Addressing Polarization and Unfairness in Performative Prediction

Kun Jin[*1], **Tian Xie**[*2], **Yang Liu**[3], **Xueru Zhang**[2]

[1] University of Michigan
[2] The Ohio State University
[3] University of California, Santa Cruz

Abstract

In many real-world applications of machine learning—such as recommendations, hiring, and lending—deployed models influence the data they are trained on, leading to feedback loops between predictions and data distribution. The *performative prediction* (PP) framework captures this phenomenon by modeling the data distribution as a function of the deployed model. While prior work has focused on finding *performative stable* (PS) solutions for robustness, their societal impacts, particularly regarding fairness, remain underexplored. We show that PS solutions can lead to severe polarization and prediction performance disparities, and that conventional fairness interventions in previous works often fail under model-dependent distribution shifts due to failing the PS criteria. To address these challenges in PP, we introduce novel fairness mechanisms that provably ensure **both stability and fairness**, validated by theoretical analysis and empirical results.

Code — https://github.com/osu-srml/FairPP

Extended version — https://arxiv.org/abs/2406.16756

1 Introduction

Modern supervised learning has achieved remarkable success in static environments, where the data distribution remains unaffected by model deployment. However, in many real-world applications—such as digital platforms, hiring, or lending—models influence user behavior, causing feedback loops that shift the data distribution. These model-induced shifts often render standard training methods unstable or unfair, and they are prevalent in real-world applications. Examples include strategic individuals manipulating their data (in school admission, hiring, or lending) to game the ML system into making favorable predictions (Hardt et al. 2016), consumers changing their retention and participation choices (in digital platforms) based on their perception toward the ML model they are subject to (Zhang et al. 2019; Chi et al. 2022).

To make predictions in the presence of model-induced distribution shifts, Perdomo et al. (2020) proposed **performative prediction** (PP), a framework that explicitly considers the target data distribution $\mathcal{D}(\boldsymbol{\theta})$ as a function of the ML model parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ to be optimized in a compact domain. While PP captures the impact of ML model on target data, the distribution $\mathcal{D}(\boldsymbol{\theta})$ is solely determined by the model regardless of the original data distribution. A subsequent study (Brown, Hod, and Kalemaj 2022) extended PP and proposed a more generalized **state-dependent performative prediction** (SDPP) framework, which considers the impacts of both model and initial data distribution. Specifically, given the deployed ML model parameter $\boldsymbol{\theta}$ and initial data distribution $\mathcal{D}$, SDPP models the resulting target data distribution $\mathcal{D}' = \mathrm{T}(\boldsymbol{\theta}; \mathcal{D})$ using some transition mapping function T. Since PP is a special case of SDPP, we focus on SDPP in this paper. When the transition map T is 1-jointly sensitive (details are in Def. 2.3), $\mathrm{T}(\boldsymbol{\theta}; \cdot)$ is contractive and repeatedly deploying $\boldsymbol{\theta}$ will cause the induced distributions to converge to a fixed point distribution $\mathcal{D}_{\boldsymbol{\theta}}$. The learning objective of SDPP is to minimize **performative risk** (PR) evaluated on $\mathcal{D}_{\boldsymbol{\theta}}$, i.e.,

$$\boldsymbol{\theta}^{\mathrm{PO}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{PR}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}_{Z \sim \mathcal{D}_{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}; Z)] \quad \text{s.t. } \mathcal{D}_{\boldsymbol{\theta}} = \mathrm{T}(\boldsymbol{\theta}; \mathcal{D}_{\boldsymbol{\theta}})$$

where $\ell(\boldsymbol{\theta}; Z)$ is the loss function, Z is the data sampled from the *fixed point* distribution $\mathcal{D}_{\boldsymbol{\theta}}$. The minimizer $\boldsymbol{\theta}^{\mathrm{PO}}$ is named as **performative optimal** (PO) solution. Because the target data distribution itself is a function of variable $\boldsymbol{\theta}$ to be optimized, finding $\boldsymbol{\theta}^{\mathrm{PO}}$ is often challenging (Perdomo et al. 2020; Brown, Hod, and Kalemaj 2022). Instead, existing works have mostly focused on finding **performative stable** (PS) solution $\boldsymbol{\theta}^{\mathrm{PS}}$, which minimizes the **decoupled performative risk** $\mathrm{DPR}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{PS}})$ defined as follows,

$$\boldsymbol{\theta}^{\mathrm{PS}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{DPR}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{PS}}) \stackrel{\text{def}}{=} \mathbb{E}_{Z \sim \mathrm{T}(\boldsymbol{\theta}^{\mathrm{PS}}; \mathcal{D}^{\mathrm{PS}})}[\ell(\boldsymbol{\theta}; Z)] \quad (1)$$

where $\mathcal{D}^{\mathrm{PS}}$ is the fixed point data distribution induced by $\boldsymbol{\theta}^{\mathrm{PS}}$ that satisfies $\mathcal{D}^{\mathrm{PS}} = \mathrm{T}(\boldsymbol{\theta}^{\mathrm{PS}}; \mathcal{D}^{\mathrm{PS}})$. Unlike $\mathrm{PR}(\boldsymbol{\theta})$ where data distribution $\mathrm{T}(\boldsymbol{\theta}; \mathcal{D})$ depends on variable $\boldsymbol{\theta}$ to be optimized, $\mathrm{DPR}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\mathrm{PS}})$ decouples the two, i.e., data distribution is induced by $\boldsymbol{\theta}^{\mathrm{PS}}$ while the variable to be optimized is $\boldsymbol{\theta}$. Although in general $\boldsymbol{\theta}^{\mathrm{PS}} \neq \boldsymbol{\theta}^{\mathrm{PO}}$, $\boldsymbol{\theta}^{\mathrm{PS}}$ is the fixed point of (1) and stabilizes the system: at $\boldsymbol{\theta}^{\mathrm{PS}}$, data distribution $\mathcal{D}^{\mathrm{PS}}$ also remains fixed. Many algorithms have been proposed in the literature to find $\boldsymbol{\theta}^{\mathrm{PS}}$. A prime example is *repeated risk minimization* (RRM) (Perdomo et al. 2020), an iterative algorithm that finds $\boldsymbol{\theta}^{\mathrm{PS}}$ (under certain conditions) by repeatedly

[*]These authors contributed equally.

updating the model $\boldsymbol{\theta}^{(t)}$ that minimizes risk on the fixed distribution $\mathcal{D}^{(t-1)}$ induced by the previous model $\boldsymbol{\theta}^{(t-1)}$, i.e.,

$$\begin{aligned} \boldsymbol{\theta}^{(t)} &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{Z\sim\mathcal{D}^{(t-1)}}[\ell(\boldsymbol{\theta}; Z)], \\ \mathcal{D}^{(t)} &= \mathrm{T}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}). \end{aligned} \tag{2}$$

However, the societal impact of PS solutions is less understood and it is unclear whether PS solutions can cause harm and violate social norms such as fairness.

In this paper, we examine the fairness properties of PS solutions. We consider scenarios where an ML model is used to make decisions about people from multiple social groups, and the population data distribution changes based on the ML model. We find that $\boldsymbol{\theta}^{\mathrm{PS}}$ can 1) incur severe polarization effects: entire population $\mathcal{D}^{\mathrm{PS}}$ is dominated by certain groups, leaving the rest marginalized and almost diminished in the system; 2) be biased when deployed on $\mathcal{D}^{\mathrm{PS}}$ and people from different groups will experience different losses. Because in many important domains such as job or loan applications, it is critical to ensure equal quality of ML predictions and population diversity, we investigate under what conditions and by what algorithms we can simultaneously achieve stability and fairness in SDPP.

