(\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}), \mathcal{D}_i(\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\mathcal{N}_i}')\right) \leq \varepsilon_i\sqrt{\textstyle\sum_{j=1}^n P_{ij}\left\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j'\right\|_2^2}$$

for all $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}), (\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\mathcal{N}_i}') \in \mathbb{R}^{p_i} \times \mathbb{R}^{\sum_{j\in\mathcal{N}_i} p_j}$.

The parameter ε_i is the bound on the *sensitivity* of the i-th population w.r.t. the shift of the deployed prediction models. Note that for each $i \in [n]$, P_{ij} can be viewed as the weight on the shift of the j-th prediction model. Similar to $\boldsymbol{A}$ in (2), some forms of normalization are often imposed to avoid ambiguity with the parameter ε_i. Among others, we may consider the special case when $P_{ij} \in \{0, 1\}$ and $P_{ii} = 1$ for all $i, j \in [n]$, i.e., the model shifts in the neighbour nodes are assigned with equal weights, then Assumption 2.2 is in line with the other multi-agent performative prediction literature, e.g., (Narang et al., 2022).

[1] Note that even if $\sum_{j=1}^n A_{ij} = b_i \neq 1$, we can simply scale ρ_i by b_i and A_{ij} by $1/b_i$, and obtain an equivalent formulation with normalized weights, i.e., $\sum_{j=1}^n (A_{ij}/b_i) = 1$.

We conclude this section by describing a motivating example for the Multi-PP game under the above settings.

Example 2.3. Consider a multi-agent classification game, where there are n agents that represent banks aiming to predict whether the loan applicants are creditworthy. Each bank trains a personalized logistic regression model according to (2) with the loss function

$$\ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) = -y_i \boldsymbol{\theta}_i^\top \boldsymbol{x}_i + \log\left(1 + e^{\boldsymbol{\theta}_i^\top \boldsymbol{x}_i}\right), \tag{3}$$

where $\boldsymbol{Z}_i = (\boldsymbol{x}_i, y_i) \in \mathbb{R}^{p_i} \times \{0, 1\}$ is the feature-label pair of the i-th applicant. On the one hand, banks located in regions near to each other tend to deploy similar loan policies, while banks that are far away from each other tend to make independent decisions. This gives rises to a inter-bank cooperation network $\mathcal{G}^{\mathsf{A}}$ formed according to their geographical distances, and thus the risk function given by (2). On the other hand, each applicant in some region may be affected by neighboring banks' policies and strategically manipulate their features to increase the chances of successfully applying for the loan. Specifically, suppose that $\overline{\boldsymbol{Z}}_i = (\bar{\boldsymbol{x}}_i, \bar{y}_i) \in \mathbb{R}^{p_i} \times \{0, 1\}$ follow some base distribution $\overline{\mathcal{D}}_i$, then each data sample $\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})$ is generated by perturbing $\overline{\mathcal{D}}_i$ as follows:

$$\boldsymbol{x}_i = \begin{cases} \bar{\boldsymbol{x}}_i + \bar{\varepsilon}_i \sum_{j=1}^n P_{ij} \boldsymbol{\theta}_j, & \text{if } \bar{y}_i = 0, \\ \bar{\boldsymbol{x}}_i, & \text{if } \bar{y}_i = 1, \end{cases} \tag{4}$$

$$y_i = \bar{y}_i, \tag{5}$$

where $\bar{\varepsilon}_i \in \mathbb{R}$. Thus, the applicant population network arises from the bank-applicant influences. Note that when $n = 1$, this setting reduces to the strategic classification problem studied in Hardt et al. (2016); Dong et al. (2018); Perdomo et al. (2020); Zrnic et al. (2021). It is worth mentioning that this example satisfies Assumptions 2.1 and 2.2; see Appendix A for detailed verification.

3. Equilibrium(s) of the Multi-PP Game

This section presents the main results on the equilibrium(s) of the game (1) resulted from the cooperation/competition among agents and the agent-population pairs. Compared to prior works on Multi-PP (Narang et al., 2022; Piliouras & Yu, 2022; Li et al., 2022), we notice that (1) depends on the graph structure of $\mathcal{G}^{\mathsf{A}}$ and $\mathcal{G}^{\mathsf{P}}$ where the interactions between agents occur simultaneously on both graphs.

The first focus of our study is on the *existence* of equilibrium(s) to (1). Our results shall highlight the contributions of graph structure to the existence condition of equilibrium(s). When the latter condition holds, we also suggest a stochastic gradient based procedure to finding an equilibrium solution.

Similar to Narang et al. (2022), we concentrate on two concepts of equilibrium solution for (1) below.

Definition 3.1. A vector $\boldsymbol{\theta}^{\mathrm{pse}} = [\boldsymbol{\theta}_1^{\mathrm{pse}}; \ldots; \boldsymbol{\theta}_n^{\mathrm{pse}}] \in \mathbb{R}^p$ is called a performative stable equilibrium (PSE) of the game (1) if it holds for all $i \in [n]$ that

$$\boldsymbol{\theta}_i^{\mathrm{pse}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}_i \in \mathbb{R}^{p_i}} \left\{ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i^{\mathrm{pse}}, \boldsymbol{\theta}_{\mathcal{N}_i}^{\mathrm{pse}})} \left[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}^{\mathrm{pse}}; \boldsymbol{Z}_i) \right] \right\}.$$

Definition 3.2. A vector $\boldsymbol{\theta}^{\mathrm{ne}} = [\boldsymbol{\theta}_1^{\mathrm{ne}}; \ldots; \boldsymbol{\theta}_n^{\mathrm{ne}}] \in \mathbb{R}^p$ is called a Nash equilibrium (NE) of the game (1) if it holds for all $i \in [n]$ that

$$\boldsymbol{\theta}_i^{\mathrm{ne}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}_i \in \mathbb{R}^{p_i}} \left\{ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}^{\mathrm{ne}})} \left[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}^{\mathrm{ne}}; \boldsymbol{Z}_i) \right] \right\}.$$

A subtle yet important difference between $\boldsymbol{\theta}^{\mathrm{pse}}$ and $\boldsymbol{\theta}^{\mathrm{ne}}$ lies in the observation that $\boldsymbol{\theta}_i^{\mathrm{ne}}$ is a global minimizer of the performative risk $F_i(\boldsymbol{\theta}_i, [\boldsymbol{\theta}_j^{\mathrm{ne}}]_{j \in \mathcal{M}_i \cup \mathcal{N}_i})$ that jointly optimizes the sampled risk and the decision-dependent distribution.

3.1. Perfomative Stable Equilibrium (PSE)

We first derive a sufficient condition for the existence of PSE by construction. In particular, to facilitate our analysis, we consider the strategy of repeated risk minimization (RRM), which is an iterative mechanism such that each agent repeatedly minimizes its own expected risk while fixing all other agents' decisions and the induced data distribution. In iteration t, agent i does

$$\begin{aligned} \boldsymbol{\theta}_i^{t+1} &= \mathcal{T}_i\left([\boldsymbol{\theta}_j^t]_{j \in \{i\} \cup \mathcal{M}_i \cup \mathcal{N}_i} \right) \\ &:= \operatorname*{arg\,min}_{\boldsymbol{\theta}_i \in \mathbb{R}^{p_i}} \left\{ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_{\mathcal{N}_i}^t)} \left[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}^t; \boldsymbol{Z}_i) \right] \right\} \end{aligned} \tag{6}$$

and compete with other agents to minimize the local risk upon deploying the current prediction model. From an optimization perspective, the RRM (6) is a Jacobi-type *coordinate descent* method. With the sampled risk function $f_i(\cdot)$ defined in (2), the mechanism is similar to the coordinate descent method in Bellet et al. (2018) when the data distribution $\mathcal{D}_i$ is independent of $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}$.

Observe that the fixed point(s) of (6) leads to a PSE solution. Hence, studying the convergence of (6) results in the following sufficient condition that implies the existence of PSE, whose proof can be found in Appendix C:

Theorem 3.3. *Suppose that* $\sum_{j=1}^n A_{ij} = 1$ *and* $\mu_i + \rho_i > 0$ *for all* $i \in [n]$*, Assumptions 2.1 and 2.2 hold. Let* $\boldsymbol{\mu} := [\mu_i]_{i=1}^n$ *and* $\boldsymbol{\rho} := [\rho_i]_{i=1}^n$*. Under the condition*

$$\sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n \frac{P_{ij} L_i^2 \varepsilon_i^2}{(\mu_i + \rho_i)^2} \right\}} + \left\| \mathrm{Diag}\left(\frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \right) \boldsymbol{A} \right\|_2 < 1, \tag{7}$$

where $\boldsymbol{\rho}/(\boldsymbol{\mu} + \boldsymbol{\rho})$ *stands for an element-wise division, then we have:* (i) *the game* (1) *admits a unique PSE, and* (ii) *the RRM* (6) *converges linearly to the PSE.*

In the above theorem, (7) gives a sufficient condition for the existence of PSE in terms of the distributional sensitivity w.r.t. the prediction models ($\{\varepsilon_i\}_{i=1}^n$). This is a general result with loss functions satisfying Assumption 2.1 and decision-dependent distributions satisfying Assumption 2.2. Moreover, the condition is directly determined by properties of the graphs $\mathcal{G}^{\mathtt{A}}$ and $\mathcal{G}^{\mathtt{P}}$.

Condition (7) subsumes several results in the literature as special cases. For instance, when $n = 1$, $\rho_1 = 0$, and $\mu_1 > 0$, then (7) reduces to the condition $\varepsilon_1 < \mu_1/L_1$, which coincides with the result for single-agent performative prediction (Perdomo et al., 2020, Theorem 3.5). For the special case with a fully connected population network in the absence of the graph regularization in (2), i.e., $\boldsymbol{P} = \mathbf{1}_n\mathbf{1}_n^\top$, and $\rho_i = 0$, $\mu_i > 0$ for all $i \in [n]$,[2] then condition (7) can be simplified to

$$\textstyle\sum_{i=1}^n L_i^2\varepsilon_i^2/\mu_i^2 < 1, \tag{8}$$

which coincides with Narang et al. (2022, Theorem 2). As a special case, if $\varepsilon_i = \varepsilon$, $L_i = L$, $\mu_i = \mu$, this condition reduces to $\varepsilon < \mu/(\sqrt{n}L)$, showing that the sensitivity requirement becomes more stringent by a factor of $\sqrt{n}$.

Next, we consider the effects of incorporating *cooperation* in (2), i.e., when $\rho_i > 0$. We highlight that the situation under consideration differs from that in Li et al. (2022) which enforces *exact consensus* in their algorithm and shows a relaxed condition for the existence of performative stable solution akin to Perdomo et al. (2020). Here, we do not enforce *exact consensus* where $\rho_i < \infty$ and the equilibrium results from the competition among agents.

To analyze this situation using Theorem 3.3, we shall adopt some simplifications by setting $\mu_i = \mu$ and $\rho_i = \rho$ for all $i \in [n]$. In this case, (7) can be implied by

$$\max_{j\in[n]} \sqrt{\textstyle\sum_{i=1}^n P_{ij}L_i^2\varepsilon_i^2} < \mu - \rho(\|\boldsymbol{A}\|_2 - 1), \tag{9}$$

where the above condition can be further simplified into $L\sqrt{\|\boldsymbol{P}\|_\infty}\max_{i\in[n]}\varepsilon_i < \mu - \rho(\|\boldsymbol{A}\|_2 - 1)$ if $L_i = L$, such that $\|\boldsymbol{P}\|_\infty = \max_{j\in[n]}\sum_{i=1}^n P_{ij}$ corresponds to the maximum weighted out-degree of the graph $\mathcal{G}^{\mathtt{P}}$.

The left hand side of (9) extends from the previously discussed case in (8) by incorporating the population network $\mathcal{G}^{\mathtt{P}}$. In fact, it relaxes the sensitivity requirement factor from $\sqrt{n}$ to $\sqrt{\|\boldsymbol{P}\|_\infty}$, showing that a more *localized* population network with less edges can be beneficial.

Meanwhile, the right hand side of (9) reveals an intriguing property on the role of cooperation in Multi-PP (1) and agent network $\mathcal{G}^{\mathtt{A}}$. Notice that if $\boldsymbol{A}$ is further assumed to be symmetric, i.e., $\boldsymbol{A}$ is a doubly stochastic matrix, one has $\|\boldsymbol{A}\|_2 = 1$.[3] As a consequence, the condition (9) is independent of the parameter ρ. However, in general, one has $\|\boldsymbol{A}\|_2 \geq 1$. In the latter case, increasing ρ may lead to violation of (7) and destabilize the RRM dynamics. We observe that while increasing ρ promotes cooperation among agents and thus better generalization performance of the trained model, in the presence of performativity in population, it may lead to undesirable outcomes in the training procedure.

[2] In this setting, $\boldsymbol{A}$ can be arbitrary since $\rho_i = 0$ for all $i \in [n]$.

Algorithm 1 Stochastic Gradient Greedy Deployment

1: **Input:** $\boldsymbol{\theta}_i^0$ for $i \in [n]$, step size $\gamma_t > 0$ for $t \geq 1$.
2: **for** $t = 0, 1, \ldots$ **do**
3: Deploy the models $\{\boldsymbol{\theta}_i^t\}_{i=1}^n$ at the population.
4: **for** $i = 1$ **to** n **do** {`executed in parallel`}
5: Sample $\boldsymbol{Z}_i^{t+1} \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_{\mathcal{N}_i}^t)$
6: Set $\boldsymbol{g}^t = \nabla\ell_i(\boldsymbol{\theta}_i^t; \boldsymbol{Z}_i^{t+1}) + \rho_i\sum_{j=1}^n A_{ij}\left(\boldsymbol{\theta}_i^t - \boldsymbol{\theta}_j^t\right)$
7: Set $\boldsymbol{\theta}_i^{t+1} = \boldsymbol{\theta}_i^t - \gamma_{t+1}\boldsymbol{g}^t$
8: **end for**
9: **end for**

Stochastic Algorithm for PSE. We conclude our discussions on the PSE solution to (1) through studying a stochastic gradient based algorithm. Our algorithm design follows from Mendler-Dünner et al. (2020); Drusvyatskiy & Xiao (2022) with greedy deployment, i.e., the prediction models under training are directly deployed, then a sample (batch) is drawn to compute the stochastic gradient estimate for (2) that is used to inform the update of $\boldsymbol{\theta}^t$; see Algorithm 1. Notice that except for line 5, the algorithm describes a natural implementation of personalized learning (Bellet et al., 2018) based on stochastic gradients.

To analyze the convergence of the stochastic-gradient greedy-deployment (SG-GD) mechanism described in Algorithm 1, we consider the following assumption on the variance of the stochastic gradient.

Assumption 3.4. There exists $\sigma_0, \sigma_1 \geq 0$ such that for any given $\boldsymbol{\theta} \in \mathbb{R}^p$, it holds

$$\mathbb{E}[\|\nabla\boldsymbol{\ell}(\boldsymbol{\theta}; \boldsymbol{Z}) - \mathbb{E}[\nabla\boldsymbol{\ell}(\boldsymbol{\theta}; \boldsymbol{Z})]\|_2^2] \leq \sigma_0^2 + \sigma_1^2\|\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2,$$

where $\nabla\boldsymbol{\ell}(\boldsymbol{\theta}; \boldsymbol{Z}) = [\nabla\ell_1(\boldsymbol{\theta}_1; \boldsymbol{Z}_1); \cdots; \nabla\ell_i(\boldsymbol{\theta}_n; \boldsymbol{Z}_n)]$ concatenates the local stochastic gradients, and the expectations are taken w.r.t. $\boldsymbol{Z} = (\boldsymbol{Z}_i)_{i=1}^n$ with $\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})$.