Focusing on group-wise *loss disparity* (Martinez, Bertran, and Sapiro 2020; Diana et al. 2021; Khalili, Zhang, and Abroshan 2023) and *participation disparity* (Zhang et al. 2019; Raab et al. 2024) fairness measures, we first explore whether existing fairness mechanisms commonly used in supervised learning can help mitigate unfairness in SDPP settings; this includes *regularization methods* (adding fairness violation as a penalty term to the objective function of unconstrained optimization, e.g., (Khan, Herasymuk, and Stoyanovich 2023; Zhang et al. 2021)) and *reweighting methods* (adjusting weights and importance of samples in learning objective, e.g., (Jung et al. 2023; Duchi and Namkoong 2018; Duchi, Hashimoto, and Namkoong 2023)). We show that common choices of penalty terms (e.g., group-wise loss difference) and re-weighting designs (e.g., standard distributionally robust optimization) that are effective in traditional supervised learning may fail in SDPP by disrupting the stability of the system. Using repeated risk minimization (RRM) shown in (2) as an example, this means that applying such fairness mechanisms at each round of RRM can disrupt the convergence of the iterative algorithm and $(\boldsymbol{\theta}^{(t)}, \mathcal{D}^{(t)})$ may diverge to an unexpected state. We thus propose novel fairness mechanisms, which can be easily adopted and incorporated into iterative algorithms such as RRM. We theoretically show that the proposed mechanism can effectively improve fairness while maintaining the stability of the system.

It is worth noting that although a few recent works also studied fairness issues under model-induced distribution shifts, they all make rather strong assumptions about the distribution shifts and do not apply to the general SDPP framework. For example, Mishler and Dalmasso (2022) pointed out the fairness issues under performative settings without providing solutions to achieve fairness and stability at the same time. Zezulka and Genin (2023); Hu and Zhang (2022); Raab et al. (2024); Somerstep, Ritov, and Sun (2024) assumed there exists a causal model that depicts how data distribution would shift based on the ML model, and these causal models need to be fully known for the fairness mechanisms to work. Raab et al. (2024) studied a special type of model-induced distribution shift where only the group proportion changes. In App. A, we discuss more related works.

The rest of the paper is organized as follows. Section 2 provides the background of SDPP. Section 3 formulates the problem and demonstrates the unfairness and polarization issues of PS solutions. Section 4 highlights the difficulties of simultaneously achieving fairness and stability in SDPP, where we first show that existing fairness mechanisms commonly used in supervised learning may fail in SDPP settings and then propose a novel fairness mechanism. In Section 5, we conduct the theoretical analysis and show that our method can effectively improve fairness while maintaining stability. Finally, Section 6 empirically validates the proposed method on both synthetic and real data.

2 Preliminaries

Iterative algorithms to find PS solutions. As mentioned in Section 1, the original goal of SDPP is to find $\boldsymbol{\theta}^{\mathrm{PO}}$ that minimizes $\mathrm{PR}(\boldsymbol{\theta}) = \mathbb{E}_{Z\sim\mathcal{D}_{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}; Z)]$, the loss over the population induced by the deployed model. However, solving this optimization is often challenging because the data distribution $\mathcal{D}_{\boldsymbol{\theta}}$ depends on the variable $\boldsymbol{\theta}$ being optimized. Thus, prior studies such as (Perdomo et al. 2020; Brown, Hod, and Kalemaj 2022) have mostly focused on finding performative stable solution $\boldsymbol{\theta}^{\mathrm{PS}}$, which is the fixed point of Eqn. (1) and can be found through an iterative process of *data sampling* and *model deployment*. Specifically, denote $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \mathbb{E}_{Z\sim\mathcal{D}}[\ell(\boldsymbol{\theta}; Z)]$ and let $(\boldsymbol{\theta}^{(t)}, \mathcal{D}^{(t)})$ be the model parameter and data distribution at round t of the iterative algorithm, then repeatedly updating the model $\boldsymbol{\theta}^{(t)}$ according to Eqn. (3) could lead $(\boldsymbol{\theta}^{(t)}, \mathcal{D}^{(t)})$ converging to PS solution $(\boldsymbol{\theta}^{\mathrm{PS}}, \mathcal{D}^{\mathrm{PS}})$ under certain conditions (Perdomo et al. 2020; Brown, Hod, and Kalemaj 2022).

$$\boldsymbol{\theta}^{(t)} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^{(t-1)}), \quad \mathcal{D}^{(t)} = \mathrm{Tr}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}). \tag{3}$$

where Tr may not be the same as transition map T that drives evolution of data. Depending on how frequently the model is deployed compared to the change of data, Tr is defined differently based on **repeated deployment schema**. Common examples include:

$$\begin{aligned} \textbf{conventional: } & \mathrm{Tr}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}) = \mathrm{T}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}) \\ \textbf{k-delayed: } & \mathrm{Tr}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}) = \mathrm{T}^{k}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}) \\ & = \underbrace{\mathrm{T}\Big(\boldsymbol{\theta}^{(t)}; \ldots \mathrm{T}\big(\boldsymbol{\theta}^{(t)}; \mathrm{T}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)})\big)\Big)}_{k \text{ times}} \\ \textbf{delayed: } & \mathrm{Tr}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}) = \mathrm{T}^{\lceil r \rceil + 1}(\boldsymbol{\theta}^{(t)}; \mathcal{D}^{(t-1)}) \end{aligned}$$

where the *repeated risk minimization* (RRM) (Perdomo et al. 2020) introduced in Section 1 corresponds to conventional

deployment schema. By customizing the time interval between two deployments, we can get variants including *delayed RRM* and k*-delayed RRM* (Brown, Hod, and Kalemaj 2022). Note that the delayed deployment schema is a special case of k-delayed deployment schema, where the number of repeated deployments r is chosen to ensure the output distribution $\mathcal{D}^{(t)}$ is sufficiently close to the fixed point distribution when $\boldsymbol{\theta}^{(t)}$ keeps being deployed on the population $\mathcal{D}^{(t-1)}$.

Technical conditions for iterative algorithms to converge to PS solutions. As shown in (Perdomo et al. 2020; Brown, Hod, and Kalemaj 2022), PS solutions exist and are unique only when ℓ and T satisfy certain conditions. Moreover, iterative algorithms introduced in Eqn. (3) can converge to the PS solution. We introduce these conditions below, where Θ, $\mathcal{Z}$, and $\triangle(\mathcal{Z})$ denote the parameter space, sample space, and space of distributions over samples.

Definition 2.1 (Strong convexity of loss function)**.** $\ell(\boldsymbol{\theta}; Z)$ is γ-strongly convex if and only if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and $Z \in \mathcal{Z}$, we have

$$\ell(\boldsymbol{\theta}; Z) \geq \ell(\boldsymbol{\theta}'; Z) + \langle \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}'; Z), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle + \frac{\gamma}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2.$$

Definition 2.2 (Joint smoothness of loss function)**.** $\ell(\boldsymbol{\theta}; Z)$ is β-jointly smooth if the gradient with respect to $\boldsymbol{\theta}$ is β-Lipschitz in $\boldsymbol{\theta}$ and Z, i.e., $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and $\forall Z, Z' \in \mathcal{Z}$, we have

$$\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; Z) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}'; Z)\|_2 \leq \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$
$$\|\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; Z) - \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; Z')\|_2 \leq \beta \|Z - Z'\|_2$$

Definition 2.3 (Joint sensitivity of transition map)**.** Let $\mathcal{W}_1$ denote the Wasserstein-1 distance measure. The transition map T is ϵ-jointly sensitive if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and $\mathcal{D}, \mathcal{D}' \in \triangle(\mathcal{Z})$, we have

$$\mathcal{W}_1(\mathrm{T}(\boldsymbol{\theta}; \mathcal{D}), \mathrm{T}(\boldsymbol{\theta}'; \mathcal{D})) \leq \epsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$
$$\mathcal{W}_1(\mathrm{T}(\boldsymbol{\theta}; \mathcal{D}), \mathrm{T}(\boldsymbol{\theta}; \mathcal{D}')) \leq \epsilon \mathcal{W}_1(\mathcal{D}, \mathcal{D}')$$

Lemma 2.4 (Existence of a unique PS solution (Brown, Hod, and Kalemaj 2022; Perdomo et al. 2020))**.** *SDPP problem is guaranteed to have a unique PS solution if* ***all*** *of the following hold: (i) $\ell(\boldsymbol{\theta}; Z)$ is γ-strongly convex; (ii) $\ell(\boldsymbol{\theta}; Z)$ is β-joint smooth; (iii) T is ϵ-joint sensitive and $\epsilon(1 + 2\beta/\gamma) < 1$.*

Lemma 2.5 (Convergence of iterative algorithms)**.** *If conditions (i)(ii)(iii) in Lemma 2.4 are all satisfied, iterative algorithms are guaranteed to converge to the unique solution. However, there is no convergence guarantee when $\ell(\boldsymbol{\theta}; Z)$ is non-convex even if both (ii) and (iii) are satisfied.*

Lemma 2.4 has been shown in Theorem 8 of Brown, Hod, and Kalemaj (2022), while we prove Lemma 2.5 based on Perdomo et al. (2020) in App. D.1.

3 Unfairness and Polarization in SDPP

Problem formulation. In this work, we study SDPP with different demographic groups, where an ML model $\boldsymbol{\theta}$ is trained to make predictions about individuals from multiple groups distinguished by a sensitive attribute $s \in \mathcal{S}$ (e.g., gender, age, race), whose data distribution changes based on the deployed ML model and such model-induced distribution shift can be captured by transition map T. Suppose individuals from group s follow the identical data distribution $\mathcal{D}_s^{(t)}$ at the round t of an iterative algorithm, and let $p_s^{(t)}$ be the size of group s as the fraction of entire population at t. Then the data distribution of the entire population is $\mathcal{D}^{(t)} = \sum_{s \in \mathcal{S}} p_s^{(t)} \mathcal{D}_s^{(t)}$ with $\sum_{s \in \mathcal{S}} p_s^{(t)} = 1$.

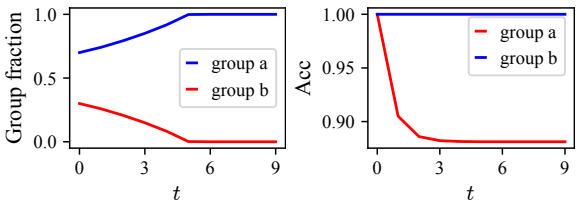


Figure 1: Illustrating examples of polarization effects and unfairness of $\boldsymbol{\theta}^{\mathrm{PS}}$ in Prop. 3.1 (left) and 3.2 (right): dynamics of the group fraction (left) and group-wise accuracy (right) under RRM: the system converges to $(\boldsymbol{\theta}^{\mathrm{PS}}; \mathcal{D}^{\mathrm{PS}})$ that is unfair (details are in App. D.2).

Note that the above SDPP with multiple groups is a general framework. By specifying the transition mapping T, many problems studied in prior works can be regarded as a special case. This includes:

1. **Strategic classification** (Hardt et al. 2016; Zhang et al. 2022): individuals in high-stakes applications such as lending, hiring, and college admission may manipulate their data based on ML model strategically to increase their chances of receiving favorable decisions, leading to changes in group distribution $\mathcal{D}_s^{(t)}$.
2. **Decision-making systems under user retention dynamics** (Zhang et al. 2019; Duchi and Namkoong 2018; Hashimoto et al. 2018a): ML models in recognition or recommendation systems may attract more users if they experience high accuracy but drive away those with less satisfaction, causing the group proportion $p_s^{(t)}$ to change.

We first explore the fairness properties of PS solutions in SDPP, i.e., examining whether $\boldsymbol{\theta}^{\mathrm{PS}}$ in Eqn. (1) have disparate impacts on different demographic groups. Specifically, we consider two fairness metrics: group-wise *loss disparity* $\triangle_{\mathcal{L}}^{(t)}$ (Martinez, Bertran, and Sapiro 2020; Hashimoto et al. 2018b) and *participation disparity* $\triangle_p^{(t)}$ (Raab et al. 2024), which measure the difference of group loss $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_s^{(t)})$ and fraction $p_s^{(t)}$ across different groups at round t of an iterative algorithm, respectively. In examples with two groups $\mathcal{S} = \{a, b\}$, the unfairness can be quantified as:

$$\triangle_{\mathcal{L}}^{(t)} := \left| \mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_a^{(t)}) - \mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_b^{(t)}) \right|,$$
$$\triangle_p^{(t)} := \left| p_a^{(t)} - p_b^{(t)} \right| \tag{4}$$

Unfairness & polarization effects in SDPP. We first show that PS solutions $\boldsymbol{\theta}^{\mathrm{PS}}$ in SDPP may have disparate im-

pacts on different groups. Specifically, when finding $\boldsymbol{\theta}^{\text{PS}}$ using iterative algorithms introduced in Section 2, the process may incur severe *polarization effects* and exhibit *unfairness*, i.e., certain groups may get more and more marginalized, and group-wise loss disparity gets exacerbated during the iterative process.

Proposition 3.1 (Polarization effects of $\boldsymbol{\theta}^{\text{PS}}$)**.** *Consider population from multiple groups with fixed group distribution $\mathcal{D}_s$ whose participation $p_s^{(t+1)}$ in an ML system depends on their perceived group loss $\mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_s)$. Suppose the deployment of system $\boldsymbol{\theta}^{(t)} = \arg\min_{\boldsymbol{\theta}} \sum_{s\in\mathcal{S}} p_s^{(t)} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_s)$ follows the conventional RRM schema, then there exist $\mathcal{D}^{(0)}$ and T such that as $t \to \infty$, $p_s^{(t)}$ changes monotonically and certain groups diminish entirely from the system.*

Proposition 3.2 (Exacerbated group-wise loss disparity)**.** *Consider population from multiple groups with fixed group proportion p_s; each individual is subject to a binary ML decision $\widehat{Y}^{(t)} = \mathbf{1}(X^{(t)} \geq \boldsymbol{\theta}^{(t)})$ and may strategically manipulate the data to increase the chance of receiving positive decisions. Suppose the deployment of ML system $\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} \Pr\left(\widehat{Y}^{(t)} = Y\right)$ follows the conventional RRM schema, and individuals manipulate features according to $X^{(t)} = X^{(t)} + \eta\boldsymbol{\theta}^{(t)}$ without changing Y(Perdomo et al. 2020), then there exists $\mathcal{D}^{(0)}$ such that group-wise loss disparity $\Delta_{\mathcal{L}}^{(t)}$ increases.*

To prove Prop. 3.1 and 3.2, it is sufficient to provide two examples to illustrate the polarization and unfairness effects in SDPP settings. We construct the examples with details in App. D.2 and visualize them in Figure 1. This shows that even though RRM can converge to a stable solution $\boldsymbol{\theta}^{\text{PS}}$, the solution is unfair and loss disparity $\triangle_{\mathcal{L}}^{(t)}$ and participation disparity $\triangle_p^{(t)}$ may get exacerbated.