This is a standard assumption on the stochastic gradient estimates with a growth condition. The following theorem establishes the convergence of SG-GD to $\boldsymbol{\theta}^{\mathrm{pse}}$:

Theorem 3.5. *Under the same conditions in Theorem 3.3*

[3] On the one hand, $\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{v}\neq\boldsymbol{0}}\frac{\|\boldsymbol{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2} \geq \frac{\|\boldsymbol{A}\mathbf{1}_n\|_2}{\|\mathbf{1}_n\|_2} = 1$ by row-stochasticity. On the other hand, $\|\boldsymbol{A}\|_2 \leq \sqrt{\|\boldsymbol{A}\|_1\|\boldsymbol{A}\|_\infty} = 1$ due to both row-stochasticity and symmetry.

with $L_i = L, \mu_i = \mu, \varepsilon_i = \varepsilon, \rho_i = \rho$. *Let*

$$\widetilde{\mu} \coloneqq \mu + \rho(1 - \|\boldsymbol{A}\|_2) - L\varepsilon\sqrt{\|\boldsymbol{P}\|_\infty},$$
$$\widetilde{\sigma}^2 \coloneqq \sigma_1^2 + 2\left(L^2\varepsilon^2\|\boldsymbol{P}\|_\infty + (L + \rho\|\boldsymbol{I}_n - \boldsymbol{A}\|_2)^2\right).$$

Suppose that Assumption 3.4 and condition (7) hold, and the step sizes satisfy

$$\sup_{t\geq 1}\gamma_t \leq \min\left\{\frac{\widetilde{\mu}}{2\widetilde{\sigma}^2}, \frac{2}{\widetilde{\mu}}\right\} \text{ and } \frac{\gamma_t}{\gamma_{t+1}} \leq 1 + \frac{\widetilde{\mu}\gamma_{t+1}}{2} \ \forall t \geq 1.$$

Then, the iterates generated by Algorithm 1 satisfy that for all $t \geq 1$,

$$\mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2] \leq \prod_{s=0}^{t}(1-\widetilde{\mu}\gamma_{s+1})\Delta_0 + \frac{4\sigma_0^2}{\widetilde{\mu}}\gamma_{t+1}, \quad (10)$$

where $\Delta_0 \coloneqq \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2$ *is the initial gap.*

See the proof in Appendix D. The step size conditions are general, e.g., they can be satisfied with a constant step size, or with a standard diminishing rule such as $\gamma_t = \frac{a_0}{a_1+t}$ with appropriate $a_0, a_1 > 0$. In the latter case, (10) shows that the convergence behavior of Algorithm 1 *towards* $\boldsymbol{\theta}^{\mathrm{pse}}$ is similar to that of SGD applied to strongly convex objective functions, i.e., at the rate of $\mathcal{O}(1/t)$ (Moulines & Bach, 2011). However, we observe that the convergence depends on the graph regularization parameter and the sensitivity of the population to model shifts, thus SG-GD may not converge even if $\mu > 0$.

Lastly, we remark that besides including the graph regularization term, our result extends over Narang et al. (2022) by the use of diminishing step sizes; and Bellet et al. (2018) by incorporating stochastic samples and the performative effects for interaction with the population network.

3.2. Nash Equilibrium (NE)

Lastly, we discuss the existence and uniqueness of the NE for the Multi-PP game (1). We focus on the case when (1) is a C^1-smooth convex game, i.e., for all $i \in [n]$, $\nabla_i F_i(\boldsymbol{\theta})$ (the partial gradient w.r.t. $\boldsymbol{\theta}_i$) is C^1-smooth and $F_i(\boldsymbol{\theta})$ is convex w.r.t. $\boldsymbol{\theta}_i$ for any fixed $[\boldsymbol{\theta}_j]_{j\in\mathcal{M}_i\cup\mathcal{N}_i}$.

We first impose the following regularity condition:

Assumption 3.6. For each $i \in [n]$ and $\boldsymbol{\theta} \in \mathbb{R}^p$, the mapping $\mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\cdot,\boldsymbol{\theta}_{\mathcal{N}_i})}[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i)]$ is differentiable at $\boldsymbol{\theta}_i$ and its derivative is continuous in $[\boldsymbol{\theta}_i; \boldsymbol{\theta}_{\mathcal{N}_i}]$.

Under Assumption 3.6, we can define the following mapping: for any $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^p$,

$$H^i_{\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) \coloneqq \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i}\mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{u}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i)]\Big|_{\boldsymbol{u}_i=\boldsymbol{\delta}_i},$$
$$H_{\boldsymbol{\delta}}(\boldsymbol{\theta}) \coloneqq [H^i_{\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i})]_{i=1}^n.$$

Then, we consider the following assumption:

Assumption 3.7. For any $i \in [n]$ and $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^p$, the mapping $H_{\boldsymbol{\delta}}(\boldsymbol{\theta})$ is monotone w.r.t. $\boldsymbol{\delta}$, i.e.,

$$\langle H_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\delta}\rangle \geq 0.$$

Moreover, we define the *decoupled* expected gradients

$$G^i_{\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) \coloneqq \mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}[\nabla_i f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i)],$$
$$G_{\boldsymbol{\delta}}(\boldsymbol{\theta}) \coloneqq [G^i_{\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i})]_{i=1}^n$$

for $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^p$. Using the product rule of derivatives, we obtain

$$\nabla_i F_i(\boldsymbol{\theta}) = G^i_{\boldsymbol{\theta}_i,\boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) + H^i_{\boldsymbol{\theta}_i,\boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}). \quad (11)$$

To study the NE of (1), we rely on a classical result that a *strongly monotone game* over a non-empty, closed and convex set admits a unique NE; see, e.g., Facchinei & Pang (2003, Theorem 2.3.3(b)). Specifically, let $\Phi_F(\boldsymbol{\theta}) \coloneqq [\nabla_1 F_1(\boldsymbol{\theta}), \ldots, \nabla_n F_n(\boldsymbol{\theta})]$, then the Multi-PP game (1) is called strongly monotone if there exists $\beta > 0$ such that

$$\langle \Phi_F(\boldsymbol{\theta}) - \Phi_F(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta}\rangle \geq \beta\|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 \quad (12)$$

for all $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^p$. In view of (11), we have

$$\Phi_F(\boldsymbol{\theta}) = G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + H_{\boldsymbol{\theta}}(\boldsymbol{\theta}). \quad (13)$$

Below, we show that the strong monotonicity property (12) can be satisfied under appropriate conditions on the sensitivity and graph regularization parameters.

Theorem 3.8. *Suppose that* $\sum_{j=1}^n A_{ij} = 1$ *for all* $i \in [n]$ *and Assumptions 2.1, 2.2, 3.6, and 3.7 hold. Let* $\mu_{\min} \coloneqq \min_{i\in[n]}\{\mu_i\}$ *and* $\rho_{\min} \coloneqq \min_{i\in[n]}\{\rho_i\}$. *If it holds that*

$$\sqrt{\max_{j\in[n]}\left\{\sum_{i=1}^n\left(\frac{P_{ij}L_i\varepsilon_i}{\mu_{\min}+\rho_{\min}}\right)^2\right\}} + \left\|\mathrm{Diag}\left(\frac{\boldsymbol{\rho}}{\mu_{\min}+\rho_{\min}}\right)\boldsymbol{A}\right\|_2 < 1 - \frac{\max_{i\in[n]}\{L_i\varepsilon_i\}}{\mu_{\min}+\rho_{\min}}, \quad (14)$$

then the Multi-PP game (1) is strongly monotone and admits a unique NE.

The proof of the above theorem is provided in Appendix E.

It is worthwhile to note that for Multi-PP games with fully connected population network (i.e., $\boldsymbol{P} = \mathbf{1}_n\mathbf{1}_n^\top$) and non-cooperating agents (i.e., $\rho_i = 0$ for all $i \in [n]$), Theorem 3.8 yields a weaker sufficient condition than Narang et al. (2022, Theorem 5). Specifically, suppose that $\mu_i = \mu > 0$ for all $i \in [n]$, then the condition (14) can be reduced to

$$\sqrt{\textstyle\sum_{i=1}^n L_i^2\varepsilon_i^2} + \max_{i\in[n]}\{L_i\varepsilon_i\} \leq \mu. \quad (15)$$

Since $\max_{i\in[n]}\{L_i\varepsilon_i\} \leq \sqrt{\sum_{i=1}^n L_i^2\varepsilon_i^2}$, our condition is strictly weaker than the condition $2\sqrt{\sum_{i=1}^n L_i^2\varepsilon_i^2} \leq \mu$ required by Narang et al. (2022, Theorem 5).

The NE can be found with the *best response (BR) dynamics*, i.e., in iteration t, each agent i does

$$\begin{aligned}\boldsymbol{\theta}_i^{t+1} &= \mathcal{B}_i\left(\left[\boldsymbol{\theta}_j^t\right]_{j\in\mathcal{M}_i\cup\mathcal{N}_i}\right) \\ &:= \arg\min_{\boldsymbol{\theta}_i\in\mathbb{R}^{p_i}} \mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}^t_{\mathcal{N}_i})}\left[f_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}^t_{\mathcal{M}_i};\boldsymbol{Z}_i)\right].\end{aligned} \tag{16}$$

We remark that the stochastic algorithms for finding $\boldsymbol{\theta}^{\mathrm{ne}}$ can be readily extended from existing works (Narang et al., 2022; Izzo et al., 2021; Miller et al., 2021). For example, the NE can be found by adapting the algorithms suggested in Narang et al. (2022, Section 6), where the derivative free algorithm exhibits a convergence rate of $\mathcal{O}(1/\sqrt{t})$. The detailed discussion is omitted due to space limitation.

4. Case Studies & Numerical Illustrations

This section presents the case studies on Multi-PP game (1) and analyze their equilibrium solution(s). Notice that Theorems 3.3 and 3.8 only give the *sufficient* conditions for the existence and uniqueness of PSE and NE, respectively. Here, we will tighten these results to *necessary and sufficient* conditions for specific examples. We also derive the closed-form solutions of the PSE and NE, where the effects of network structure are explicit. Lastly, we provide numerical examples to illustrate our findings.

4.1. Mean Squared Error Minimization

We first consider the following special case of (1). Let the model dimension be $p_i = \bar{p} \in \mathbb{Z}_{++}$ and $\mathcal{Z}_i = \mathbb{R}^{\bar{p}}$ for all $i \in [n]$. The loss function of agent i is

$$\ell_i(\boldsymbol{\theta}_i;\boldsymbol{Z}_i) = \frac{1}{2}\|\boldsymbol{\theta}_i - \boldsymbol{Z}_i\|_2^2. \tag{17}$$

Therefore, the local risk function corresponds to the mean squared error (MSE) in estimating the mean of $\boldsymbol{Z}_i$. The graph regularization parameters takes $\rho_i = \rho$ for all $i \in [n]$ for some $\rho \geq 0$. Next, a sample $\boldsymbol{Z}_i$ drawn from the distribution $\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})$ satisfies

$$\boldsymbol{Z}_i = \bar{\boldsymbol{Z}}_i + \bar{\varepsilon}\textstyle\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j, \tag{18}$$

where $\bar{\varepsilon} \in \mathbb{R}$ is the sensitivity parameter, $\bar{\boldsymbol{Z}}_i$ is a random variable with mean $\boldsymbol{m}_i$ and covariance $\sigma\boldsymbol{I}_{\bar{p}}$. We remark that this distribution mapping is motivated by a scenario when the i-th population generates samples by maximizing a linear utility. For brevity, we denote $\boldsymbol{m} = [\boldsymbol{m}_1;\cdots;\boldsymbol{m}_n]$.

The Multi-PP game (1) considered under this setting extends the *full-information revenue maximization* game in Narang et al. (2022) and the *multi-agent mean estimation* problem in Li et al. (2022) by taking into account the multiplex network structure. The following proposition provides an exact characterization of the PSE and NE:

Proposition 4.1. *Consider the Multi-PP game (1) instantiated by (17) and (18). Suppose that $\sum_{j=1}^n A_{ij} = 1$ for all $i \in [n]$. Then, the following hold:*

i) There exists a unique PSE with closed form

$$\boldsymbol{\theta}^{\mathrm{pse}} = \left([(1+\rho)\boldsymbol{I}_n - \rho\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}]^{-1} \otimes \boldsymbol{I}_{\bar{p}}\right)\boldsymbol{m}, \tag{19}$$

if and only if $(1+\rho)\boldsymbol{I}_n - \rho\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}$ is invertible. Moreover, the RRM dynamics (6) converges to the PSE if and only if

$$\chi\left(\frac{\rho}{1+\rho}\boldsymbol{A} + \frac{\bar{\varepsilon}}{1+\rho}\boldsymbol{P}\right) < 1, \tag{20}$$

where $\chi(\cdot)$ denotes the spectral radius (i.e., the maximum absolute eigenvalue) of a matrix.

ii) There exists a unique NE with closed form

$$\boldsymbol{\theta}^{\mathrm{ne}} = \left(\left[\left(1+\frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{I}_n - \frac{\rho}{1-\bar{\varepsilon}}\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}\right]^{-1} \otimes \boldsymbol{I}_{\bar{p}}\right)\boldsymbol{m},$$

if and only if $\left(1+\frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{I}_n - \frac{\rho}{1-\bar{\varepsilon}}\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}$ is invertible. Moreover, the BR dynamics (16) converges to the NE if and only if

$$\chi\left(\frac{\rho}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{A} + \frac{(1-\bar{\varepsilon})\bar{\varepsilon}}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{P}\right) < 1. \tag{21}$$

We remark that the necessary and sufficient conditions for the existence of the PSEs and NEs can be found along with the proof in Appendix F.1.

Proposition 4.1 reveals that the agent and population networks collaboratively contribute to the properties of the PSE and NE with different weights. Furthermore, (20) and (21) provide tight conditions for the stability of the PSE and NE. Below, we illustrate Proposition 4.1 under proper simplifications. For simplicity, we let $\bar{p} = 1$ and focus on the PSE solution.

Structure of the PSE Solution. We first focus on the structure of the PSE solutions. With $\bar{p} = 1$, the PSE solution (19) reduces to

$$\boldsymbol{\theta}^{\mathrm{pse}} = ((1+\rho)\boldsymbol{I}_n - \bar{\varepsilon}\boldsymbol{P} - \rho\boldsymbol{A})^{-1}\boldsymbol{m}. \tag{22}$$

Observe that if $\bar{\varepsilon} > 0$, then the PSE solution at agent i, i.e., $\boldsymbol{\theta}_i^{\mathrm{pse}}$, is the *weighted Katz-Bonacich centrality* (Jackson, 2010; Jackson & Zenou, 2015) of node i in a weighted graph $\mathcal{G}^{\mathtt{W}}$ with adjacency matrix $\boldsymbol{W}(\bar{\varepsilon},\rho) := \frac{\rho}{1+\rho}\boldsymbol{A} + \frac{\bar{\varepsilon}}{1+\rho}\boldsymbol{P}$, i.e., $\mathcal{G}^{\mathtt{W}}$ is a weighted combination of $\mathcal{G}^{\mathtt{A}}$ and $\mathcal{G}^{\mathtt{P}}$.

Moreover, we study how the PSE solution of each agent will be affected if the local data distribution is perturbed. Specifically, suppose that the j-th mean $\boldsymbol{m}_j$ is perturbed by κ and let $\tilde{\boldsymbol{\theta}}^{\mathrm{pse}}(j) \in \mathbb{R}^n$ be the new PSE. Then,

$$\tilde{\boldsymbol{\theta}}^{\mathrm{pse}}(j) - \boldsymbol{\theta}^{\mathrm{pse}} = \frac{\kappa}{1+\rho}\left(\boldsymbol{I}_n - \boldsymbol{W}(\bar{\varepsilon},\rho)\right)^{-1}\boldsymbol{e}_j. \tag{23}$$

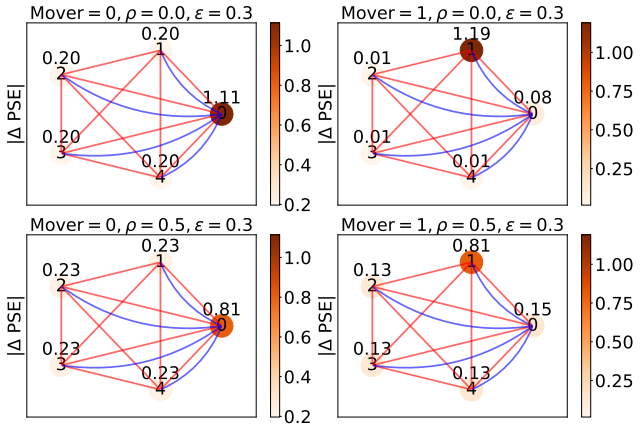

Figure 2: Illustrating the magnitude of perturbation $|\tilde{\boldsymbol{\theta}}^{\text{pse}}(j) - \boldsymbol{\theta}^{\text{pse}}(j)|$ for the PSE of mean estimation problem (17) when the mean of one of the local populations ('Mover') is perturbed. ($\mathcal{G}^{\mathtt{A}}$: red, $\mathcal{G}^{\mathtt{P}}$: blue.)

The changes in the PSE solution at agent i after perturbing the j-th population is given by

$$\Delta_{ij} \coloneqq \tilde{\boldsymbol{\theta}}_i^{\text{pse}}(j) - \boldsymbol{\theta}_i^{\text{pse}} = \frac{\kappa}{1+\rho}[(\boldsymbol{I}_n - \boldsymbol{W}(\bar{\varepsilon},\rho))^{-1}]_{ij} = \frac{\kappa}{1+\rho}\sum_{k=1}^{\infty}[(\boldsymbol{W}(\bar{\varepsilon},\rho))^k]_{ij}. \quad (24)$$

If $\bar{\varepsilon} > 0$ and $\rho = 0$, then $\boldsymbol{W}(\bar{\varepsilon},\rho) = \bar{\varepsilon}\boldsymbol{P}$ and Δ_{ij} represents the total number of walks from node i to node j in $\mathcal{G}^{\mathtt{P}}$. Note that even when $\rho = 0$, i.e., the graph regularization is ignored in (2) and the agents are not directly cooperating, such a distribution shift at the local population of another agent can still affect the PSE and NE solutions across the network.

In Figure 2, we illustrate (24) on a simple configuration with $\mathcal{G}^{\mathtt{A}}$ being an undirected complete graph and $\mathcal{G}^{\mathtt{P}}$ being an undirected star graph. The weights on $\boldsymbol{A}$ are assigned such that $A_{ij} = 1/\deg(i)$ if (i,j) is an edge in $\mathcal{G}^{\mathtt{A}}$. In the figure, we compute (24) at different combinations of $(\bar{\varepsilon},\rho)$ and perturb population j (denoted as 'Mover' in the figure). Observe that the pattern of changes $\{|\boldsymbol{\Delta}_{ij}|\}_{i\in[n]}$ depends on the location of the perturbed population j, i.e., $|\boldsymbol{\Delta}_{ij}|$ increases if agent i is closer to agent j on the combined graph $\mathcal{G}^{\mathtt{W}}$, corroborating the calculation (24). Moreover, increasing ρ has the effect of making the variations of $|\boldsymbol{\Delta}_{ij}|$ more uniform across the network. This is reasonable due to the consensus inducing effect of graph regularization.

Effects of Cooperation on Stability of PSE. We notice that (20) provides a *necessary and sufficient* condition for the RRM to converge to the PSE. The condition is tight such that if (20) is violated, the RRM may diverge and the PSE solution becomes unstable. Our focus below is to investigate the satisfaction of (20) under different settings of $\boldsymbol{A}, \boldsymbol{P}$ and the cooperation strength ρ via numerical illustrations.

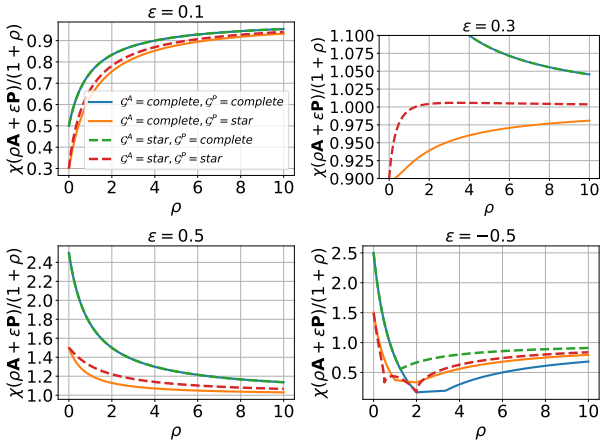

Figure 3: Evaluating the necessary and sufficient condition (20) for stability of RRM on Multi-PP game with (17), (18).

Intuitively, it may seem that increasing the strength of regularization ρ would promote stability (of PSE) as it forces agents to reach consensus while optimizing for the local risk in (6). However, we also recall from our discussions on Theorem 3.3 that our *sufficient* condition for stability (cf. (9)) maybe violated for $\rho \gg 1$ when $\boldsymbol{A}$ is asymmetric[4].

This motivated us to investigate the above phenomena by evaluating the tight condition (20) with different combinations of $\boldsymbol{A}, \boldsymbol{P}$ and parameters $\bar{\varepsilon}, \rho$. Our results are shown in Figure 3. We set $\mathcal{G}^{\mathtt{A}}$ as either a complete graph or a star graph. The weighted adjacency matrix $\boldsymbol{A}$ follows the same design as in Figure 2. Note that $\boldsymbol{A} \neq \boldsymbol{A}^\top$ if $\mathcal{G}^{\mathtt{A}}$ is not regular. From the figure, we observe that for small (resp. large) sensitivity parameter, $\bar{\varepsilon} = 0.1$ (resp. $\bar{\varepsilon} = 0.5$), the condition (20) is always satisfied (resp. violated) irrespective of the value of ρ. However, at $\bar{\varepsilon} = 0.3$, we note that increasing ρ would lead to violation of (20) for the case when both $\mathcal{G}^{\mathtt{A}}, \mathcal{G}^{\mathtt{P}}$ are star graphs. This coincides with the previous observation that $\rho \gg 1$ can destabilize the PSE when $\boldsymbol{A}$ is asymmetric. On the other hand, at $\bar{\varepsilon} = -0.5$, increasing ρ can stabilize the PSE, i.e., satisfying (20). Our results indicate that the stability of PSE depends jointly on the multiplex network structure. Additional results are available in Appendix G.

4.2. Logistic Regression

Our second case study focuses on a multi-agent logistic regression game under performative data. The problem setup has been described in Example 2.3. Particularly, we set $\rho_i = \rho$, $\mathcal{Z}_i = \mathbb{R}^{\bar{p}}$ for all $i \in [n]$ and consider the loss function given by (3). Suppose that $\bar{\boldsymbol{Z}}_i = (\bar{\boldsymbol{x}}_i, \bar{y}_i) \in \mathbb{R}^{\bar{p}} \times \{0,1\}$ follows the base distribution $\bar{\mathcal{D}}_i$ satisfying $\mathbb{P}(\bar{y}_i = 0) = q \in (0,1)$, $\mathbb{E}[\bar{\boldsymbol{x}}_i \mid \bar{y}_i = 0] = \boldsymbol{m}_i^0 \in \mathbb{R}^{\bar{p}}$, and $\mathbb{E}[\bar{\boldsymbol{x}}_i^1 \mid \bar{y}_i = 0] = \boldsymbol{m}_i^1 \in \mathbb{R}^{\bar{p}}$. Meanwhile, sample $\boldsymbol{Z}_i = (\boldsymbol{x}_i, y_i) \in \mathbb{R}^{\bar{p}} \times \{0,1\}$ from the i-th population follows

[4] We remark that as $\chi(\boldsymbol{A}) = 1$, a *sufficient* condition for (20) is $|\bar{\varepsilon}| < 1/\chi(\boldsymbol{P})$, which is slightly weaker than (9).

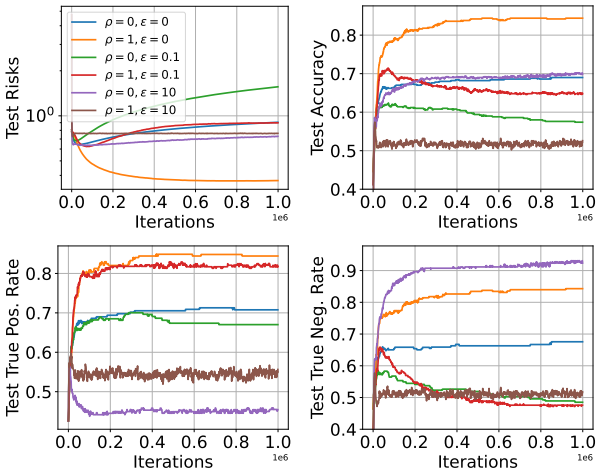

Figure 4: Learning dynamics of Multi-PP logistic regression game using SG-GD. ($\mathcal{G}^{\mathtt{A}}$: complete, $\mathcal{G}^{\mathtt{P}}$: star)

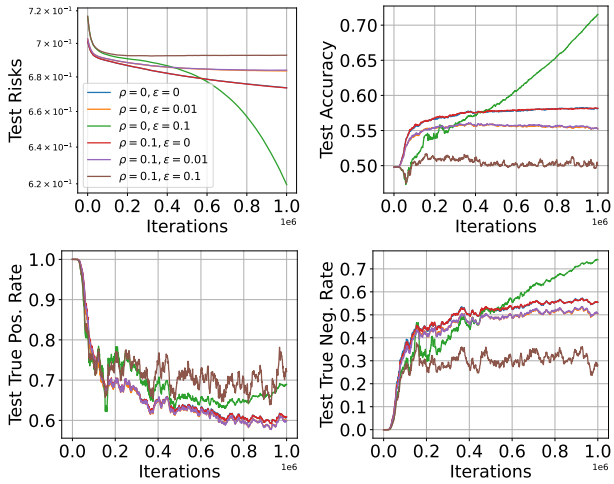

Figure 5: Learning dynamics of Multi-PP logistic regression game using SG-GD on Kaggle Give Me Some Credit dataset. ($\mathcal{G}^{\mathtt{A}}$: complete, $\mathcal{G}^{\mathtt{P}}$: star)

distribution $\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})$ as described in (4) and (5).

Since the logistic loss is highly nonlinear, we do not expect to obtain closed-form solutions for the PSE and NE. As a preliminary investigation, our remedy is to study an approximate equilibrium using Taylor expansion and investigate the latter's dependence on the network structure. Denote $\bar{\boldsymbol{m}}_i = (1-q)\boldsymbol{m}_i^1 - q\boldsymbol{m}_i^0$ and $\bar{\boldsymbol{m}} = \frac{1}{2}[\bar{\boldsymbol{m}}_1; \cdots; \bar{\boldsymbol{m}}_n]$. Focusing on the PSE, we apply Taylor approximation for the logistic loss (3) around $\mathbf{0}$ and linearize the latter. If the true PSE solution is close to $\mathbf{0}$, then it can be shown (see Appendix F.2) that the approximate PSE $\widehat{\boldsymbol{\theta}}^{\mathrm{pse}}$ satisfies

$$\frac{1}{2}\bar{\boldsymbol{m}}_i + \frac{q\bar{\varepsilon}_i}{2}\sum_{j=1}^n P_{ij}\widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} + \rho_i \sum_{j=1}^n A_{ij}(\widehat{\boldsymbol{\theta}}_i^{\mathrm{pse}} - \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}}) = \mathbf{0},$$

for $i \in [n]$, and thus

$$\widehat{\boldsymbol{\theta}}^{\mathrm{pse}} = \Big(\big[(\mathrm{Diag}(\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) + \mathrm{Diag}\left(\frac{q}{2}\bar{\boldsymbol{\varepsilon}}\right)\boldsymbol{P}\big]^{-1} \otimes \boldsymbol{I}_{\bar{p}}\Big)\bar{\boldsymbol{m}}.$$

The above expression illustrates that the PSE for the logistic regression game admits a similar dependence on the network structure as the MSE game in Section 4.1, with the exception that the effect of $\boldsymbol{P}$ will be weighted by q. Nonetheless, we should mention that this is only a crude characterization of the PSE when the latter is close to $\mathbf{0}$. In fact, extending the analysis by approximation around other points such as $\tilde{\boldsymbol{\theta}}^0 \neq \mathbf{0}$ reveals further nonlinear dependence on $\boldsymbol{P}$.

Numerical Illustration. We examine the network effects on the multi-agent logistic regression game via simulating the SG-GD algorithm. We first describe the data generation process with $n = 5$ agents. Similar to Bellet et al. (2018), each agent holds a training dataset of size $1 \leq S_i \leq 100$ and a testing dataset of size 100. For each $i \in [n]$, a target hyperplane is first generated as $\boldsymbol{m}_i^\star \sim N(\boldsymbol{m}^\star, 10^{-1}\boldsymbol{I})$ for some fixed $\boldsymbol{m}^\star$, S_i random feature-label pairs are then generated with $\boldsymbol{x}_i^s \sim N(\mathbf{0}, \boldsymbol{I})$ and $y_i^s = \frac{\mathrm{sign}(\langle \boldsymbol{x}_i^s \boldsymbol{m}_i^\star \rangle)+1}{2}$ for $s = 1, \ldots, S_i$ as the training set, and the testing set samples $\{(\bar{\boldsymbol{x}}_i^s, \bar{y}_i^s)\}_{s=1}^{100}$ are generated in a similar manner.

We apply the SG-GD procedure for the distributed learning task. To account for the performative effect in Example 2.3, for any deployed $\boldsymbol{\theta}^t$, the drawn sample, $(\boldsymbol{X}_i, Y_i) \sim \mathcal{D}_i(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_{\mathcal{N}_i}^t)$, follows the generation rule $Y_i = y_i^s$, $\boldsymbol{X}_i = \boldsymbol{x}_i^s + (1-y_i^s)\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^t$ with $s \sim \mathcal{U}\{1, \ldots, S_i\}$. Meanwhile, the prediction model is evaluated with the performative effect on the *testing data*, i.e., the samples are modified by similar rule as above except for taking $(\bar{\boldsymbol{x}}_i^s, \bar{y}_i^s)$ instead.

From Figure 4, we observe that while enabling graph regularization (with $\rho = 1$) allows the agents to maintain a high accuracy in classification in general ($\bar{\varepsilon} \in \{0, 0.1\}$), under large distribution shifts ($\bar{\varepsilon} = 10$) of negative samples, it may lead to degraded performance.

Finally, we validate the results of this paper in a semi-realistic setting by sampling from a Kaggle dataset (Give Me Some Credit[5]). Similar to the previous experiment, we observe from Figure 5 that when the distribution shift is large ($\bar{\varepsilon} = 0.1$), there is a significant degradation of classification accuracy from $\rho = 0$ to $\rho = 0.1$.

Conclusions. We have studied a new setting for Multi-PP games over multiplex networks. We analyze the existence and uniqueness of the equilibriums to the game, and provide insights on the role of network structures to these equilibriums. Our results also indicate an issue that increased strength of cooperation may destabilize the Multi-PP game for some topology configurations, which deserves future investigation.

Acknowledgement. This work is supported in part by the Hong Kong RGC Project #24203520.

[5] https://www.kaggle.com/c/GiveMeSomeCredit

References

Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. Networks, shocks, and systemic risk. Technical report, National Bureau of Economic Research, 2015.

Allen, J. M., Skeldon, A. C., and Hoyle, R. B. Social influence preserves cooperative strategies in the conditional cooperator public goods game on a multiplex network. *Physical Review E*, 98(6):062305, 2018.

Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. Personalized and private peer-to-peer machine learning. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pp. 473–481. PMLR, 2018.

Bramoullé, Y., Kranton, R., and D'amours, M. Strategic interaction and networks. *American Economic Review*, 104(3):898–930, 2014.

Candogan, O., Bimpikis, K., and Ozdaglar, A. Optimal pricing in networks with externalities. *Operations Research*, 60(4):883–905, 2012.

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., and Arenas, A. Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022, 2013.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Drusvyatskiy, D. and Xiao, L. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

Facchinei, F. and Pang, J.-S. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., and Yariv, L. Network games. *The Review of Economic Studies*, 77(1):218–244, 2010.

Gómez-Gardenes, J., Reinares, I., Arenas, A., and Floría, L. M. Evolution of cooperation in multiplex networks. *Scientific Reports*, 2(1):1–6, 2012.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

Izzo, Z., Ying, L., and Zou, J. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of 38th International Conference on Machine Learning*, pp. 4641–4650. PMLR, 2021.

Jackson, M. O. *Social and Economic Networks*. Princeton University Press, 2010.

Jackson, M. O. and Zenou, Y. Games on networks. In *Handbook of Game Theory with Economic Applications*, volume 4, pp. 95–163. Elsevier, 2015.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Li, Q., Yau, C.-Y., and Wai, H. T. Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *Advances in Neural Information Processing Systems 35*, 2022.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems 30*, 2017.

Liu, S., Pan, S. J., and Ho, Q. Distributed multi-task relationship learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 937–946, 2017.

Mazumdar, E., Ratliff, L. J., and Sastry, S. S. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems 33*, pp. 4929–4939, 2020.

Miller, J. P., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of 38th International Conference on Machine Learning*, pp. 7710–7720. PMLR, 2021.

Moulines, E. and Bach, F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, 2011.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. Multiplayer performative prediction: Learning in decision-dependent games. *arXiv preprint arXiv:2201.03398*, 2022.

Nassif, R., Vlaski, S., Richard, C., Chen, J., and Sayed, A. H. Multitask learning over graphs: An approach for distributed, streaming machine learning. *IEEE Signal Processing Magazine*, 37(3):14–25, 2020.

Nedic, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Parise, F. and Ozdaglar, A. A variational inequality framework for network games: Existence, uniqueness, convergence and sensitivity analysis. *Games and Economic Behavior*, 114:47–82, 2019.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Piliouras, G. and Yu, F.-Y. Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv preprint arXiv:2201.10483*, 2022.