4 Finding Fair-PS Solutions

Section 3 shows that without fairness consideration, PS solutions of SDPP may incur polarization effects and have disparate impacts on different groups. This section tackles unfairness issues in SDPP. One straightforward idea is to directly apply the fairness mechanisms at every round of the iterative algorithms introduced in Section 2. However, we will show that although such methods are effective in conventional supervised learning, they can disrupt the stability and the iterative algorithms may no longer converge.

4.1 Fair-PS Solutions

Many fairness mechanisms have been proposed in supervised learning to mitigate group-wise loss disparity and participation disparity. We consider two categories commonly used in the literature: *regularization method* and *sample reweighting method*, as detailed below.

1. **Fairness via regularization:** It adds a regularization or penalty term to the original learning objective function $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$, which penalizes the violation of fairness (Khan, Herasymuk, and Stoyanovich 2023; Zhang et al. 2021). The fair objective function is
$$\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}, \rho) := \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) + \mathcal{P}(\boldsymbol{\theta}; \mathcal{D}, \rho), \qquad (5)$$
where $\mathcal{P}(\boldsymbol{\theta}; \mathcal{D}, \rho)$ is the fair penalty term and the scalar $\rho > 0$ controls the strength of the penalty.
2. **Fairness via sample re-weighting:** It adjusts the weights of samples (possibly adversarially) and increases weights for disadvantaged groups (Jung et al. 2023; Duchi and Namkoong 2018; Duchi, Hashimoto, and Namkoong 2023). An example is distributionally robust optimization (DRO) (Hashimoto et al. 2018a), which minimizes worst-case loss and the fair objective is
$$\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}, \rho) := \max_{\tilde{\mathcal{D}} \in \mathcal{B}(\mathcal{D}, r(\rho))} \mathcal{L}(\boldsymbol{\theta}; \tilde{\mathcal{D}}), \qquad (6)$$
where $\mathcal{B}(\mathcal{D}, r(\rho)) := \{\tilde{\mathcal{D}} | d(\mathcal{D}, \tilde{\mathcal{D}}) \leq r(\rho)\}$ denotes a distribution ball centered at $\mathcal{D}$ with radius $r(\rho)$ derived from the fair mechanism strength ρ, and d is a distribution distance metric.

By optimizing a fair objective $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}, \rho)$, existing fairness mechanisms can effectively mitigate unfairness in supervised learning with static data distribution $\mathcal{D}$. However, it remains unclear how these methods would perform in SDPP when the model itself causes the data distribution shifts. Specifically, consider iterative algorithms introduced in Eqn. (3) that find PS solutions (e.g., RRM). Suppose we apply the above fairness mechanism at every round when updating the model parameter $\boldsymbol{\theta}$, i.e., replacing $\boldsymbol{\theta}^{(t)} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^{(t-1)})$ with fair version $\boldsymbol{\theta}^{(t)} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}^{(t-1)}, \rho)$ in iterative algorithms. We ask: *can such new iterative algorithms mitigate group-wise loss and participation disparity in SDPP and converge to a fair and stable solution?*

Before answering the above question, we first define **Fair-PS solutions** for SDPP, at which both ML system and population distribution reach stability and unfairness is mitigated.

Definition 4.1 (Fair-PS solution)**.** We define $(\mathcal{D}_{\mathsf{fair}}^{\text{PS}}, \boldsymbol{\theta}_{\mathsf{fair}}^{\text{PS}})$ as the Fair-PS solution to $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}, \rho)$ if
$$\mathcal{D}_{\mathsf{fair}}^{\text{PS}} = \mathrm{T}(\boldsymbol{\theta}_{\mathsf{fair}}^{\text{PS}}; \mathcal{D}_{\mathsf{fair}}^{\text{PS}}), \;\; \boldsymbol{\theta}_{\mathsf{fair}}^{\text{PS}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}_{\mathsf{fair}}^{\text{PS}}, \rho).$$

4.2 Existing Designs Fail to Converge to Fair-PS Solutions

We will use two examples to illustrate that the popular choices of fairness mechanisms used in existing literature may fail to achieve stability and fairness in SDPP. This includes (i) regularization method (Khan, Herasymuk, and Stoyanovich 2023; Zhang et al. 2021) with *group loss variance* $\mathcal{P}(\boldsymbol{\theta}; \mathcal{D}, \rho) := \rho \sum_{s\in\mathcal{S}} p_s \cdot [\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_s) - \mathcal{L}(\boldsymbol{\theta}; \mathcal{D})]^2$ as the penalty term in Eqn. (5); and (ii) *distributionally robust optimization* (DRO) method (Hashimoto et al. 2018b; Peet-Pare, Hegde, and Fyshe 2023) with χ^2-distance metric $d(\mathcal{D}, \tilde{\mathcal{D}}) := \int \left(\frac{d\mathcal{D}}{d\tilde{\mathcal{D}}} - 1\right)^2 d\tilde{\mathcal{D}}$ in Eqn. (6).

Example 4.2 (Group loss variance as penalty term)**.** *Consider two groups a, b with data $Z = (X, Y)$ and fixed group fractions $p_a = p_b = 0.5$. Suppose all samples in group a*

are $(1,-1)$ *and all samples in group* b *are* $(2,1)$*. Consider squared loss function* $\ell(z;\theta) = (y - h_\theta(x))^2 = (y-\theta x)^2$ *which is strongly convex and jointly smooth. Under group loss variance penalty* $\mathcal{P}(\boldsymbol{\theta};\mathcal{D},\rho)$*, we have* $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D},\rho) = 2.5\theta^2 - \theta + 1 + \rho\cdot(2.25\theta^4 + 9\theta^2 - 9\theta^3)$*. When* $\rho = 0.6$*, the second-order derivative* $\nabla^2_\theta \mathcal{L}_{\mathsf{fair}} = 16.2\theta^2 - 32.4\theta + 15.8$ *and is negative when* $\theta = 1$*. The negative second-order gradient means that* $\mathcal{L}_{\mathsf{fair}}$ *is nonconvex. According to Lemma 2.4, we cannot* ***ensure*** *the iterative algorithms converge to Fair-PS solutions when* ℓ *is nonconvex. Thus, adding group loss variance as a penalty at each round possibly disrupts the stability. App. C.5 verifies the non-convergence with empirical results.*