Vanhaesebrouck, P., Bellet, A., and Tommasi, M. Decentralized collaborative learning of personalized models over networks. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 509–517. PMLR, 2017.

Zrnic, T., Mazumdar, E., Sastry, S., and Jordan, M. Who leads and who follows in strategic classification? In *Advances in Neural Information Processing Systems 34*, pp. 15257–15269, 2021.

A. Verifying Assumptions 2.1 and 2.2 for Example 2.3

Indeed, Assumption 2.1 has been verified in Perdomo et al. (2020, Section G). Here, we only provide the verification of Assumption 2.2. Suppose that $\boldsymbol{Z}_i = (\boldsymbol{x}_i, y_i) \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})$ and $\boldsymbol{Z}_i' = (\boldsymbol{x}_i', y_i') \sim \mathcal{D}_i(\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\mathcal{N}_i}')$. Let $\mathbf{1}_A$ denotes the indicator random variable for some event A. Then, it is implied by (4) and (5) that

$$\boldsymbol{x}_i = \left(\bar{\boldsymbol{x}}_i + \bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j\right)\mathbf{1}_{\{\bar{y}_i=0\}} + \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=1\}} \text{ and } y_i = \bar{y}_i,$$
$$\boldsymbol{x}_i' = \left(\bar{\boldsymbol{x}}_i' + \bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j'\right)\mathbf{1}_{\{\bar{y}_i'=0\}} + \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=1\}} \text{ and } y_i' = \bar{y}_i',$$

for some $\bar{\boldsymbol{Z}}_i = (\bar{\boldsymbol{x}}_i, \bar{y}_i) \sim \overline{\mathcal{D}}_i$ and $\bar{\boldsymbol{Z}}_i' = (\bar{\boldsymbol{x}}_i', \bar{y}_i') \sim \overline{\mathcal{D}}_i$. Thus, we have

$$\begin{aligned}
&\|\boldsymbol{Z}_i - \boldsymbol{Z}_i'\|_2 = \left\| \begin{bmatrix} \boldsymbol{x}_i \\ y_i \end{bmatrix} - \begin{bmatrix} \boldsymbol{x}_i' \\ y_i' \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \left(\bar{\boldsymbol{x}}_i + \bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j\right)\mathbf{1}_{\{\bar{y}_i=0\}} + \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=1\}} \\ \bar{y}_i \end{bmatrix} - \begin{bmatrix} \left(\bar{\boldsymbol{x}}_i' + \bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j'\right)\mathbf{1}_{\{\bar{y}_i'=0\}} + \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=1\}} \\ \bar{y}_i' \end{bmatrix} \right\|_2 \\
&\le \left\| \begin{bmatrix} \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=0\}} - \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=0\}} + \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=1\}} - \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=1\}} \\ \bar{y}_i - \bar{y}_i' \end{bmatrix} \right\|_2 + \left\| \begin{bmatrix} \left(\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j\right)\mathbf{1}_{\{\bar{y}_i=0\}} - \left(\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j'\right)\mathbf{1}_{\{\bar{y}_i'=0\}} \\ \bar{y}_i - \bar{y}_i' \end{bmatrix} \right\|_2.
\end{aligned}$$

This, together with the definition of Wasserstein 1-distance, gives

$$\begin{aligned}
&W_1(\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}), \mathcal{D}_i(\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\mathcal{N}_i}')) = \inf_{\pi \in \Pi(\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}), \mathcal{D}_i(\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\mathcal{N}_i}'))} \mathbb{E}_{(\boldsymbol{Z}_i, \boldsymbol{Z}_i') \sim \pi}[\|\boldsymbol{Z}_i - \boldsymbol{Z}_i'\|_2] \\
\le\ & \inf_{\pi \in \Pi(\overline{\mathcal{D}}_i, \overline{\mathcal{D}}_i)} \mathbb{E}_{(\bar{\boldsymbol{Z}}_i, \bar{\boldsymbol{Z}}_i') \sim \pi} \left\| \begin{bmatrix} \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=0\}} - \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=0\}} + \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=1\}} - \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=1\}} \\ \bar{y}_i - \bar{y}_i' \end{bmatrix} \right\|_2 \\
&+ \inf_{\pi \in \Pi(\overline{\mathcal{D}}_i, \overline{\mathcal{D}}_i)} \mathbb{E}_{(\bar{\boldsymbol{Z}}_i, \bar{\boldsymbol{Z}}_i') \sim \pi} \left\| \begin{bmatrix} \left(\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j\right)\mathbf{1}_{\{\bar{y}_i=0\}} - \left(\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j'\right)\mathbf{1}_{\{\bar{y}_i'=0\}} \\ \bar{y}_i - \bar{y}_i' \end{bmatrix} \right\|_2 \\
\le\ & \mathbb{E}_{(\bar{\boldsymbol{Z}}_i, \bar{\boldsymbol{Z}}_i') \sim \pi_0} \left\| \begin{bmatrix} \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=0\}} - \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=0\}} + \bar{\boldsymbol{x}}_i \mathbf{1}_{\{\bar{y}_i=1\}} - \bar{\boldsymbol{x}}_i' \mathbf{1}_{\{\bar{y}_i'=1\}} \\ \bar{y}_i - \bar{y}_i' \end{bmatrix} \right\|_2 \\
&+ \mathbb{E}_{(\bar{\boldsymbol{Z}}_i, \bar{\boldsymbol{Z}}_i') \sim \pi_0} \left\| \begin{bmatrix} \left(\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j\right)\mathbf{1}_{\{\bar{y}_i=0\}} - \left(\bar{\varepsilon}_i \sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j'\right)\mathbf{1}_{\{\bar{y}_i'=0\}} \\ \bar{y}_i - \bar{y}_i' \end{bmatrix} \right\|_2
\end{aligned}$$

where $\Pi(\mathcal{D}, \mathcal{D}')$ denotes the set of joint distributions with marginals $\mathcal{D}$ and $\mathcal{D}'$ on the first and second factors, respectively and π_0 is an arbitrary joint distribution in $\Pi(\overline{\mathcal{D}}_i, \overline{\mathcal{D}}_i)$. Taking π_0 be a joint distribution satisfying

$$(\bar{\boldsymbol{Z}}_i, \bar{\boldsymbol{Z}}_i') \sim \pi_0 \Rightarrow \bar{\boldsymbol{x}}_i = \bar{\boldsymbol{x}}_i' \text{ and } \bar{y}_i = \bar{y}_i',$$

then the above can be further bounded as follows:

$$W_1(\mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}), \mathcal{D}_i(\boldsymbol{\theta}_i', \boldsymbol{\theta}_{\mathcal{N}_i}')) \le 0 + \left\| \bar{\varepsilon}_i \sum_{j=1}^n P_{ij}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_j') \right\|_2 \le \bar{\varepsilon}_i \sqrt{|\mathcal{N}_i|} \sqrt{\sum_{j=1}^n P_{ij} \|\boldsymbol{\theta}_j - \boldsymbol{\delta}_j\|_2^2},$$

where the second inequality follows from the fact that $\|\sum_{i=1}^m \boldsymbol{a}_i\|_2^2 \le m \sum_{i=1}^m \|\boldsymbol{a}_i\|_2^2$ for vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m \ne \mathbf{0}$. Therefore, Assumption 2.2 is satisfied by Example 2.3 with $\varepsilon_i = \bar{\varepsilon}_i \sqrt{|\mathcal{N}_i|}$.

B. Auxiliary Lemmas

We first introduce the following technical lemma, which is adapted from Perdomo et al. (2020, Lemma D.4).

Lemma B.1. *Suppose that Assumption 2.1 ii) holds. Then, for any $\boldsymbol{\theta}_i \in \mathbb{R}^{p_i}$ and two probability measures $\mathcal{P}, \mathcal{P}'$ on $\mathcal{Z}_i$, we have*

$$\left\|\mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{P}}\left[\nabla\ell_i(\boldsymbol{\theta}_i;\boldsymbol{Z}_i)\right]-\mathbb{E}_{\boldsymbol{Z}_i'\sim\mathcal{P}'}\left[\nabla\ell_i(\boldsymbol{\theta}_i;\boldsymbol{Z}_i')\right]\right\|_2 \le L_i W_1\left(\mathcal{P},\mathcal{P}'\right).$$

Next, for $\boldsymbol{\theta}=[\boldsymbol{\delta}_1;\ldots;\boldsymbol{\delta}_n]\in\mathbb{R}^p$, we define

$$J^i_{\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i) \coloneqq \mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}\left[\nabla\ell_i(\boldsymbol{\theta}_i;\boldsymbol{Z}_i)\right] \text{ and } J_{\boldsymbol{\delta}}(\boldsymbol{\theta}) \coloneqq [J^1_{\boldsymbol{\delta}_1,\boldsymbol{\delta}_{\mathcal{N}_1}}(\boldsymbol{\theta}_1);\ldots;J^i_{\boldsymbol{\delta}_n,\boldsymbol{\delta}_{\mathcal{N}_n}}(\boldsymbol{\theta}_n)].$$

Then, we introduce the following lemma that will be used multiple times in our analysis.

Lemma B.2. *Suppose that Assumptions 2.1 and 2.2 hold. Then, for any $\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\delta}'\in\mathbb{R}^p$ and $\boldsymbol{\alpha}\coloneqq(\alpha_1,\ldots,\alpha_n)\in\mathbb{R}^n$, we have*

$$\left\|\boldsymbol{\alpha}\odot\left(J_{\boldsymbol{\theta}}(\boldsymbol{\delta}')-J_{\boldsymbol{\delta}}(\boldsymbol{\delta}')\right)\right\|_2 \le \sqrt{\max_{j\in[n]}\left\{\sum_{i=1}^n P_{ij}\alpha_i^2L_i^2\varepsilon_i^2\right\}}\,\left\|\boldsymbol{\theta}-\boldsymbol{\delta}\right\|_2.$$

Proof. It follows from Lemma B.1 that

$$\begin{aligned}
\left\|\boldsymbol{\alpha}\odot\left(J_{\boldsymbol{\theta}}(\boldsymbol{\delta}')-J_{\boldsymbol{\delta}}(\boldsymbol{\delta}')\right)\right\|_2^2 &= \sum_{i=1}^n\left\|\alpha_i\left(J^i_{\boldsymbol{\theta}_i,\boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i')-J^i_{\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i')\right)\right\|_2^2\\
&= \sum_{i=1}^n\alpha_i^2\left\|\mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}_{\mathcal{N}_i})}\left[\nabla\ell_i(\boldsymbol{\delta}_i';\boldsymbol{Z}_i)\right]-\mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}\left[\nabla\ell_i(\boldsymbol{\delta}_i';\boldsymbol{Z}_i)\right]\right\|_2^2\\
&\le \sum_{i=1}^n\alpha_i^2L_i^2W_1^2\left(\mathcal{D}_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}_{\mathcal{N}_i}),\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})\right).
\end{aligned}$$

Further using the Lipchitzness property of the distribution mapping $\mathcal{D}_i(\cdot)$ according to Assumption 2.2, we obtain

$$\begin{aligned}
\left\|\boldsymbol{\alpha}\odot\left(J_{\boldsymbol{\theta}}(\boldsymbol{\delta}')-J_{\boldsymbol{\delta}}(\boldsymbol{\delta}')\right)\right\|_2^2 &\le \sum_{i=1}^n\alpha_iL_i^2\varepsilon_i^2\sum_{j=1}^n P_{ij}\left\|\boldsymbol{\theta}_j-\boldsymbol{\delta}_j\right\|_2^2\\
&= \sum_{j=1}^n\sum_{i=1}^n P_{ij}\alpha_iL_i^2\varepsilon_i^2\left\|\boldsymbol{\theta}_j-\boldsymbol{\delta}_j\right\|_2^2\\
&\le \max_{j\in[n]}\left\{\sum_{i=1}^n P_{ij}\alpha_iL_i^2\varepsilon_i^2\right\}\sum_{j=1}^n\left\|\boldsymbol{\theta}_j-\boldsymbol{\delta}_j\right\|_2^2.
\end{aligned}$$

This implies that

$$\left\|\boldsymbol{\alpha}\odot\left(J_{\boldsymbol{\theta}}(\boldsymbol{\delta}')-J_{\boldsymbol{\delta}}(\boldsymbol{\delta}')\right)\right\|_2 \le \sqrt{\max_{j\in[n]}\left\{\sum_{i=1}^n \frac{P_{ij}L_i^2\varepsilon_i^2}{\left(\mu_i+\rho_i\right)^2}\right\}}\,\left\|\boldsymbol{\theta}-\boldsymbol{\delta}\right\|_2,$$

as desired. □

Lastly, the following lemma adapted from Li et al. (2022, Lemma 6) will be used in the proof of Theorem 3.5.

Lemma B.3. *Let $\{\gamma_t\}_{t\ge1}$ be a non-negative sequence and $a>0$. Suppose that $\sup_{t\ge1}\gamma_t<2/a$ and $\gamma_t/\gamma_{t+1}\le 1+\gamma_{t+1}a/2$ for all $t\ge1$. Then, we have*

$$\sum_{j=0}^t\gamma_{j+1}^2\prod_{k=j+1}^t\left(1-a\gamma_{k+1}\right)\le\frac{2}{a}\gamma_{t+1}.$$

C. Proof of Theorem 3.3

Proof. Let $\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\delta}, \boldsymbol{\delta}' \in \mathbb{R}^p$ be such that $\boldsymbol{\theta}' = \mathcal{T}(\boldsymbol{\theta})$ and $\boldsymbol{\delta}' = \mathcal{T}(\boldsymbol{\delta})$, i.e., it holds for all $i \in [n]$ that

$$\begin{aligned}\boldsymbol{\theta}_i' &= \underset{\boldsymbol{u}_i \in \mathbb{R}^{p_i}}{\arg\min}\ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})} \left[f_i(\boldsymbol{u}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i) \right] \\ &= \underset{\boldsymbol{u}_i \in \mathbb{R}^{p_i}}{\arg\min}\ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})} \left[\ell_i(\boldsymbol{u}_i; \boldsymbol{Z}_i) \right] + \frac{\rho_i}{2} \sum_{j=1}^n A_{ij} \|\boldsymbol{u}_i - \boldsymbol{\theta}_j\|_2^2, \end{aligned} \tag{25}$$

and

$$\begin{aligned}\boldsymbol{\delta}_i' &= \underset{\boldsymbol{u}_i \in \mathbb{R}^{p_i}}{\arg\min}\ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[f_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{M}_i}; \boldsymbol{Z}_i) \right] \\ &= \underset{\boldsymbol{u}_i \in \mathbb{R}^{p_i}}{\arg\min}\ \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\ell_i(\boldsymbol{u}_i; \boldsymbol{Z}_i) \right] + \frac{\rho_i}{2} \sum_{j=1}^n A_{ij} \|\boldsymbol{u}_i - \boldsymbol{\delta}_j\|_2^2. \end{aligned} \tag{26}$$

Since f_i is convex for $i \in [n]$ according to Assumption 2.1, then the first-order optimality conditions for Problems (25) and (26) read

$$\begin{aligned} J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i') + \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\theta}_i' - \boldsymbol{\theta}_j \right) &= \mathbf{0}, \\ J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') + \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) &= \mathbf{0}. \end{aligned}$$

These, together with the fact that $\ell_i(\cdot; \boldsymbol{Z}_i)$ is μ_i-strongly convex according to Assumption 2.1, imply

$$\begin{aligned} \mu_i \left\| \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\|_2^2 &\leq \left\langle J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\ &= \left\langle \left(J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i') + \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\theta}_i' - \boldsymbol{\theta}_j \right) \right) - \left(J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') + \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\ &\quad + \left\langle \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) - \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\theta}_i' - \boldsymbol{\theta}_j \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\ &= \left\langle \left(J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') + \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) \right) - \left(J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') + \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\ &\quad + \left\langle \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) - \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\theta}_i' - \boldsymbol{\theta}_j \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle. \end{aligned}$$