Example 4.3 (Repeated DRO with χ^2-distance metric)**.** *Consider two groups* a, b *with fixed data* $z_a = 1$ *and* $z_b = -1$ *for all samples but group fractions* $p_a^{(t)} = 0.5\cdot(1+\theta^{(t)})$ *and* $p_b^{(t)} = 0.5\cdot(1-\theta^{(t)})$*.* $p_a^{(0)} = 0.4, p_b^{(0)} = 0.6$*. Consider a mean estimation task where the model parameter* $\theta^{(t+1)} := \arg\min_{\boldsymbol{\theta}} \max_{\tilde{\mathcal{D}}\in\mathcal{B}(\mathcal{D}^{(t)},r)} \mathcal{L}(\theta;\tilde{\mathcal{D}})$ *is updated using DRO with mean squared error and* χ^2*-distance bound* $r = 1/6$*. Denote* $q_a^{(t)}, q_b^{(t)}$ *as the group fractions of the "worst-case" distribution* $\tilde{\mathcal{D}}$*. We have* $q_a^{(0)} = 0.6$*,* $q_b^{(0)} = 0.4$ *and* $\theta^{(1)} = \arg\min_\theta q_a^{(0)}(1-\theta)^2 + q_b^{(0)}(1+\theta)^2 = 0.2$*, which results in* $p_a^{(1)} = 0.6, p_b^{(1)} = 0.4$*. Since* $\mathcal{L}_a(\theta^{(1)};z_a) < \mathcal{L}_b(\theta^{(1)};z_b)$*, DRO should minimize the risk of the "worst-case" distribution with* $q_a^{(1)} = 0.4$*,* $q_b^{(1)} = 0.6$*, i.e.,* $\theta^{(2)} = \arg\min_\theta q_a^{(1)}(1-\theta)^2 + q_b^{(1)}(1+\theta)^2 = -0.2$*. Repeating the procedure we will get* $p_a^{(2)} = 0.4, p_b^{(2)} = 0.6$ *and* $q_a^{(2)} = 0.6, q_b^{(2)} = 0.4$*,* $\theta^{(3)} = 0.2$*. It turns out that repeated DRO results in* $\theta^{(t)}$ *oscillating between* 0.2 *and* -0.2 *and it never converges.*

It is worth noting that DRO methods have been used in Hashimoto et al. (2018b) to mitigate group fraction disparity in repeated optimizations. However, it only improves fairness without any convergence guarantees to stable solutions. Peet-Pare, Hegde, and Fyshe (2023) repeatedly used DRO to improve fairness under PP settings, it only converges to a PS solution under stronger assumptions where the distributionally robust objective must be strongly convex and jointly smooth, and the transition map T also needs to be ϵ-sensitive with respect to the worst-case distribution. Under milder conditions in Lemma 2.4, it may fail to converge as Example 4.3 illustrated.

4.3 Novel Designs for Fairness Mechanism

Next, we introduce three novel fair objective functions $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D},\rho)$ for fairness mechanisms, of which two belong to *regularization method* and one is a *sample re-weighting method*. By replacing $\boldsymbol{\theta}^{(t)} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta};\mathcal{D}^{(t-1)})$ with the proposed fair update $\boldsymbol{\theta}^{(t)} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D}^{(t-1)},\rho)$ in Eqn. (3), the resulting iterative algorithms can converge to Fair-PS solutions.

Proposed fair regularization (with and without demographics). Let $\mathcal{D}^{(0)}$ denote the initial population distribution. Depending on whether sensitive attributes S are accessible during training, we propose two fairness penalty terms, as detailed below.

1. Group level fairness penalty: It updates $\boldsymbol{\theta}^{(t)}$ by minimizing $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D}^{(t-1)},\rho)$ defined as follows
$$\mathcal{L}(\boldsymbol{\theta};\mathcal{D}^{(t-1)}) + \rho\sum_{s\in\mathcal{S}} p_s^{(t-1)}[\mathcal{L}(\boldsymbol{\theta};\mathcal{D}_s^{(t-1)})]^2 \tag{7}$$
2. Sample level fairness penalty without demographics: It updates $\boldsymbol{\theta}^{(t)}$ by minimizing $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D}^{(t-1)},\rho)$ defined as follows, which does not require access to sensitive attribute values.
$$\mathcal{L}(\boldsymbol{\theta};\mathcal{D}^{(t-1)}) + \rho\mathbb{E}_{Z\sim\mathcal{D}^{(t-1)}}\left[[\ell(\boldsymbol{\theta};Z)]^2\right]. \tag{8}$$

Proposed fair sample re-weighting. At the first round, it performs risk minimization. Starting from $t=2$, it updates $\boldsymbol{\theta}^{(t)}$ by minimizing $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D}^{(t-1)},\rho)$ defined as follows
$$\textstyle\sum_{s\in\mathcal{S}} q_s^{(t-1)}\mathcal{L}(\boldsymbol{\theta};\mathcal{D}_s^{(t-1)}) \tag{9}$$
with
$$\boldsymbol{q}^{(t-1)} = \left[q_s^{(t-1)}\right]_{s\in\mathcal{S}} = \frac{\boldsymbol{p}^{(t-1)} + \rho\boldsymbol{l}^{(t-1)}}{\|\boldsymbol{p}^{(t-1)} + \rho\boldsymbol{l}^{(t-1)}\|_1}$$
$$\boldsymbol{l}^{(t-1)} = \left[p_s^{(t-1)}\mathcal{L}(\boldsymbol{\theta}^{(t-1)};\mathcal{D}_s^{(t-2)})\right]_{s\in\mathcal{S}}$$
Unlike DRO in Peet-Pare, Hegde, and Fyshe (2023); Hashimoto et al. (2018b) that requires solving a min-max optimization at the current round, our re-weighting method only adjusts the weights for each group based on the group-wise losses in the previous round, which is more computationally efficient.

Comparison & discussion. Intuitively, compared to original $\mathcal{L}(\boldsymbol{\theta};\mathcal{D}^{(t-1)})$, all three proposed fair objective functions $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D}^{(t-1)},\rho)$ improves fairness at each round by assigning more weights to disadvantaged groups/samples (i.e., those experiencing higher losses) in the upcoming update. Indeed, both *sample level fairness penalty* and *fair sample re-weighting* can be regarded as modifications of *group level fairness penalty*. Comparing Eqn. (7) and (8), the two penalty terms get similar when most individual samples in the disadvantaged (resp. advantaged) groups are also similarly disadvantaged (resp. advantaged). This is more likely to happen when each group has a distribution with a small variance. For example, if the distribution of each group is a point mass, Eqn. (7) and (8) are identical.

Comparing Eqn. (7) and (9), we can rewrite Eqn. (9) as the following:
$$\begin{aligned}&\|\boldsymbol{p}^{(t-1)} + \rho\boldsymbol{l}^{(t-1)}\|_1\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta};\mathcal{D}^{(t-1)},\rho)\\ =&\sum_s p_s^{(t-1)}(1+\rho\mathcal{L}(\boldsymbol{\theta}^{(t-1)};\mathcal{D}_s^{(t-2)}))\,\mathcal{L}(\boldsymbol{\theta};\mathcal{D}_s^{(t-1)})\end{aligned}$$
We can see that the right-hand side will be a multiple of Eqn. (7) if we replace $\mathcal{L}(\boldsymbol{\theta}^{(t-1)};\mathcal{D}_s^{(t-2)})$ with $\mathcal{L}(\boldsymbol{\theta};\mathcal{D}_s^{(t-1)})$. This means both equations yield the same $\boldsymbol{\theta}^{(t)}$ when the population distribution does not change from $t-2$ to $t-1$, suggesting that the two approaches become more similar when the sensitivity of the transition map T is smaller, or equivalently, the distribution shift is milder.