Further simplifying gives

$$\begin{aligned} &\mu_i \left\| \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\|_2^2 \\ &\leq \left\langle J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \left\langle \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\delta}_i' - \boldsymbol{\delta}_j \right) - \rho_i \sum_{j=1}^n A_{ij} \left(\boldsymbol{\theta}_i' - \boldsymbol{\theta}_j \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\ &= \left\langle J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \left\langle \rho_i \sum_{j=1}^n A_{ij} (\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle - \left\langle \rho_i \sum_{j=1}^n A_{ij} (\boldsymbol{\theta}_i' - \boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\ &= \left\langle J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \left\langle \rho_i \sum_{j=1}^n A_{ij} (\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle - \left(\rho_i \sum_{j=1}^n A_{ij} \right) \left\| \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\|_2^2, \\ &= \left\langle J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \left\langle \rho_i \sum_{j=1}^n A_{ij} (\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle - \rho_i \left\| \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\|_2^2, \end{aligned} \tag{27}$$

where equality (27) holds due to the fact that $\sum_{j=1}^n A_{ij} = 1$. Let $\boldsymbol{\mu} + \boldsymbol{\rho} \coloneqq (\tau_1, \ldots, \tau_n)$ with $\tau_i \coloneqq \mu_i + \rho_i$ for $i \in [n]$, then it follows from (27) that

$$(\mu_i + \rho_i) \left\|\boldsymbol{\theta}_i' - \boldsymbol{\delta}_i'\right\|_2^2 \leq \left\langle J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i'), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \left\langle \rho_i \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle$$

$$\iff \left\|\boldsymbol{\theta}_i' - \boldsymbol{\delta}_i'\right\|_2^2 \leq \left\langle \frac{1}{\mu_i + \rho_i} \left(J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \left\langle \frac{\rho_i}{\mu_i + \rho_i} \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \tag{28}$$

Let $\boldsymbol{\phi} \coloneqq (\phi_1, \ldots, \phi_n)$ with $\phi_i \coloneqq \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j)$ for $i \in [n]$, $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_n)$, and $\boldsymbol{\rho} \coloneqq (\rho_i, \ldots, \rho_n)$. Then, (28) implies that

$$\begin{aligned}
\left\|\boldsymbol{\theta}' - \boldsymbol{\delta}'\right\|_2^2 &= \sum_{i=1}^n \left\|\boldsymbol{\theta}_i' - \boldsymbol{\delta}_i'\right\|_2^2 \\
&\leq \sum_{i=1}^n \left\langle \frac{1}{\mu_i + \rho_i} \left(J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i') \right), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle + \sum_{i=1}^n \left\langle \frac{\rho_i}{\mu_i + \rho_i} \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i' - \boldsymbol{\delta}_i' \right\rangle \\
&= \left\langle \frac{\mathbf{1}_n}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \left(J_{\boldsymbol{\delta}}(\boldsymbol{\delta}') - J_{\boldsymbol{\theta}}(\boldsymbol{\delta}') \right), \boldsymbol{\theta}' - \boldsymbol{\delta}' \right\rangle + \left\langle \frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \boldsymbol{\phi}, \boldsymbol{\theta}' - \boldsymbol{\delta}' \right\rangle \\
&= \left\| \frac{\mathbf{1}_n}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \left(J_{\boldsymbol{\delta}}(\boldsymbol{\delta}') - J_{\boldsymbol{\theta}}(\boldsymbol{\delta}') \right) \right\|_2 \left\|\boldsymbol{\theta}' - \boldsymbol{\delta}'\right\|_2 + \left\| \frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \boldsymbol{\phi} \right\|_2 \left\|\boldsymbol{\theta}' - \boldsymbol{\delta}'\right\|_2,
\end{aligned} \tag{29}$$

where $\odot$ denote Hadamard product.

To upper bound the first term in (29), we apply Lemma B.2 with $\boldsymbol{\alpha} = \mathbf{1}_n / (\boldsymbol{\mu} + \boldsymbol{\rho})$ and obtain

$$\left\| \frac{\mathbf{1}_n}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \left(J_{\boldsymbol{\theta}}(\boldsymbol{\delta}') - J_{\boldsymbol{\delta}}(\boldsymbol{\delta}') \right) \right\|_2 \leq \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n \frac{P_{ij} L_i^2 \varepsilon_i^2}{(\mu_i + \rho_i)^2} \right\}} \left\|\boldsymbol{\theta} - \boldsymbol{\delta}\right\|_2. \tag{30}$$

To upper bound the second term in (29), we let $\widetilde{\boldsymbol{A}}$ be given by $\widetilde{A}_{ij} = \frac{\rho_i A_{ij}}{\mu_i + \rho_i}$ for $i, j \in [n]$, then we obtain

$$\begin{aligned}
\left\| \frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \boldsymbol{\phi} \right\|_2^2 &= \sum_{i=1}^n \left\| \sum_{j=1}^n \frac{\rho_i A_{ij}}{\mu_i + \rho_i} (\boldsymbol{\theta}_j - \boldsymbol{\delta}_j) \right\|_2^2 \\
&= \left\| (\widetilde{\boldsymbol{A}} \otimes \boldsymbol{I}_n)(\boldsymbol{\theta} - \boldsymbol{\delta}) \right\|_2^2 \\
&\leq \left\| \widetilde{\boldsymbol{A}} \right\|_2^2 \left\| \boldsymbol{\theta} - \boldsymbol{\delta} \right\|_2^2,
\end{aligned}$$

which implies that

$$\left\| \frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \odot \boldsymbol{\phi} \right\|_2 \leq \left\| \widetilde{\boldsymbol{A}} \right\|_2 \left\| \boldsymbol{\theta} - \boldsymbol{\delta} \right\|_2. \tag{31}$$

Plugging (30) and (31) back into (29) and then dividing both sides by $\left\|\boldsymbol{\theta}' - \boldsymbol{\delta}'\right\|_2$ give

$$\begin{aligned}
\left\|\boldsymbol{\theta}' - \boldsymbol{\delta}'\right\|_2 &\leq \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n \frac{P_{ij} L_i^2 \varepsilon_i^2}{(\mu_i + \rho_i)^2} \right\}} \left\|\boldsymbol{\theta} - \boldsymbol{\delta}\right\|_2 + \left\| \widetilde{\boldsymbol{A}} \right\|_2 \left\| \boldsymbol{\theta} - \boldsymbol{\delta} \right\|_2 \\
&= \left(\sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n \frac{P_{ij} L_i^2 \varepsilon_i^2}{(\mu_i + \rho_i)^2} \right\}} + \left\| \mathrm{Diag}\left(\frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \right) \boldsymbol{A} \right\|_2 \right) \left\| \boldsymbol{\theta} - \boldsymbol{\delta} \right\|_2.
\end{aligned}$$

Hence, $\mathcal{T}$ is a contraction mapping provided that

$$\sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n \frac{P_{ij} L_i^2 \varepsilon_i^2}{(\mu_i + \rho_i)^2} \right\}} + \left\| \mathrm{Diag}\left(\frac{\boldsymbol{\rho}}{\boldsymbol{\mu} + \boldsymbol{\rho}} \right) \boldsymbol{A} \right\|_2 < 1.$$

Therefore, if the above condition holds, then game (1) admits a unique PSE according to the Banach fixed-point theorem. □

D. Proof of Theorem 3.5

Proof. We recall that the SG-GD iteration reads as follows:

$$\boldsymbol{\theta}_i^{t+1} = \boldsymbol{\theta}_i^t - \gamma_{t+1}\left(\nabla \ell_i(\boldsymbol{\theta}_i^t; \boldsymbol{Z}_i^{t+1}) + \rho\sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_i^t - \boldsymbol{\theta}_j^t)\right),$$

where $\boldsymbol{Z}_i^{t+1} \sim \mathcal{D}(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_{\mathcal{N}_i}^t)$. As we have defined $\boldsymbol{\theta}^t = [\boldsymbol{\theta}_1^t; \boldsymbol{\theta}_2^t; \cdots; \boldsymbol{\theta}_n^t]$, the above can be written compactly as

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \gamma_{t+1}\left(\nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^t\right).$$

Let $\mathbb{E}_t[\cdot]$ denote the expectation conditioning on the filtration $\mathcal{F}_t := \sigma(\boldsymbol{\theta}^s, 0 \le s \le t)$, then we have

$$\mathbb{E}_t[\nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1})] = J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t). \tag{32}$$

We proceed with the proof by observing that

$$\begin{aligned}
\mathbb{E}_t[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2] &= \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 - 2\gamma_{t+1}\mathbb{E}_t[\langle \nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^t, \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle] \\
&\quad + \gamma_{t+1}^2\mathbb{E}_t[\|\nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^t\|_2^2] \\
&= \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 - 2\gamma_{t+1}\langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^t, \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle \\
&\quad + \gamma_{t+1}^2\mathbb{E}_t[\|\nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^t\|_2^2].
\end{aligned} \tag{33}$$

Notice that the PSE solution $\boldsymbol{\theta}^{\mathrm{pse}}$ satisfies

$$J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^{\mathrm{pse}}) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^{\mathrm{pse}} = \boldsymbol{0}. \tag{34}$$

We observe the following lower bound on the inner product:

$$\begin{aligned}
&\langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t, \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle \\
&= \langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle + \langle J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^{\mathrm{pse}}), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle + \langle \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)(\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle \\
&\ge \langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle + \mu\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + \rho\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 - \rho\|\boldsymbol{A}\|_2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 \\
&= \langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle + (\mu + \rho(1 - \|\boldsymbol{A}\|_2))\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2.
\end{aligned} \tag{35}$$

Moreover, applying Lemma B.2 with $\boldsymbol{\alpha} = \boldsymbol{1}_n$, we have

$$\begin{aligned}
\|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t)\|_2 &\le \sqrt{\max_{j\in[n]}\sum_{i=1}^n P_{ij}L^2\varepsilon^2}\,\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2 \\
&= L\varepsilon\sqrt{\|\boldsymbol{P}\|_\infty}\,\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2,
\end{aligned} \tag{36}$$

which, together with the Cauchy-Schwarz inequality, implies that

$$\begin{aligned}
|\langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t), \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle| &\le \|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t)\|_2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2 \\
&\le L\varepsilon\sqrt{\|\boldsymbol{P}\|_\infty}\,\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2.
\end{aligned}$$

Plugging this back in to (35) gives

$$\begin{aligned}
\langle J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^t, \boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\rangle &\ge \left(\mu + \rho(1 - \|\boldsymbol{A}\|_2) - L\varepsilon\sqrt{\|\boldsymbol{P}\|_\infty}\right)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 \\
&= \widetilde{\mu}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2,
\end{aligned} \tag{37}$$

where $\widetilde{\mu} = \mu + \rho(1 - \|\boldsymbol{A}\|_2) - L\varepsilon\sqrt{\|\boldsymbol{P}\|_\infty}$. Note that $\widetilde{\mu} > 0$ under condition (7) in Theorem 3.3. We further observe that by (32), the last term in (33) can be decomposed as

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2] \\
&= 2\mathbb{E}_t[\|\nabla \boldsymbol{\ell}(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) - J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t)\|_2^2] + 2\|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2 \\
&\le 2\sigma_0^2 + 2\sigma_1^2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|^2 + 2\|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2,
\end{aligned} \tag{38}$$

where the inequality follows from Assumption 3.4. Moreover,

$$\begin{aligned}
&\|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2 \\
&= \|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t) + J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2 \\
&\le 2\|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t)\|_2^2 + 2\|J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2 \\
&= 2\|J_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}^t) - J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t)\|_2^2 + 2\|J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^t) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t - (J_{\boldsymbol{\theta}^{\mathrm{pse}}}(\boldsymbol{\theta}^{\mathrm{pse}}) + \rho\big((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p\big)\boldsymbol{\theta}^{\mathrm{pse}})\|_2^2 \\
&\le 2L^2\varepsilon^2\|\boldsymbol{P}\|_\infty\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + 2\left(L\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2 + \rho\|\boldsymbol{I}_n - \boldsymbol{A}\|_2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2\right)^2 \\
&= 2\left(L^2\varepsilon^2\|\boldsymbol{P}\|_\infty + (L + \rho\|\boldsymbol{I}_n - \boldsymbol{A}\|_2)^2\right)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2, \qquad (39)
\end{aligned}$$

where the second equality uses (34) and the second inequality follows from (36) and Assumption 2.1 ii). Substituting (39) back to (38) gives

$$\begin{aligned}
&\mathbb{E}_t[\|\nabla\ell(\boldsymbol{\theta}^t; \boldsymbol{Z}^{t+1}) + \rho((\boldsymbol{I}_n - \boldsymbol{A}) \otimes \boldsymbol{I}_p)\boldsymbol{\theta}^t\|_2^2] \\
&\le 2\sigma_0^2 + 2\sigma_1^2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|^2 + 4\left(L^2\varepsilon^2\|\boldsymbol{P}\|_\infty + (L + \rho\|\boldsymbol{I}_n - \boldsymbol{A}\|_2)^2\right)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|^2 \\
&= 2\sigma_0^2 + 2\widetilde{\sigma}^2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|^2, \qquad (40)
\end{aligned}$$

where $\widetilde{\sigma}^2 = \sigma_1^2 + 2\left(L^2\varepsilon^2\|\boldsymbol{P}\|_\infty + (L + \rho\|\boldsymbol{I}_n - \boldsymbol{A}\|_2)^2\right)$. Combining (37) and (40) with (33) yields

$$\begin{aligned}
\mathbb{E}_t[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2] &\le \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 - 2\gamma_{t+1}\widetilde{\mu}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + 2\gamma_{t+1}^2\sigma_0^2 + 2\gamma_{t+1}^2\widetilde{\sigma}^2\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|^2 \\
&= (1 - 2\widetilde{\mu}\gamma_{t+1} + 2\widetilde{\sigma}^2\gamma_{t+1}^2)\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + 2\sigma_0^2\gamma_{t+1}^2 \\
&\le (1 - \widetilde{\mu}\gamma_{t+1})\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + 2\sigma_0^2\gamma_{t+1}^2,
\end{aligned}$$

where the last inequality holds due to $\sup_{t\ge1}\gamma_t \le \widetilde{\mu}/(2\widetilde{\sigma}^2)$. Solving the above recursion gives

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2] &\le \prod_{s=0}^{t}(1 - \widetilde{\mu}\gamma_{s+1})\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + 2\sigma_0^2\sum_{j=0}^{t}\gamma_{j+1}^2\prod_{k=j+1}^{t}(1 - \widetilde{\mu}\gamma_{k+1}) \\
&\le \prod_{s=0}^{t}(1 - \widetilde{\mu}\gamma_{s+1})\|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^{\mathrm{pse}}\|_2^2 + \frac{4\sigma_0^2}{\widetilde{\mu}}\gamma_{t+1},
\end{aligned}$$

where the last inequality follows from $\sup_{t\ge1}\gamma_t < 2/\widetilde{\mu}$ and $\gamma_t/\gamma_{t+1} \le 1 + \widetilde{\mu}\gamma_{t+1}/2$ for all $t \ge 1$, and then applying Lemma B.3. □

E. Proof of Theorem 3.8

E.1. Useful Technical Results

In this subsection, we provide some technical results that are used in the proof of Theorem 3.8.