5 Theoretical Analysis

Algorithm 1: Fair repeated risk minimization (Fair-RRM)

Require: $t = 0$, initial data distribution $\mathcal{D}^{(0)}$, strength of fair mechanism ρ, initial model parameter $\boldsymbol{\theta}^{(0)}$, stopping criteria τ
Choose repeated deployment schema and fair mechanism;
repeat
$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}^{(t)}, \rho)$;
Get $\mathcal{D}^{(t+1)} = \mathrm{Tr}(\boldsymbol{\theta}^{(t+1)}; \mathcal{D}^{(t)})$ from the chosen schema;
$t \leftarrow t + 1$;
until $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|_2 \leq \tau$

In this section, we will show that by replacing $\boldsymbol{\theta}^{(t)} = \mathrm{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}^{(t-1)})$ in Eqn. (3) with the fair update $\boldsymbol{\theta}^{(t)} = \mathrm{argmin}_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}^{(t-1)}, \rho)$ we proposed in Section 4.3, Fair-PS solutions (Definition 4.1) exist under certain conditions and the resulting iterative algorithms (Algorithm 1) can converge to such Fair-PS solutions. We assume conditions in Lemma 2.4 hold in this section and there exists a unique PS solution in the original SDPP problem. We first define a parameter $\tilde{\beta}$ which will be frequently used in the theorems introduced below.

$$\tilde{\beta} := \begin{cases} (2\rho\bar{\ell} + 1)\beta + 2\rho\tilde{\ell}^2, & \text{for regularization method (7) or (8)} \\ (\rho\bar{\ell} + 1)\beta, & \text{for sample re-weighting method (9)} \end{cases}$$

where $\bar{\ell} := \sup_{\boldsymbol{\theta}, Z} \|\ell(\boldsymbol{\theta}; Z)\|$ and $\tilde{\ell} := \sup_{\boldsymbol{\theta}, Z}\{\|\nabla \ell_{\boldsymbol{\theta}}(\boldsymbol{\theta}; Z)\|, \|\nabla \ell_Z(\boldsymbol{\theta}; Z)\|\}$. We can identify conditions under which a unique Fair-PS solution exists.

Proposition 5.1 (Existence of unique Fair-PS solution)**.** *For a given population with initial distribution $\mathcal{D}^{(0)}$ and the proposed fair mechanism $\mathcal{L}_{\mathsf{fair}}$ with strength ρ, there is a unique Fair-PS solution if $\epsilon(1+\tilde{\beta}/\gamma) < 1$. Moreover, $(\mathcal{D}^{PS}_{\mathsf{fair}}, \boldsymbol{\theta}^{PS}_{\mathsf{fair}})$ is independent of the choice of repeated deployment schema.*

Although the choice of repeated deployment schema does not affect the Fair-PS solution $(\mathcal{D}^{\mathrm{PS}}_{\mathsf{fair}}, \boldsymbol{\theta}^{\mathrm{PS}}_{\mathsf{fair}})$, it influences the convergence rate of the iterative algorithms in Theorem 5.2.

Theorem 5.2 (Convergence of Fair-RRM)**.** *Algorithm 1 converges to a Fair-PS solution under the following deployment schemas: (i) under **conventional** deployment schema, it converges to the Fair-PS solution at a linear rate if $\epsilon(1 + \tilde{\beta}/\gamma) < 1$; (ii) under k-**delayed** deployment schema, it converges to the Fair-PS solution at a linear rate for any k if $\epsilon(1 + \tilde{\beta}/\gamma) < 1 - \epsilon$; (iii) under **delayed** deployment schema when $r = \log^{-1}\left(\frac{1}{\epsilon}\right) \log\left(\frac{\mathcal{W}_1(\mathcal{D}^{(0)}, \mathcal{D}^{(1)})}{\delta}\right)$, it converges to within a radius δ of the Fair-PS solution (i.e., $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{PS}_{\mathsf{fair}}\|_2 \leq \delta$ and $\mathcal{W}_1(\mathcal{D}(\boldsymbol{\theta}^{(t)}), \mathcal{D}^{PS}_{\mathsf{fair}}) \leq \delta$) in $\mathcal{O}(\log^2 \frac{1}{\delta})$ steps if $\epsilon(1 + \tilde{\beta}/\gamma) < 1$.*

In App. B.3, we extend Algorithm 1 to *fair repeated empirical risk minimization (Fair-RERM)*. Although the theorems are for fairness mechanisms in Section 4, the proofs can be easily extended to a general class of $\mathcal{L}_{\mathsf{fair}}$ with convex and smooth fairness penalty terms (details in App. B.4).

Fairness guarantee. Thm. 5.2 and App. B.3 show that repeatedly minimizing $\mathcal{L}_{\mathsf{fair}}(\boldsymbol{\theta}; \mathcal{D}^{(t-1)}, \rho)$ on evolving data sequence can lead the system converging to a Fair-PS solution. Note that ρ controls the strength of fairness and different ρ could result in different $(\boldsymbol{\theta}^{\mathrm{PS}}_{\mathsf{fair}}, \mathcal{D}^{\mathrm{PS}}_{\mathsf{fair}})$. In conventional supervised learning with static data distribution $\mathcal{D}$, it is trivial to see that larger ρ will lead to a fairer solution (Martinez, Bertran, and Sapiro 2020). However, in SDPP settings, the impact of ρ on unfairness is less straightforward. Since both data distribution $\mathcal{D}^{\mathrm{PS}}_{\mathsf{fair}}$ and model $\boldsymbol{\theta}^{\mathrm{PS}}_{\mathsf{fair}}$ depend on ρ, analyzing how loss disparity would change as ρ varies can be highly complex. However, we manage to study fair mechanisms in Eqn. (7) with *group-level fairness penalty* to prove and quantify the fairness guarantee at the Fair-PS solution. We focus on a special case of user retention dynamics (Hashimoto et al. 2018b; Zhang et al. 2019) with two groups $s \in \{a, b\}$, where the group distribution $\mathcal{D}_s$ is fixed but the fraction $p_s^{(t)}$ changes based on group loss $\mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_s)$ during repeated risk minimization process.

Assumption 5.3. The majority group always experiences lower expected loss, i.e., $\arg\max_{s\in\{a,b\}} p_s^t = \arg\min_{s\in\{a,b\}} \mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_s)$. For two models deployed on population $\mathcal{D}^{(t)}$, the model with larger loss disparity $\triangle_{\mathcal{L}}^{(t)}$ leads to higher group fraction disparity $\triangle_p^{(t+1)}$ at time $t+1$.

Assumption 5.3 is natural: in applications such as recommendation systems, the minority group often suffers from higher loss and has a lower retention rate. Denote Fair-PS solution $\left(\boldsymbol{\theta}^{\mathrm{PS}}_{\mathsf{fair}}(\rho), \mathcal{D}^{\mathrm{PS}}_{\mathsf{fair}}(\rho)\right)$ as a function of $\rho \geq 0$ and the original PS solution as $(\boldsymbol{\theta}^{\mathrm{PS}}, \mathcal{D}^{\mathrm{PS}})$. Let $\triangle^{\mathrm{PS}}_{\mathsf{fair},\mathcal{L}}(\rho)$ be the group loss disparity of $\boldsymbol{\theta}^{\mathrm{PS}}_{\mathsf{fair}}(\rho)$ evaluated on distribution $\mathcal{D}^{\mathrm{PS}}_{\mathsf{fair}}(\rho)$ and $\Delta^{\mathrm{PS}}_{\mathcal{L}}$ be the group loss disparity of $\boldsymbol{\theta}^{\mathrm{PS}}$ evaluated at $\mathcal{D}^{\mathrm{PS}}$. We can first prove that a larger ρ leads to stronger fairness in Thm. 5.4.