Lemma E.1. *Suppose that Assumptions 2.1, 2.2, and 3.6 hold. Then, for any $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^p$, we have*

$$\langle G_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta}\rangle \ge \left(\min_{i\in[n]}\{\mu_i\} + \min_{i\in[n]}\{\rho_i\} - \|\operatorname{Diag}(\boldsymbol{\rho})\boldsymbol{A}\|_2\right)\|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2.$$

Proof. By Assumption 2.1 (i), we have

$$\begin{aligned}
&\langle G_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta}\rangle \\
&= \sum_{i=1}^{n}\left\langle G^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) - G^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{M}_i}), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i\right\rangle \\
&= \sum_{i=1}^{n}\left\langle \left(J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i) + \rho_i\sum_{j=1}^{n}A_{ij}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)\right) - \left(J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i) + \rho_i\sum_{j=1}^{n}A_{ij}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j)\right), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i\right\rangle \\
&= \sum_{i=1}^{n}\left(\langle J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i) - J^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i\rangle + \left\langle \rho_i\sum_{j=1}^{n}A_{ij}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) - \rho_i\sum_{j=1}^{n}A_{ij}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i\right\rangle\right). \qquad (41)
\end{aligned}$$

For each $i \in [n]$, the strongly convexity of $\ell_i(\cdot, \boldsymbol{Z}_i)$ implies that

$$\langle J^i_{\boldsymbol{\theta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i) - J^i_{\boldsymbol{\theta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\delta}_i), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i \rangle \geq \mu_i \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2^2. \tag{42}$$

Besides, for each $i \in [n]$, we have

$$\begin{aligned}
&\left\langle \rho_i \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) - \rho_i \sum_{j=1}^n A_{ij}(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i \right\rangle \\
&= \left\langle \rho_i \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_i - \boldsymbol{\delta}_i) - \rho_i \sum_{j=1}^n A_{ij}(\boldsymbol{\theta}_j - \boldsymbol{\delta}_j), \boldsymbol{\theta}_i - \boldsymbol{\delta}_i \right\rangle \\
&= \left(\rho_i \sum_{j=1}^n A_{ij} \right) \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2^2 - \rho_i \sum_{j=1}^n A_{ij} \langle \boldsymbol{\theta}_j - \boldsymbol{\delta}_j, \boldsymbol{\theta}_i - \boldsymbol{\delta}_i \rangle \\
&\geq \left(\rho_i \sum_{j=1}^n A_{ij} \right) \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2^2 - \rho_i \sum_{j=1}^n A_{ij} \|\boldsymbol{\theta}_j - \boldsymbol{\delta}_j\|_2 \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2 \\
&= \rho_i \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2^2 - \sum_{j=1}^n \rho_i A_{ij} \|\boldsymbol{\theta}_j - \boldsymbol{\delta}_j\|_2 \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2,
\end{aligned} \tag{43}$$

where (43) follows from the fact that $\sum_{j=1}^n A_{ij} = 1$. Plugging (42) and (43) back into (41) gives

$$\begin{aligned}
&\langle G_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
&\geq \sum_{i=1}^n \mu_i \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2^2 + \rho_i \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2^2 - \sum_{j=1}^n \rho_i A_{ij} \|\boldsymbol{\theta}_j - \boldsymbol{\delta}_j\|_2 \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2 \\
&\geq \min_{i \in [n]} \{\mu_i\} \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 + \min_{i \in [n]} \{\rho_i\} \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 - \sum_{i=1}^n \sum_{j=1}^n \rho_i A_{ij} \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2 \|\boldsymbol{\theta}_j - \boldsymbol{\delta}_j\|_2.
\end{aligned} \tag{44}$$

To proceed, we further bound the last term in (44). Let $\bar{\boldsymbol{A}} \in \mathbb{R}^{n \times n}$ be a matrix given by $\bar{A}_{ij} = \rho_i A_{ij}$ for $i, j \in [n]$ and $\boldsymbol{w} \in \mathbb{R}^n$ be a vector with $w_i \coloneqq \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2$ for $i \in [n]$, then we have

$$\sum_{i=1}^n \sum_{j=1}^n \rho_i A_{ij} \|\boldsymbol{\theta}_i - \boldsymbol{\delta}_i\|_2 \|\boldsymbol{\theta}_j - \boldsymbol{\delta}_j\|_2 = \boldsymbol{w}^\top \bar{\boldsymbol{A}} \boldsymbol{w} \leq \|\bar{\boldsymbol{A}}\|_2 \|\boldsymbol{w}\|_2^2 = \|\operatorname{Diag}(\boldsymbol{\rho}) \boldsymbol{A}\|_2 \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2, \tag{45}$$

where the inequality follows from the definition of matrix spectral norm. Plugging (45) back into (44) gives

$$\begin{aligned}
&\langle G_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
&\geq \min_{i \in [n]} \{\mu_i\} \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 + \min_{i \in [n]} \{\rho_i\} \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 - \max_{i \in [n]} \{\rho_i\} \|\boldsymbol{A}\|_2 \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 \\
&= \left(\min_{i \in [n]} \{\mu_i\} + \min_{i \in [n]} \{\rho_i\} - \|\operatorname{Diag}(\boldsymbol{\rho}) \boldsymbol{A}\|_2 \right) \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2,
\end{aligned}$$

as desired. □

Lemma E.2. *Suppose that Assumptions 2.1, 2.2, and 3.6 hold. Then, for any $\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\delta} \in \Theta$ and $i \in [n]$, we have*

$$H_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta}') \leq \max_{i \in [n]} \{L_i \varepsilon_i\} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 .$$

Proof. Suppose that $\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\delta}$. For each $i \in [n]$, we have

$$\begin{aligned}
&H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) - H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_{\mathcal{M}_i}) \\
&= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i) - f_i(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_{\mathcal{M}_i}; \boldsymbol{Z}_i) \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i} \\
&= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) - \ell_i(\boldsymbol{\theta}'_i; \boldsymbol{Z}_i) + \frac{\rho}{2} \sum_{i=1}^n A_{ij} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2 - \frac{\rho}{2} \sum_{i=1}^n A_{ij} \|\boldsymbol{\theta}'_i - \boldsymbol{\theta}'_j\|_2^2 \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i} \\
&= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) - \ell_i(\boldsymbol{\theta}'_i; \boldsymbol{Z}_i) \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i}. \qquad (46)
\end{aligned}$$

For each $i \in [n]$, we let $h_i(s) = \boldsymbol{\theta}'_i + s(\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i)$ for $s \in (0, 1)$. Then, we have

$$\begin{aligned}
\ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) - \ell_i(\boldsymbol{\theta}'_i; \boldsymbol{Z}_i) &= \int_0^1 \langle \nabla \ell_i \left(\boldsymbol{\theta}'_i + s(\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i); \boldsymbol{Z}_i \right), \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i \rangle \, \mathrm{d}s \\
&= \int_0^1 \langle \nabla \ell_i(h_i(s); \boldsymbol{Z}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i \rangle \, \mathrm{d}s. \qquad (47)
\end{aligned}$$

Plugging (47) into (46) gives

$$\begin{aligned}
H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) - H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_{\mathcal{M}_i}) &= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\int_0^1 \langle \nabla \ell_i(h_i(s); \boldsymbol{Z}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i \rangle \, \mathrm{d}s \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i} \\
&= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\langle \nabla \ell_i(h_i(s); \boldsymbol{Z}_i), \boldsymbol{\theta}_i - \boldsymbol{\theta}'_i \rangle \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i} \mathrm{d}s. \qquad (48)
\end{aligned}$$

Lemma B.1 implies that the function $\mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\cdot, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\nabla \ell_i(h_i(s); \boldsymbol{Z}_i) \right]$ is $L_i \varepsilon_i$-Lipschitz continuous, thus its gradient satisfies

$$\left\| \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\nabla \ell_i(h_i(s); \boldsymbol{Z}_i) \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i} \right\|_2 \leq L_i \varepsilon_i. \qquad (49)$$

Hence, combing (48) and (49), we have

$$\begin{aligned}
\left\| H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) - H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_{\mathcal{M}_i}) \right\|_2 &\leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{u}_i} \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{u}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\nabla \ell_i(h_i(s); \boldsymbol{Z}_i) \right] \Big|_{\boldsymbol{u}_i = \boldsymbol{\delta}_i} \right\|_2 \|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2 \, \mathrm{d}s \\
&\leq L_i \varepsilon_i \|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2,
\end{aligned}$$

where the first inequality holds due to the Cauchy-Schwartz inequality. This further implies that

$$\begin{aligned}
\|H_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta}')\|_2 &= \sqrt{\sum_{i=1}^n \left\| H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) - H^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_{\mathcal{M}_i}) \right\|_2^2} \\
&\leq \sqrt{\sum_{i=1}^n L_i^2 \varepsilon_i^2 \|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2^2} \\
&\leq \max_{i \in [n]} \{L_i \varepsilon_i\} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2,
\end{aligned}$$

as desired. □

E.2. Proving Theorem 3.8

Proof. For $i \in [n]$, since $\nabla_i f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i) = \nabla \ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) + \rho_i \sum_{j=1}^n (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)$, then we have

$$\begin{aligned}
&\left\| G^i_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) - G^i_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}) \right\|_2 \\
=& \left\| \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})} \left[\nabla_i f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}_i) \right] - \mathbb{E}_{\boldsymbol{Z}'_i \sim \mathcal{D}_i(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\nabla_i f_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{M}_i}; \boldsymbol{Z}'_i) \right] \right\|_2 \\
=& \left\| \mathbb{E}_{\boldsymbol{Z}_i \sim \mathcal{D}_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i})} \left[\nabla \ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}_i) \right] - \mathbb{E}_{\boldsymbol{Z}'_i \sim \mathcal{D}_i(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i})} \left[\nabla \ell_i(\boldsymbol{\theta}_i; \boldsymbol{Z}'_i) \right] \right\|_2 \\
=& \left\| J_{\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i) - J_{\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i}}(\boldsymbol{\theta}_i) \right\|_2 .
\end{aligned}$$

Then, we have

$$\begin{aligned}
|\langle G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle| &\leq \|G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\theta})\|_2 \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2 \\
&= \|J_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - J_{\boldsymbol{\delta}}(\boldsymbol{\theta})\|_2 \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2 \\
&\leq \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n P_{ij} L_i^2 \varepsilon_i^2 \right\}} \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2 ,
\end{aligned}$$

where the last inequality follows from Lemma B.2 with $\boldsymbol{\alpha} = \mathbf{1}_n$. This, together with Lemma E.1, yields

$$\begin{aligned}
&\langle G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
=\ & \langle G_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle + \langle G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
\geq\ & (\mu_{\min} + \rho_{\min} - \|\operatorname{Diag}(\boldsymbol{\rho}) \boldsymbol{A}\|_2) \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 - \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n P_{ij} L_i^2 \varepsilon_i^2 \right\}} \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2, \\
=\ & \left(\mu_{\min} + \rho_{\min} - \|\operatorname{Diag}(\boldsymbol{\rho}) \boldsymbol{A}\|_2 - \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n P_{ij} L_i^2 \varepsilon_i^2 \right\}} \right) \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2. \qquad (50)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\langle H_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta})), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle &= \langle H_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta})), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle + \langle H_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
&\geq \langle H_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta})), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
&\geq -\|H_{\boldsymbol{\delta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta}))\|_2 \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2 \\
&\geq -\max_{i \in [n]} \{L_i \varepsilon_i\} \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2, \qquad (51)
\end{aligned}$$

where the first inequality due to $\langle H_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \geq 0$ by Assumption 3.7 and (51) follows from Lemma E.2. Then, combining (50) and (51) gives

$$\begin{aligned}
\langle \Phi_F(\boldsymbol{\theta}) - \Phi_F(\boldsymbol{\delta}) \rangle &= \langle G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - G_{\boldsymbol{\delta}}(\boldsymbol{\delta}), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle + \langle H_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - H_{\boldsymbol{\delta}}(\boldsymbol{\theta})), \boldsymbol{\theta} - \boldsymbol{\delta} \rangle \\
&\geq \left(\mu_{\min} + \rho_{\min} - \|\operatorname{Diag}(\boldsymbol{\rho}) \boldsymbol{A}\|_2 - \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n P_{ij} L_i^2 \varepsilon_i^2 \right\}} \right) \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 - \max_{i \in [n]} \{L_i \varepsilon_i\} \, \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2 \\
&= \left(\mu_{\min} + \rho_{\min} - \max_{i \in [n]} \{L_i \varepsilon_i\} - \|\operatorname{Diag}(\boldsymbol{\rho}) \boldsymbol{A}\|_2 - \sqrt{\max_{j \in [n]} \left\{ \sum_{i=1}^n P_{ij} L_i^2 \varepsilon_i^2 \right\}} \right) \|\boldsymbol{\theta} - \boldsymbol{\delta}\|_2^2,
\end{aligned}$$

where the first equality follows from (13). Thus, according to (12), the Multi-PP game (1) is strongly monotone if

$$\mu_{\min}+\rho_{\min}-\max_{i\in[n]}\{L_i\varepsilon_i\}-\|\operatorname{Diag}(\boldsymbol{\rho})\boldsymbol{A}\|_2-\sqrt{\max_{j\in[n]}\left\{\sum_{i=1}^n P_{ij}L_i^2\varepsilon_i^2\right\}}>0$$

$$\iff \sqrt{\max_{j\in[n]}\left\{\sum_{i=1}^n\left(\frac{P_{ij}L_i\varepsilon_i}{\mu_{\min}+\rho_{\min}}\right)^2\right\}}+\left\|\operatorname{Diag}\left(\frac{\boldsymbol{\rho}}{\mu_{\min}+\rho_{\min}}\right)\boldsymbol{A}\right\|_2<1-\frac{\max_{i\in[n]}\{L_i\varepsilon_i\}}{\mu_{\min}+\rho_{\min}}.$$

Lastly, the strong monotonicity property implies that the Multi-PP game (1) admits a unique NE (Facchinei & Pang, 2003). □

F. Missing Proofs in Section 4

F.1. Proof of Proposition 4.1

Proof. Let $\boldsymbol{Z}_i' := \boldsymbol{Z}_i - \boldsymbol{m}_i - \varepsilon\sum_{j=1}^n A_{ij}\boldsymbol{\theta}_j$, then $\mathbb{E}[\boldsymbol{Z}_i']=\mathbf{0}$ and $\operatorname{Var}[\boldsymbol{Z}_i']=\bar{p}\sigma^2$. We compute the following expectation w.r.t. distribution $\mathcal{D}(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})$ for some given $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_{\mathcal{N}_i}$:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Z}_i\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}\left[\frac{1}{2}\|\boldsymbol{Z}_i-\boldsymbol{\theta}_i\|_2^2\right]\\
&=\frac{1}{2}\mathbb{E}\left[\left\|\boldsymbol{Z}_i'+\boldsymbol{\mu}_i+\varepsilon\sum_{j=1}^n A_{ij}\boldsymbol{\delta}_j-\boldsymbol{\theta}_i\right\|_2^2\right]\\
&=\frac{1}{2}\operatorname{Var}\left[\boldsymbol{Z}_i'+\boldsymbol{\mu}_i+\varepsilon\sum_{j=1}^n A_{ij}\boldsymbol{\delta}_j-\boldsymbol{\theta}_i\right]+\frac{1}{2}\left\|\mathbb{E}\left[\boldsymbol{Z}_i'+\boldsymbol{m}_i+\varepsilon\sum_{j=1}^n A_{ij}\boldsymbol{\delta}_j-\boldsymbol{\theta}_i\right]\right\|_2^2\\
&=\frac{\bar{p}\sigma^2}{2}+\frac{1}{2}\left\|\boldsymbol{m}_i+\bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\delta}_j-\boldsymbol{\theta}_i\right\|_2^2.
\end{aligned}$$

i) The PSE satisfies

$$\begin{aligned}
&\boldsymbol{\theta}_i^{\mathrm{pse}}=\operatorname*{arg\,min}_{\boldsymbol{\theta}_i\in\mathbb{R}^{\bar{p}}}\left\{\frac{1}{2}\left\|\boldsymbol{m}_i+\bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^{\mathrm{pse}}-\boldsymbol{\theta}_i\right\|_2^2+\frac{\rho}{2}\sum_{j=1}^n A_{ij}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_j^{\mathrm{pse}}\|_2^2\right\}\\
&\iff \boldsymbol{\theta}_i^{\mathrm{pse}}-\boldsymbol{m}_i-\bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^{\mathrm{pse}}+\rho\sum_{j=1}^n A_{ij}\left(\boldsymbol{\theta}_i^{\mathrm{pse}}-\boldsymbol{\theta}_j^{\mathrm{pse}}\right)=\mathbf{0}\\
&\iff \left(1+\rho\sum_{j=1}^n A_{ij}-\bar{\varepsilon}P_{ii}\right)\boldsymbol{\theta}_i^{\mathrm{pse}}-\sum_{j\neq i}\rho A_{ij}\boldsymbol{\theta}_j^{\mathrm{pse}}-\sum_{j\neq i}\bar{\varepsilon}P_{ij}\boldsymbol{\theta}_j^{\mathrm{pse}}=\boldsymbol{m}_i\\
&\iff \left(\boldsymbol{I}_{\bar{p}n}+\rho(\boldsymbol{I}_n-\boldsymbol{A})\otimes\boldsymbol{I}_{\bar{p}}-\bar{\varepsilon}\boldsymbol{P}\otimes\boldsymbol{I}_{\bar{p}}\right)\boldsymbol{\theta}^{\mathrm{pse}}=\boldsymbol{m}.
\end{aligned}$$

Then, there exists a unique PSE if and only if $(1+\rho)\boldsymbol{I}_n-\rho\boldsymbol{A}-\bar{\varepsilon}\boldsymbol{P}$ is invertible (note that the PSE exists if and only if $\boldsymbol{m}\in\operatorname{range}\left(\boldsymbol{I}_{\bar{p}n}+\rho(\boldsymbol{I}_n-\boldsymbol{A})\otimes\boldsymbol{I}_{\bar{p}}-\bar{\varepsilon}\boldsymbol{P}\otimes\boldsymbol{I}_{\bar{p}}\right)$). Moreover, the closed-form solution of PSE reads

$$\begin{aligned}
\boldsymbol{\theta}^{\mathrm{pse}}&=\left(\boldsymbol{I}_{\bar{p}n}+\rho(\boldsymbol{I}_n-\boldsymbol{A})\otimes\boldsymbol{I}_{\bar{p}}-\bar{\varepsilon}\boldsymbol{P}\otimes\boldsymbol{I}_{\bar{p}}\right)^{-1}\boldsymbol{m}\\
&=\left([(1+\rho)\boldsymbol{I}_n-\rho\boldsymbol{A}-\bar{\varepsilon}\boldsymbol{P}]^{-1}\otimes\boldsymbol{I}_{\bar{p}}\right)\boldsymbol{m}.
\end{aligned}$$