Theorem 5.4. *Under Assumption 5.3, $\triangle^{PS}_{\mathsf{fair},\mathcal{L}}(\rho)$ is non-increasing in ρ.*

Furthermore, we can quantify the fairness improvement.

Theorem 5.5. *Denote p_s^{PS} as the fraction of group s at $\mathcal{D}^{PS}$ under retention dynamics. Assume $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ is twice continuously differentiable. For sufficiently small $\rho > 0$, we have*

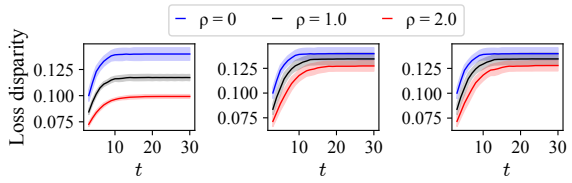
$$\Delta^{PS}_{\mathcal{L}} - \Delta^{PS}_{\mathcal{L},\mathsf{fair}}(\rho) = 2\rho\, p_a^{PS} p_b^{PS}\, \Delta^{PS}_{\mathcal{L}} \cdot \boldsymbol{v}_{\mathsf{fair}}^\top H^{-1} \boldsymbol{v}_{\mathsf{fair}} + \mathcal{O}(\rho^2),$$

where $\boldsymbol{v}_{\mathsf{fair}} = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{PS}; \mathcal{D}_a) - \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{PS}; \mathcal{D}_b)$, and $H = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{PS}; \mathcal{D}^{PS})$.

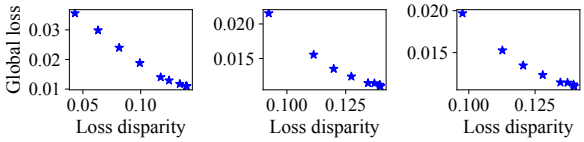
Since H is positive definite due to the strong convexity of $\mathcal{L}_{\mathsf{fair}}$ (Lemma D.4, Thm. 5.5 reveals the fairness improvement is positive when ρ is sufficiently small even without Assumption 5.3.

6 Numerical Results

This section empirically evaluates the proposed methods on synthetic and real data (including credit data (Perdomo et al. 2020; Kaggle 2012) and MNIST (Deng 2012)) under semi-synthesized performative shifts. We run all experiments with multiple random seeds and visualize the standard errors.

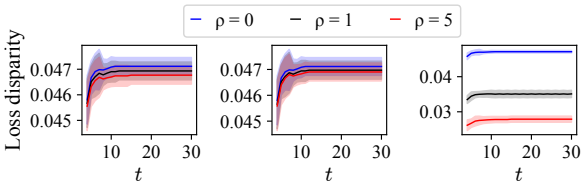


(a) Dynamics of loss disparity for Performative Gaussian mean estimation when $\rho \in \{0, 1.0, 2.0\}$. When $\rho = 0$, the policy reduces to standard RERM.

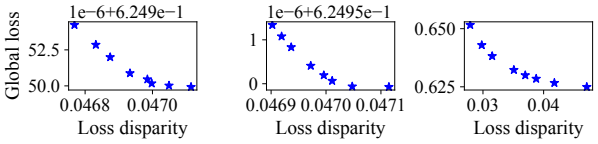


(b) Tradeoff between fairness and the global loss at the PS solution. "Stars" are produced by varying ρ.

Figure 2: Results on Gaussian data: Fair-RERM-RW (left), Fair-RERM-GLP (middle), Fair-RERM-SLP (right).



(a) Dynamics of group-wise loss disparity for Credit data when $\rho \in \{0, 1.0, 5.0\}$. When $\rho = 0$, the policy is just RERM.



(b) Tradeoff between fairness and the global loss at the PS solution. "Stars" are produced by varying ρ.

Figure 3: Results on Credit data: Fair-RERM-RW (left), Fair-RERM-GLP (middle), Fair-RERM-SLP (right).

Performative Gaussian mean estimation. We generate a synthetic dataset of 10000 samples from $s \in \{a, b\}$ with initial group fractions $p_a^{(0)} = 0.3, p_b^{(0)} = 0.7$ and target values $y_a = 0.3 + \epsilon, y_b = 0.7 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.05)$. At each round, the decision-maker estimates the mean of the current distribution as $\boldsymbol{\theta}^{(t)}$, and the loss function is the mean squared error with l_2 regularization. We consider user retention dynamics similar to Hashimoto et al. (2018b) where $p_s^{(t+1)} = \frac{\mathcal{R}(s,t)}{\sum_{s' \in \{a,b\}} \mathcal{R}(s',t)}$ changes based on group-wise loss. Here $\mathcal{R}(s,t) = \left(1 - \sum_{s' \in \{a,b\}} p_{s'}^{\min}\right) \times \frac{1}{2}\left(p_s^t + \frac{\mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_{-s}^{(t)})}{\mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_{-s}^{(t)}) + \mathcal{L}(\boldsymbol{\theta}^{(t)}; \mathcal{D}_s^{(t)})}\right) + p_s^{\min}$ where $p_s^{\min} = 0.02$ is the minimum group fraction, and $-s = \{a, b\} \setminus s$. We perform the mean estimation task empirically to train multiple linear regression models and compare RERM with Fair-RERM, including regularization methods with group-level penalty (Fair-RERM-GLP), sample-level penalty (Fair-RERM-SLP) and sample re-weighting method (Fair-RERM-RW). We perform 30 rounds of empirical risk minimization on 7 different random seeds. Fig. 2a illustrates the evolution of group-wise loss disparity $\triangle_{\mathcal{L}}^{(t)}$, where higher fairness improvement is achieved with higher ρ. The sample re-weighting method (Fair-RERM-RW) seems to yield better fairness control than others in this setting. Fig. 2b shows the tradeoff between fairness (group-wise loss disparity) and the global loss. The "stars" are produced by adjusting ρ and demonstrate a "Pareto-optimal" surface of each fairness mechanism.

Credit data retention dynamics with strategic behaviors. We use the *Give Me Some Credit* data (Kaggle 2012) consisting of features $X \in \mathbb{R}^{10}$ to measure individuals' creditworthiness $Y \in \{0, 1\}$ (Perdomo et al. 2020; Hu and Zhang 2022). We preprocessed the data similarly to Perdomo et al. (2020) and divided individuals into two groups $s \in \{a, b\}$ based on the age attribute. Next, we assume there is a newly established credit rating agency comparing RERM with Fair-RERM with logistic classification models to predict individuals' creditworthiness. Similarly, we assume the group-wise loss in round t affects the group fraction at $t+1$. Meanwhile, we assume there is a subset of features $X_s \in X$ which individuals can change strategically to X_s' based on the current model parameters θ. Specifically, $X_s' = X_s - \epsilon \cdot \theta_s$, where θ_s is the subset of θ with respect to X_s and $\epsilon = 0.1$. With the dynamics, we can visualize the evolution of group-wise loss disparity and the tradeoff between fairness and the global loss in Fig. 3. All results demonstrate the effectiveness of our methods where Fair-RERM-SLP seems to be more effective.