Moreover, suppose that $\boldsymbol{\theta}^{t+1}=\mathcal{T}(\boldsymbol{\theta}^t)$, then we have for all $i\in[n]$,

$$\boldsymbol{\theta}_i^{t+1}=\operatorname*{arg\,min}_{\boldsymbol{\theta}_i\in\mathbb{R}^{\bar{p}}}\left\{\frac{1}{2}\left\|\boldsymbol{m}_i+\bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^{t}-\boldsymbol{\theta}_i\right\|_2^2+\frac{\rho}{2}\sum_{j=1}^n A_{ij}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_j^{t}\|_2^2\right\}. \tag{52}$$

The first-order optimality condition for (52) reads

$$\begin{aligned}
&\boldsymbol{\theta}_i^{t+1} - \boldsymbol{m}_i - \bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^t + \rho\sum_{j=1}^n A_{ij}\left(\boldsymbol{\theta}_i^{t+1} - \boldsymbol{\theta}_j^t\right) = \mathbf{0} \\
\iff &\left(1 + \rho\sum_{j=1}^n A_{ij}\right)\boldsymbol{\theta}_i^{t+1} = \boldsymbol{m}_i + \rho\sum_{j=1}^n A_{ij}\boldsymbol{\theta}_j^t + \bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^t \\
\iff &(1+\rho)\,\boldsymbol{\theta}_i^{t+1} = \boldsymbol{m}_i + \rho\sum_{j=1}^n A_{ij}\boldsymbol{\theta}_j^t + \bar{\varepsilon}\sum_{j=1}^n P_{ij}\boldsymbol{\theta}_j^t,
\end{aligned}$$

for all $i \in [n]$, which can be further written in the following compact form:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{m} + \left(\frac{\rho}{1+\rho}\boldsymbol{A}\otimes\boldsymbol{I}_{\bar{p}} + \frac{\bar{\varepsilon}}{1+\rho}\boldsymbol{P}\otimes\boldsymbol{I}_{\bar{p}}\right)\boldsymbol{\theta}^t.$$

Thus, the RRM converges to the PSE if and only if

$$\max_{i\in[n]}\left\{\left|\lambda_i\left(\frac{\rho}{1+\rho}\boldsymbol{A} + \frac{\bar{\varepsilon}}{1+\rho}\boldsymbol{P}\right)\right|\right\} < 1.$$

ii) The Nash equilibrium satisfies

$$\begin{aligned}
&\boldsymbol{\theta}_i^{\mathrm{ne}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}_i\in\mathbb{R}^{\bar{p}}}\left\{\frac{1}{2}\left\|\boldsymbol{m}_i - (1-\bar{\varepsilon})\boldsymbol{\theta}_i + \bar{\varepsilon}\sum_{j\neq i}P_{ij}\boldsymbol{\theta}_j^{\mathrm{ne}}\right\|_2^2 + \frac{\rho}{2}\sum_{j=1}^n A_{ij}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j^{\mathrm{ne}}\|_2^2\right\} \\
\iff &-(1-\bar{\varepsilon})\left(\boldsymbol{m}_i - (1-\bar{\varepsilon})\boldsymbol{\theta}_i^{\mathrm{ne}} + \bar{\varepsilon}\sum_{j\neq i}P_{ij}\boldsymbol{\theta}_j^{\mathrm{ne}}\right) + \rho\sum_{j=1}^n A_{ij}\left(\boldsymbol{\theta}_i^{\mathrm{ne}} - \boldsymbol{\theta}_j^{\mathrm{ne}}\right) = \mathbf{0} \\
\iff &\left(1 - \bar{\varepsilon} + \frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{\theta}_i^{\mathrm{ne}} - \frac{\rho}{1-\bar{\varepsilon}}\sum_{j\neq i}A_{ij}\boldsymbol{\theta}_j^{\mathrm{ne}} - \bar{\varepsilon}\sum_{j\neq i}P_{ij}\boldsymbol{\theta}_j^{\mathrm{ne}} = \boldsymbol{m}_i,
\end{aligned}$$

which can be written in the following compact form:

$$\left[\left(1 + \frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{I}_n\otimes\boldsymbol{I}_{\bar{p}} - \frac{\rho}{1-\bar{\varepsilon}}\boldsymbol{A}\otimes\boldsymbol{I}_{\bar{p}} - \bar{\varepsilon}\boldsymbol{P}\otimes\boldsymbol{I}_{\bar{p}}\right]\boldsymbol{\theta}^{\mathrm{ne}} = \boldsymbol{m}.$$

Thus, there exists a unique NE if and only if $\left(1+\frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{I}_n - \frac{\rho}{1-\bar{\varepsilon}}\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}$ is invertible (note that the NE exists if and only if $\boldsymbol{m} \in$ range $\left(\left(1+\frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{I}_n - \frac{\rho}{1-\bar{\varepsilon}}\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}\right)$). Moreover, the closed-form solution of NE reads

$$\boldsymbol{\theta}^{\mathrm{ne}} = \left(\left[\left(1 + \frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{I}_n - \frac{\rho}{1-\bar{\varepsilon}}\boldsymbol{A} - \bar{\varepsilon}\boldsymbol{P}\right]^{-1}\otimes\boldsymbol{I}_{\bar{p}}\right)\boldsymbol{m}.$$

Moreover, suppose that $\boldsymbol{\theta}_i^{t+1} = \mathcal{B}_i\left(\left[\boldsymbol{\theta}_j^t\right]_{j\in\mathcal{M}_i\cup\mathcal{N}_i}\right)$, then we have for all $i \in [n]$,

$$\boldsymbol{\theta}_i^{t+1} = \operatorname*{arg\,min}_{\boldsymbol{\theta}_i\in\mathbb{R}^{\bar{p}}}\left\{\frac{1}{2}\left\|\boldsymbol{m}_i - (1-\bar{\varepsilon})\boldsymbol{\theta}_i + \bar{\varepsilon}\sum_{j\neq i}P_{ij}\boldsymbol{\theta}_j^t\right\|_2^2 + \frac{\rho}{2}\sum_{j=1}^n A_{ij}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j^t\|_2^2\right\}. \tag{53}$$

Then, the first-order optimality condition for (53) reads

$$\begin{aligned}
&-(1-\bar{\varepsilon})\left(\boldsymbol{m}_i - (1-\bar{\varepsilon})\boldsymbol{\theta}_i^{t+1} + \bar{\varepsilon}\sum_{j\neq i}P_{ij}\boldsymbol{\theta}_j^t\right) + \rho\sum_{j=1}^n A_{ij}\left(\boldsymbol{\theta}_i^{t+1} - \boldsymbol{\theta}_j^t\right) = \mathbf{0} \\
\iff &\left(1 - \bar{\varepsilon} + \frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{\theta}_i^{t+1} = \boldsymbol{m}_i + \bar{\varepsilon}\sum_{j\neq i}P_{ij}\boldsymbol{\theta}_j^t + \frac{\rho}{1-\bar{\varepsilon}}\sum_{j=1}^n A_{ij}\boldsymbol{\theta}_j^t,
\end{aligned}$$

for all $i \in [n]$, which can be further written in the following compact form:

$$\left(1-\bar{\varepsilon}-\frac{\rho}{1-\bar{\varepsilon}}\right)\boldsymbol{\theta}^{t+1} = \boldsymbol{m} + \bar{\varepsilon}(\boldsymbol{P}\otimes\boldsymbol{I}_{\bar{p}})\boldsymbol{\theta}^t + \frac{\rho}{1-\varepsilon}(\boldsymbol{A}\otimes\boldsymbol{I}_{\bar{p}})\boldsymbol{\theta}^t$$
$$\iff \boldsymbol{\theta}^{t+1} = \frac{1-\bar{\varepsilon}}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{m} + \left(\left[\frac{(1-\bar{\varepsilon})\bar{\varepsilon}}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{P} + \frac{\rho}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{A}\right]\otimes\boldsymbol{I}_{\bar{p}}\right)\boldsymbol{\theta}^t.$$

Thus, the BR dynamics converges to the NE if and only if

$$\max_{i\in[n]}\left\{\left|\lambda_i\left(\frac{(1-\bar{\varepsilon})\bar{\varepsilon}}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{P} + \frac{\rho}{(1-\bar{\varepsilon})^2+\rho}\boldsymbol{A}\right)\right|\right\} < 1,$$

as desired. □

F.2. Derivations for the Approximations in Section 4.2

Proof. Computing the following expectation w.r.t. distribution $\mathcal{D}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i})$ for some $(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{\mathcal{N}_i})$:

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x}_i,y_i)\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}\left[\ell_i(\boldsymbol{\theta}_i;\boldsymbol{x}_i,y_i)\right]\\
=\;&\mathbb{E}_{(\boldsymbol{x}_i,y_i)\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}\left[-y_i\boldsymbol{\theta}_i^\top\boldsymbol{x}_i + \log\left(1+e^{\boldsymbol{\theta}_i^\top\boldsymbol{x}_i}\right)\right]\\
=\;&\mathbb{E}\left[\log\left(1+e^{\boldsymbol{\theta}_i^\top\boldsymbol{x}_i}\right)\;\middle|\; y_i=0\right]\Pr(y_i=0) + \mathbb{E}\left[-\boldsymbol{\theta}_i^\top\boldsymbol{x}_i + \log\left(1+e^{\boldsymbol{\theta}_i^\top\boldsymbol{x}_i}\right)\;\middle|\; y_i=1\right]\Pr(y_i=1).
\end{aligned} \tag{54}$$

Let $g(\boldsymbol{\theta}_i;\boldsymbol{x}_i) = \log\left(1+e^{\boldsymbol{\theta}_i^\top\boldsymbol{x}_i}\right)$, whose gradient is

$$\nabla g(\boldsymbol{\theta}_i;\boldsymbol{x}_i) = \frac{\boldsymbol{x}_i}{1+e^{-\boldsymbol{\theta}_i^\top\boldsymbol{x}_i}}.$$

The first-order approximations of $g(\boldsymbol{\theta}_i;\boldsymbol{x}_i)$, around $\mathbf{0}$ yields

$$g(\boldsymbol{\theta}_i;\boldsymbol{x}_i) \approx g(\mathbf{0};\boldsymbol{x}_i) + \nabla g(\mathbf{0};\boldsymbol{x}_i)^\top\boldsymbol{\theta}_i = \log(2) + \frac{1}{2}\boldsymbol{\theta}_i^\top\boldsymbol{x}_i.$$

Plugging this into (54) and using $\Pr(y_i=0)=q$, $\Pr(y_i=1)=1-q$ yield

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{x}_i,y_i)\sim\mathcal{D}_i(\boldsymbol{\delta}_i,\boldsymbol{\delta}_{\mathcal{N}_i})}\left[\ell_i(\boldsymbol{\theta}_i;\boldsymbol{x}_i,y_i)\right]\\
\approx\;&q\,\mathbb{E}\left[\frac{1}{2}\boldsymbol{\theta}_i^\top\boldsymbol{x}_i\;\middle|\; y_i=0\right] + (1-q)\,\mathbb{E}\left[-\boldsymbol{\theta}_i^\top\boldsymbol{x}_i + \frac{1}{2}\boldsymbol{\theta}_i^\top\boldsymbol{x}_i\;\middle|\; y_i=1\right] + \text{constant}\\
=\;&q\,\mathbb{E}\left[\frac{1}{2}\,\boldsymbol{\theta}_i^\top\left(\bar{\boldsymbol{x}}_i^0 + \bar{\varepsilon}_i\sum_{j=1}^n P_{ij}\boldsymbol{\delta}_j\right)\right] - (1-q)\,\mathbb{E}\left[\frac{1}{2}\boldsymbol{\theta}_i^\top\bar{\boldsymbol{x}}_i^1\right] + \text{constant}\\
=\;&\frac{q}{2}\boldsymbol{\theta}_i^\top\left(\boldsymbol{m}_i^0 + \bar{\varepsilon}_i\sum_{j=1}^n P_{ij}\boldsymbol{\delta}_j\right) - \frac{(1-q)}{2}\boldsymbol{\theta}_i^\top\boldsymbol{m}_i^1 + \text{constant}\\
=\;&\frac{q}{2}\boldsymbol{\theta}_i^\top\boldsymbol{m}_i^0 - \frac{1-q}{2}\boldsymbol{\theta}_i^\top\boldsymbol{m}_i^1 + \frac{q\bar{\varepsilon}_i}{2}\boldsymbol{\theta}_i^\top\sum_{j=1}^n P_{ij}\boldsymbol{\delta}_j + \text{constant}.
\end{aligned} \tag{55}$$

The PSE of multi-agent logistic regression satisfies the following system for $i\in[n]$:

$$\boldsymbol{\theta}_i^{\text{pse}} = \underset{\boldsymbol{\theta}_i\in\mathbb{R}^{\bar{p}}}{\arg\min}\left\{\mathbb{E}_{(\boldsymbol{x}_i,y_i)\sim\mathcal{D}(\boldsymbol{\theta}_i^{\text{pse}},\boldsymbol{\theta}_{\mathcal{N}_i}^{\text{pse}})}\left[\ell_i(\boldsymbol{\theta}_i;\boldsymbol{x}_i,y_i) + \frac{\rho_i}{2}\sum_{j=1}^n A_{ij}\left\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j^{\text{pse}}\right\|_2^2\right]\right\}.$$

In view of the first-order approximation of the expected risk in (55), the approximate PSE solution to system (55) satisfies

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_i^{\mathrm{pse}} &= \underset{\boldsymbol{\theta}_i \in \mathbb{R}^{\bar{p}}}{\arg\min} \left\{ \frac{q}{2} \boldsymbol{\theta}_i^\top \boldsymbol{m}_i^0 - \frac{1-q}{2} \boldsymbol{\theta}_i^\top \boldsymbol{m}_i^1 + \frac{q\bar{\varepsilon}_i}{2} \boldsymbol{\theta}_i^\top \sum_{j=1}^n P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} + \frac{\rho_i}{2} \sum_{j=1}^n A_{ij} \left\| \boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} \right\|_2^2 \right\} \\
&\Longleftrightarrow \frac{q}{2} \boldsymbol{m}_i^0 - \frac{1-q}{2} \boldsymbol{m}_i^1 + \frac{q\bar{\varepsilon}_i}{2} \sum_{j=1}^n P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} + \rho_i \sum_{j=1}^n A_{ij} \left(\widehat{\boldsymbol{\theta}}_i^{\mathrm{pse}} - \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} \right) = \mathbf{0} \\
&\Longleftrightarrow \left(\sum_{j=1}^n \rho_i A_{ij} \right) \widehat{\boldsymbol{\theta}}_i^{\mathrm{pse}} - \sum_{j=1}^n \rho_i A_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} + \sum_{j=1}^n \frac{q\bar{\varepsilon}_i}{2} P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} = \frac{1-q}{2} \boldsymbol{m}_i^1 - \frac{q}{2} \boldsymbol{m}_i^0 \\
&\Longleftrightarrow 2\rho_i \widehat{\boldsymbol{\theta}}_i^{\mathrm{pse}} - \sum_{j=1}^n 2\rho_i A_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} + \sum_{j=1}^n q\bar{\varepsilon}_i P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{pse}} = (1-q)\boldsymbol{m}_i^1 - q\boldsymbol{m}_i^0.
\end{aligned}$$

Since $\widehat{\boldsymbol{\theta}}^{\mathrm{pse}} = [\widehat{\boldsymbol{\theta}}_1^{\mathrm{pse}}; \ldots; \widehat{\boldsymbol{\theta}}_n^{\mathrm{pse}}]$ and $\boldsymbol{m} = [\boldsymbol{m}_1; \ldots; \boldsymbol{m}_n]$, we have

$$\left[(2\,\mathrm{Diag}(\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) + \mathrm{Diag}\left(q\bar{\boldsymbol{\varepsilon}}\right) \boldsymbol{P} \right) \otimes \boldsymbol{I}_p \right] \widehat{\boldsymbol{\theta}}^{\mathrm{pse}} = (1-q)\boldsymbol{m}^1 - q\boldsymbol{m}^0.$$

Then, if $2\,\mathrm{Diag}(\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) + \mathrm{Diag}\left(q\bar{\boldsymbol{\varepsilon}}\right) \boldsymbol{P}$ is invertible, we have

$$\widehat{\boldsymbol{\theta}}^{\mathrm{pse}} = \left(\left[(2\,\mathrm{Diag}(\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) + \mathrm{Diag}\left(q\bar{\boldsymbol{\varepsilon}}\right) \boldsymbol{P} \right]^{-1} \otimes \boldsymbol{I}_p \right) \left((1-q)\boldsymbol{m}_i^1 - q\boldsymbol{m}_i^0 \right),$$

as stated in the main paper.

By the way, we can also give the approximate NE using a similar approximation strategy. Specifically, the NE of multi-agent logistic regression can also be derived. The latter satisfies the following system for $i \in [n]$:

$$\boldsymbol{\theta}_i^{\mathrm{ne}} = \underset{\boldsymbol{\theta}_i \in \mathbb{R}^{\bar{p}}}{\arg\min} \left\{ \mathbb{E}_{(\boldsymbol{x}_i, y_i) \sim \mathcal{D}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\mathcal{N}_i}^{\mathrm{ne}})} \left[\ell_i(\boldsymbol{\theta}_i; \boldsymbol{x}_i, y_i) + \frac{\rho_i}{2} \sum_{j=1}^n A_{ij} \left\| \boldsymbol{\theta}_i - \boldsymbol{\theta}_j^{\mathrm{ne}} \right\|_2^2 \right] \right\}. \tag{56}$$

Based on the first-order approximation of the expected risk (55), the approximate NE solution to system (56) satisfies

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_i^{\mathrm{ne}} &= \underset{\boldsymbol{\theta}_i \in \mathbb{R}^{\bar{p}}}{\arg\min} \left\{ \frac{q}{2} \boldsymbol{\theta}_i^\top \boldsymbol{m}_i^0 - \frac{1-q}{2} \boldsymbol{\theta}_i^\top \boldsymbol{m}_i^1 + \frac{q\bar{\varepsilon}_i}{2} \|\boldsymbol{\theta}_i\|_2^2 + \frac{q\bar{\varepsilon}_i}{2} \boldsymbol{\theta}_i^\top \sum_{j \in \mathcal{N}_i} P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{ne}} + \frac{\rho}{2} \sum_{j \in \mathcal{M}_i} A_{ij} \left\| \boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_j^{\mathrm{ne}} \right\|_2^2 \right\} \\
&\Longleftrightarrow \frac{q}{2} \boldsymbol{m}_i^0 - \frac{1-q}{2} \boldsymbol{m}_i^1 + q\bar{\varepsilon}_i \widehat{\boldsymbol{\theta}}_i^{\mathrm{ne}} + \frac{q\bar{\varepsilon}_i}{2} \sum_{j \in \mathcal{N}_i} P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{ne}} + \rho_i \sum_{j \in \mathcal{M}_i} A_{ij} \left(\widehat{\boldsymbol{\theta}}_i^{\mathrm{ne}} - \widehat{\boldsymbol{\theta}}_j^{\mathrm{ne}} \right) = \mathbf{0} \\
&\Longleftrightarrow \left(q\bar{\varepsilon}_i + \sum_{j=1}^n \rho_i A_{ij} \right) \widehat{\boldsymbol{\theta}}_i^{\mathrm{ne}} - \sum_{j \neq i} \rho_i A_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{ne}} + \sum_{j \neq i} \frac{q\bar{\varepsilon}_i}{2} P_{ij} \widehat{\boldsymbol{\theta}}_j^{\mathrm{ne}} = \frac{1-q}{2} \boldsymbol{m}_i^1 - \frac{q}{2} \boldsymbol{m}_i^0,
\end{aligned}$$

for $i \in [n]$. Thus, we have

$$\left(\left(\mathrm{Diag}(q\bar{\boldsymbol{\varepsilon}}) + \mathrm{Diag}(\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) + \mathrm{Diag}\left(\frac{q}{2} \bar{\boldsymbol{\varepsilon}} \right) (\boldsymbol{P} - \boldsymbol{I}_n) \right) \otimes \boldsymbol{I}_p \right) \widehat{\boldsymbol{\theta}}^{\mathrm{ne}} = \frac{1-q}{2} \boldsymbol{m}_i^1 - \frac{q}{2} \boldsymbol{m}_i^0.$$

If $\mathrm{Diag}(q\bar{\boldsymbol{\varepsilon}}) + \mathrm{Diag}(\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) + \mathrm{Diag}\left(\frac{q}{2} \bar{\boldsymbol{\varepsilon}} \right) (\boldsymbol{P} - \boldsymbol{I}_n)$ is invertible, we have

$$\widehat{\boldsymbol{\theta}}^{\mathrm{ne}} = \left(\left[(\mathrm{Diag}(2q\bar{\boldsymbol{\varepsilon}}) + \mathrm{Diag}(2\boldsymbol{\rho})(\boldsymbol{I}_n - \boldsymbol{A}) - \mathrm{Diag}\left(q\bar{\boldsymbol{\varepsilon}}\right) (\boldsymbol{I}_n - \boldsymbol{P}) \right]^{-1} \otimes \boldsymbol{I}_p \right) \left((1-q)\boldsymbol{m}_i^1 - q\boldsymbol{m}_i^0 \right).$$

□

G. Additional Numerical Results

In the following, we present several additional numerical results for Section 4. Unless otherwise specified, we follow the same settings as described in the main paper, yet different network topology configurations and/or additional results will be used as described in the captions.

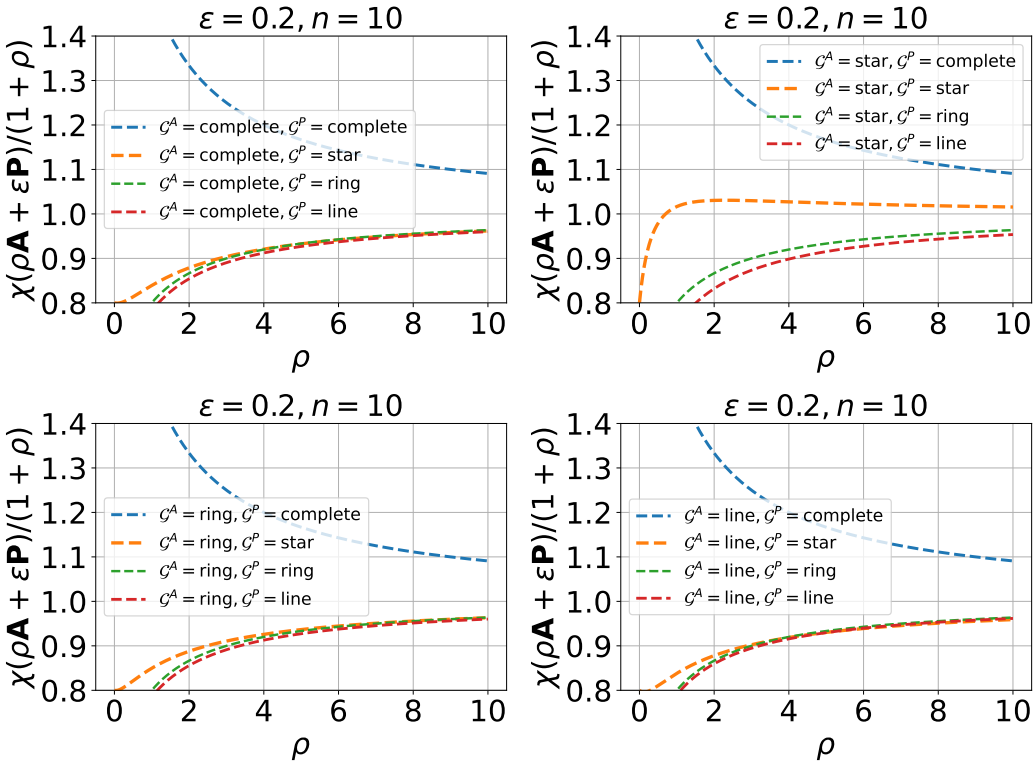

Figure 6: **Stability of PSE solution for MSE game** (17)**,** (18). Evaluating the condition (20) for different configurations of network topology with $\varepsilon = 0.2, n = 10$.

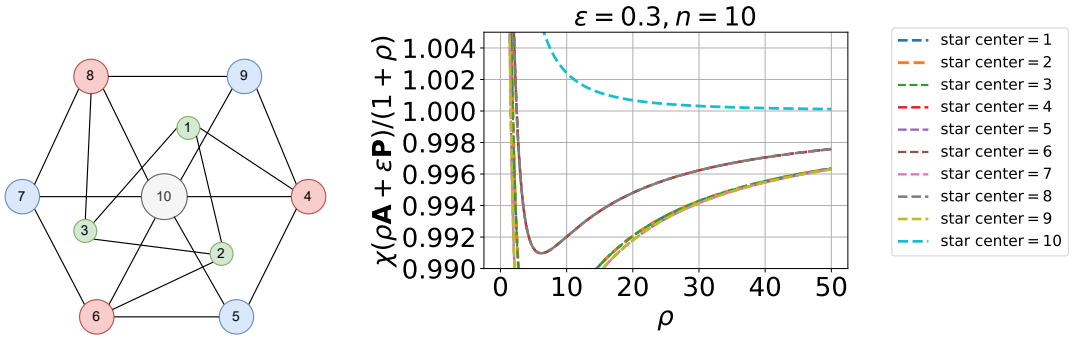

Figure 7: **Stability of PSE solution for the MSE game** (17), (18). We take $\mathcal{G}^{\mathtt{A}}$ as the Golomb graph (shown on the left), $\mathcal{G}^{\mathtt{P}}$ as the star graph, and evaluate (20) against ρ when the center node of $\mathcal{G}^{\mathtt{P}}$ is placed at different locations. Notice that when the center node of $\mathcal{G}^{\mathtt{P}}$ is '10', the PSE is never stabilized.

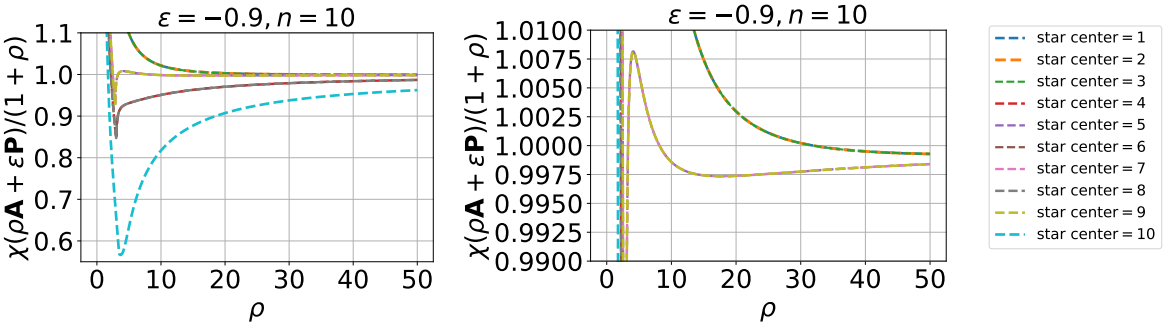

Figure 8: **Stability of PSE Solution for MSE Game** (17), (18). We take $\mathcal{G}^{\mathtt{A}}$ as the star graph, $\mathcal{G}^{\mathtt{P}}$ as the Golomb graph (shown in Figure 7), and evaluate (20) against ρ when the center node of $\mathcal{G}^{\mathtt{A}}$ is placed at different locations. Notice that when the center node of $\mathcal{G}^{\mathtt{A}}$ is '5' or '7' or '9', the PSE becomes unstable for a certain interval of ρ.

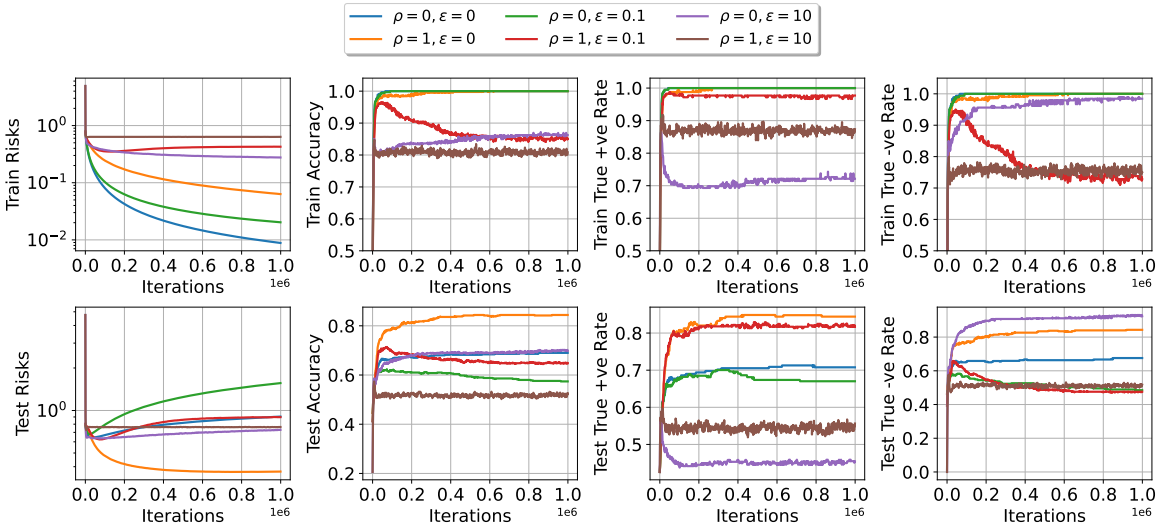

Figure 9: **Logistic regression game on synthetic dataset**. Learning dynamics of SG-GD in logistic regression problem with an ℓ_2-regularization $\frac{\lambda}{2}\|\theta_i\|^2$ and $\lambda = 10^{-4}$. ($\mathcal{G}^{\mathtt{A}}$: complete, $\mathcal{G}^{\mathtt{P}}$: star.)

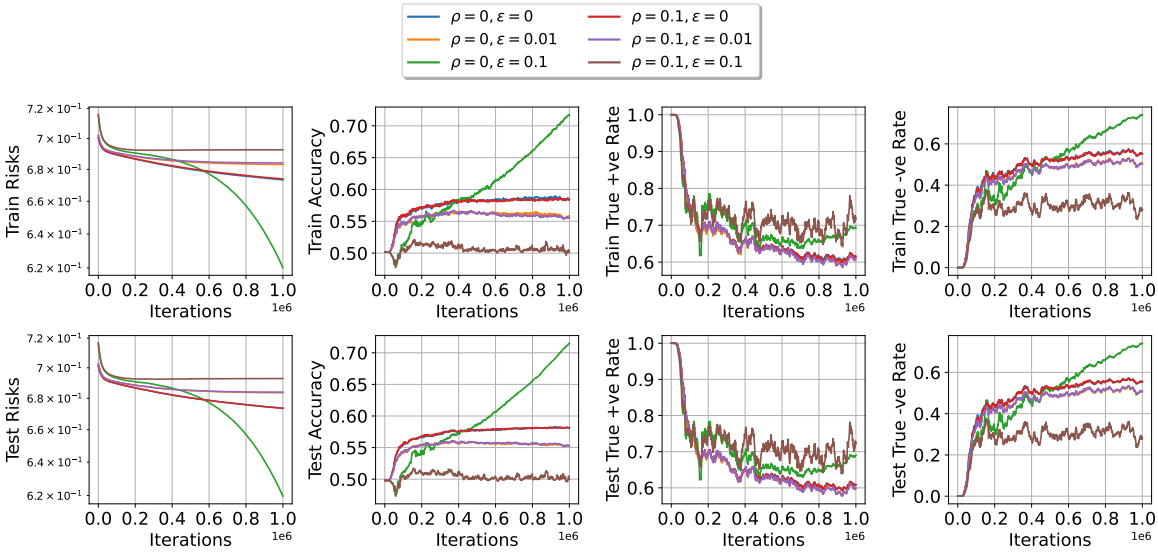

Figure 10: **Logistic Regression Game on the Kaggle Dataset**. Learning dynamics of RSGD for logistic regression on Kaggle Give Me Some Credit dataset in logistic regression problem with an ℓ_2-regularization $\frac{\lambda}{2}\|\theta_i\|^2$ and $\lambda = 10^{-4}$. ($\mathcal{G}^{\mathtt{A}}$: complete, $\mathcal{G}^{\mathtt{P}}$: star.)