Additional experiments in App. C. Due to the page limit, we defer additional experiments to App. C.1. We want to highlight that we perform experiments on MNIST data (Deng 2012) to test whether our fairness mechanisms can still be useful beyond the convex setting with a deep learning model. Remarkably, Fig. 6a in App. C.1 verifies that Fair-RERM can still improve fairness at the PS solution. Moreover, Fig. 6b in App. C.1 demonstrates that the Fair-PS solution may converge to a local stationary point better for both fairness and performative loss in non-convex PP settings. We also perform experiments on ACSIncome dataset (Ding et al. 2022) (App. C.3) and show how our fairness mechanisms influence *Equal Opportunity* and *Demographic Parity* (App. C.7).

7 Conclusions & Limitations

Our work reveals unfairness issues of the PS solutions of PP. We propose novel fairness-aware algorithms to find Fair-PS solutions with the convergence holds under mild assumptions to facilitate trustworthy machine learning. However, the theory of this work does not cover non-convex settings and it still remains an open question for the future research.

Acknowledgements

This work was funded in part by the National Science Foundation under award number IIS2202699 and IIS-2416895.

References

Brown, G.; Hod, S.; and Kalemaj, I. 2022. Performative Prediction in a Stateful World. In Camps-Valls, G.; Ruiz, F. J. R.; and Valera, I., eds., *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 6045–6061. PMLR.

Caton, S.; and Haas, C. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1–38.

Chi, J.; Shen, J.; Dai, X.; Zhang, W.; Tian, Y.; and Zhao, H. 2022. Towards Return Parity in Markov Decision Processes. In *International Conference on Artificial Intelligence and Statistics*, 1161–1178. PMLR.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Diana, E.; Gill, W.; Kearns, M.; Kenthapadi, K.; and Roth, A. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 66–76.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2022. Retiring Adult: New Datasets for Fair Machine Learning. arXiv:2108.04884.

Duchi, J.; Hashimoto, T.; and Namkoong, H. 2023. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2): 649–664.

Duchi, J.; and Namkoong, H. 2018. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, 160–171. PMLR.

Guldogan, O.; Zeng, Y.; yong Sohn, J.; Pedarsani, R.; and Lee, K. 2023. Equal Improvability: A New Fairness Notion Considering the Long-term Impact. In *The Eleventh International Conference on Learning Representations*.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hardt, M.; and Mendler-Dünner, C. 2025. Performative prediction: Past and future. *Statistical Science*, 40(3): 417–436.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018a. Fairness Without Demographics in Repeated Loss Minimization. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1929–1938. PMLR.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018b. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938. PMLR.

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The Disparate Effects of Strategic Manipulation. In *ACM Conference on Fairness, Accountability, and Transparency*. Atlanta, Georgia.

Hu, Y.; and Zhang, L. 2022. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9549–9557.

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

Jia, Z.; Wang, Y.; Dong, R.; and Hanasusanto, G. A. 2024. Distributionally Robust Performative Optimization. *arXiv preprint arXiv:2407.01344*.

Jung, S.; Park, T.; Chun, S.; and Moon, T. 2023. Reweighting Based Group Fairness Regularization via Classwise Robust Optimization. arXiv:2303.00442.

Kaggle. 2012. Give Me Some Credit. https://www.kaggle.com/c/GiveMeSomeCredit/data.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware Learning through Regularization Approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650.

Khalili, M. M.; Zhang, X.; and Abroshan, M. 2023. Loss balancing for fair supervised learning. In *International Conference on Machine Learning*, 16271–16290. PMLR.

Khan, F. A.; Herasymuk, D.; and Stoyanovich, J. 2023. On Fairness and Stability: Is Estimator Variance a Friend or a Foe? *arXiv preprint arXiv:2302.04525*.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, 6755–6764. PMLR.

Mendler-Dünner, C.; Perdomo, J.; Zrnic, T.; and Hardt, M. 2020. Stochastic optimization for performative prediction. In *Advances in neural information processing systems*, 4929–4939. Curran Associates, Inc.

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

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2018. The Social Cost of Strategic Classification. arXiv:1808.08460.

Mishler, A.; and Dalmasso, N. 2022. Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings. arXiv:2202.05049.

Mofakhami, M.; Mitliagkas, I.; and Gidel, G. 2023. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, 11079–11093. PMLR.

Peet-Pare, G. L.; Hegde, N.; and Fyshe, A. 2023. Long Term Fairness via Performative Distributionally Robust Optimization.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative prediction. In *International Conference on Machine Learning*, 7599–7609. PMLR.

Raab, R.; Boczar, R.; Fazel, M.; and Liu, Y. 2024. Fair Participation via Sequential Policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 14758–14766.

Raab, R.; and Liu, Y. 2021. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34.

Somerstep, S.; Ritov, Y.; and Sun, Y. 2024. Algorithmic Fairness in Performative Policy Learning: Escaping the Impossibility of Group Fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 616–630.

Xie, T.; Tan, X.; and Zhang, X. 2024. Algorithmic decision-making under agents with persistent improvement. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1672–1683.

Xie, T.; and Zhang, X. 2024a. Automating data annotation under strategic human agents: Risks and potential solutions. *Advances in Neural Information Processing Systems*, 37: 127436–127482.

Xie, T.; and Zhang, X. 2024b. Non-linear welfare-aware strategic learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1660–1671.

Xie, T.; Zhu, D.; Liu, J.; Khalili, M.; and Zhang, X. 2025. SPRINT: Stochastic Performative Prediction With Variance Reduction. *arXiv preprint arXiv:2509.17304*.

Xie, T.; Zuo, Z.; Khalili, M. M.; and Zhang, X. 2024. Learning under Imitative Strategic Behavior with Unforeseeable Outcomes. *Transactions on Machine Learning Research*.

Xue, S.; and Sun, Y. 2024. Distributionally Robust Performative Prediction. *arXiv preprint arXiv:2412.04346*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zezulka, S.; and Genin, K. 2023. Performativity and Prospective Fairness. *arXiv preprint arXiv:2310.08349*.

Zhang, F.; Kuang, K.; Liu, Y.; Chen, L.; Wu, C.; Wu, F.; Lu, J.; Shao, Y.; and Xiao, J. 2021. Unified group fairness on federated learning. *arXiv preprint arXiv:2111.04986*.

Zhang, X.; Khalili, M. M.; Jin, K.; Naghizadeh, P.; and Liu, M. 2022. Fairness Interventions as (Dis)Incentives for Strategic Manipulation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 26239–26264. PMLR.

Zhang, X.; Khalili, M. M.; Tekin, C.; and Liu, M. 2019. *Group Retention When Using Machine Learning in Sequential Decision Making: The Interplay between User Dynamics and Fairness*. Red Hook, NY, USA: Curran Associates Inc.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellström, H.; Zhang, K.; and Zhang, C. 2020. How Do Fair Decisions Fare in Long-Term Qualification? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Zhao, Y. 2022. Optimizing the performative risk under weak convexity assumptions. *arXiv preprint arXiv:2209.00771*.